UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Evaluación de software para pruebas automatizadas y análisis de datos

TESINA

Que para obtener el título de

Ingeniero en Computación

P R E S E N T A

Adolfo Herrera Arias

DIRECTOR DE TESINA

Ing. Marco Antonio Martínez Quintana

Ciudad Universitaria, Cd. Mx., 2017

"Lo segundo, has de poner los ojos en quien eres, procurando conocerte a ti mismo, que es el más difícil conocimiento que puede imaginarse. Del conocerte saldrá el no hincharte, como la rana que quiso igualarse al buey [...]"

Del Decálogo de Don Quijote a Sancho Panza

"Ahora pues, ve, y yo estaré en tu boca, y te enseñaré lo que has de hablar"

Éxodo 4:12

"Libre, y para mí sagrado, es el derecho de pensar. La educación es fundamental para la felicidad social, es el principio en el que descansan la libertad y el engrandecimiento de los pueblos"

Benito Juárez

Agradecimientos

A mis padres por el enorme esfuerzo que realizaron para brindarme la educación.

A mis hermanos por el buen ejemplo que seguí de ellos.

A la Universidad Nacional Autónoma De México por aceptarme en sus aulas y formarme integralmente.

A mis maestros de la Facultad de Ingeniería por instruirme en la búsqueda del conocimiento y ayudarme a formar mi propio criterio.

A la DGTIC y a la División de Colaboración y Vinculación por permitirme prestar el servicio social en sus instalaciones y por el material facilitado.

A mis compañeros por el apoyo en los trabajos.

A la comunidad de Software libre...

Índice

Introducción

El presente informe estructurado en una tesina contiene el registro de las actividades realizadas en la Dirección General de Cómputo y de Tecnologías de Información y Comunicación bajo el programa de servicio social: Desarrollo de Sistemas**,** clave: 2016-12/6-528 que tiene por objetivo:

-Diseñar, desarrollar e implementar sistemas de información que permitan cumplir con los requerimientos de los usuarios internos y externos a la UNAM.

Y actividades:

-Colaborar en el análisis y desarrollo de bases de datos que sustenten la implementación de sistemas avanzados.

- Participar en el análisis, desarrollo, implementación y seguimiento de sistemas de aplicación compleja.

- Apoyar los procesos de diseño y desarrollo de interfaces inteligentes.

- Colaborar en el diseño, construcción, operación y mantenimiento de sistemas de aplicación compleja.

-Apoyar la instrumentación de estrategias de seguimiento para el óptimo funcionamiento de los sistemas.

Asimismo, las actividades de apoyo e investigación para la implementación y seguimiento de sistemas de cómputo se presentan en dos grandes rubros principales; el software para el análisis de datos y el software para las pruebas automatizadas. Además, se agrega una sección con las actividades adicionales para el cumplimiento de los objetivos del programa de servicio social mencionado y una sección con la documentación de problemas técnicos.

En la División de Colaboración y Vinculación de la DGTIC se encuentra en desarrollo el proyecto de “Sistema de Información Universitaria” que surge de las líneas de acción para el diagnóstico de la Universidad mediante las Tecnologías de Información y Comunicación que se expresaron en el plan de desarrollo institucional manifestado por el rector Dr. Enrique Graue.

Por otra parte, el Instituto Nacional de Acceso a la Información (INAI) realizó requerimientos a la UNAM a la vez que se realizaban proyectos para la plataforma de transparencia en la DCV. Bajo el contexto descrito, se apoyó en las actividades de diseño de bases de datos y extracción de requerimientos. Por otra parte, se apoyó con la demostración de análisis de datos y pruebas mediante la evaluación de software.

Las actividades realizadas tienen un sustento teórico que se integra en el capítulo I. El capítulo I presenta temas diversos que se presentaron durante la formación profesional en las aulas de la Facultad de Ingeniería y a la vez se buscó integrar conceptos actuales para encaminar la comprensión del contexto de este documento.

Se agregaron anexos con documentación más detallada (y de mayor evidencia) de las actividades y cada capítulo hace referencia al anexo correspondiente.

En el capítulo II se documenta la actividad de investigación de pruebas de software tomando el desarrollo de una aplicación de ejemplo que permitió tener habilidades de configuración e instalación en ambientes web y de software libre (Linux).

En el capítulo III se encuentra la parte analítica investigada, esta es la parte más extensa y la principal de este trabajo.

En el capítulo IV se documentaron, en resumen, los problemas técnicos que se presentaron durante el periodo de servicio y que pueden servir de referencia para otras actividades.

En el capítulo V se integraron las actividades adicionales que sirvieron de apoyo para la implementación de sistemas avanzados.

Las conclusiones se localizan en el capítulo VI en el que se reflexiona sobre los resultados obtenidos en las asignaciones, el aporte a la sociedad y a la formación profesional.

En general las actividades desarrolladas en el servicio social fueron de demostración de casos de análisis o de pruebas en la evaluación de software para tener una base de aprendizaje en la organización para desarrollo de proyectos.

Finalmente, la bibliografía y Mesografía se presenta con el formato ISO-690 con el apoyo de libros electrónicos de pago disponibles durante un año gratuitamente en el repositorio de safari para la UNAM durante la prestación del servicio social.

Capítulo I Marco teórico

1. Arquitecturas Cliente/Servidor

Cliente/Servidor es una arquitectura de red en la que cada equipo de trabajo o proceso es un cliente o un servidor. Dentro de los servidores se cuenta con equipos potentes de trabajo y dedicados a tareas específicas dentro de la red, por ejemplo, equipos para la gestión de ficheros, impresoras, gestión de tráfico, datos o incluso aplicaciones. Por otra parte, los clientes son menos potentes y utilizan los recursos de los servidores. En esta arquitectura se presenta la relación entre el cliente que realiza peticiones de servicios a un servidor que responde con procesos.

Cabe destacar que los procesos pueden ser ejecutados en diferentes procesadores lo que permite segmentar la aplicación en los niveles o áreas más convenientes. En cuanto a las funciones, la arquitectura se clasifica en varias capas, siendo la arquitectura de tres capas la más común y que está conformada de las siguientes áreas.

-Lógica de presentación: Es la interfaz con la que el usuario envía peticiones a la lógica de negocio, recibe y muestra al cliente los resultados de la petición.

-Lógica de negocio: Gestiona los datos a nivel de procesamiento, actúa como puente entre el usuario y los datos. Envía el resultado de la petición al nivel de presentación (por ejemplo: cálculo de nóminas, control de inventario).

-Lógica de datos: Se encarga de la gestión de los datos a nivel de almacenamiento, la integridad, mantenimiento y recuperación de los mismos.

2. Virtualización

Actualmente el desarrollo de la tecnología permite aprovechar más eficientemente los recursos del hardware a través del software. La virtualización logra este objetivo mediante un proceso en el que se crea una representación basada en software, y no física. Los servidores, las aplicaciones, el almacenamiento y las redes son tecnologías candidatas a estar virtualizadas para agilizar los procesos de negocio y reducir los costos[1].

[1] Es frecuente confundir los términos computacionales emulación y simulación.
La emulación se describe mejor como imitación de una determinada plataforma o programa informático en otra plataforma o programa (Koninklijke Bibliotheek,2017).

El equipo sobre el que se virtualiza se le llama huésped (host) y lo que se pretende virtualizar se llama anfitrión (guest).

2.1 La máquina virtual

Un sistema informático virtual se denomina "máquina virtual" (VM, Virtual Machine): un contenedor de software muy aislado en el que se incluyen un sistema operativo y aplicaciones. Cada una de las VM autónomas es completamente independiente. Si se colocan múltiples VM en una única computadora, es posible la ejecución de varios sistemas operativos y varias aplicaciones en un solo servidor físico o "host" (vmware, 2017).

3. Desarrollo de aplicaciones web

3.1.1 Aplicaciones web

Son aplicaciones basadas en el modelo Cliente/Servidor que se gestionan mediante servidores web, y que utilizan como interfaz páginas web.

La comunicación entre el cliente y el servidor se realiza principalmente mediante el protocolo HTTP que es un protocolo que pertenece a la familia de protocolos TCP/IP utilizados en internet[2].

3.1.2 Componentes semánticos

Los componentes semánticos de la web permiten interaccionar con el servidor a través de lenguajes estandarizados. A continuación se enlistan las propiedades de los componentes semánticos y como están estructurados algunos de ellos.

URI: Identifica los recursos web para su acceso y manipulación

URN: Identifica de forma única el recurso, independientemente de donde resida (RFC 2141)

URL: Señala donde se encuentra exactamente un recurso y está conformado por un esquema, servidor y nombre de recurso.

Sintaxis:

esquema://[usuario];[password]@:[dominio/IP][puerto]/;[parametros]?[consulta]#[sección]

Ejemplos:

http://www.hardware.com:2000/pc/check.cgi?item=1273&model=B
ftp://jose:suclave@www.hardware.com/informacion.txt

La simulación: Reproduce el comportamiento del fenómeno natural o artificial sin que el fenómeno necesariamente ocurra realmente.

[2] La familia de protocolos TCP/IP actúa sobre la capa siete del modelo OSI (Open System Interconnection).

HTML: Es el lenguaje que permite formatear texto, integrar imágenes y otros recursos.
HTTP: Es el protocolo de comunicación bien definido entre el cliente y el servidor.

3.2 Entornos de trabajo (Frameworks)

Son estructuras de software que contienen patrones de diseño genéricos para la creación de aplicaciones, en este caso en ambientes web, que facilitan el desarrollo con mayor rapidez y adaptabilidad ante cambios. Los frameworks tienen mayor estabilidad en cuanto infraestructura y seguridad en los datos. En otras palabras, un framework se puede considerar como una aplicación genérica incompleta y configurable a la que podemos añadirle las últimas piezas para construir una aplicación concreta.

3.3 Modelo-Vista-Controlador

Algunos entornos de trabajo (frameworks) se basan en un patrón de diseño; uno de los patrones más populares en la actualidad es el modelo-vista-controlador y consiste en la separación de los procesos en tres diferentes áreas, lo cual permite una mejor modificación y adaptación de la aplicación web ante cambios.

Las tres principales áreas se explican a continuación:

-Modelo: Implanta la lógica de negocio, es responsable de recuperar, procesar, validar y asociar cualquier tarea significativa mediante la manipulación de datos significativos para el negocio.

-Vista: Es la sección que se encarga de proporcionar una interfaz de interacción con el usuario y de la visualización de los recursos.

-Controlador: Es la capa que funciona como intermediario entre la vista y el modelo, administra las peticiones y delega los trabajos específicos a cada capa como la búsqueda de datos al modelo o la respuesta más adecuada a la vista.

4. Pruebas ágiles

Las pruebas de software forman parte del ciclo de desarrollo de software y permiten verificar el buen funcionamiento y la calidad de las aplicaciones. Por la naturaleza de las aplicaciones web, existen herramientas orientadas a verificar el correcto funcionamiento, la carga de datos y el rendimiento.

En esencia, las pruebas ágiles siguen las prácticas del manifiesto ágil[3] para el desarrollo de software, puesto que en la planeación y ejecución de las mismas se utilizan herramientas ligeras, de emulación, de generación de reportes comprensibles y adaptables a cambios. Las pruebas ágiles forman parte integral del desarrollo de software y engloban prácticas que involucran a todo el equipo de pruebas.

Sin importar el tipo de metodología ágil, lenguaje y tipo de aplicación, las siguientes pruebas son las más comunes en las aplicaciones web:

-Unitarias: Se refiere a la colección de pruebas para verificar los casos de uso de cada función o módulo.

-Integración: Comprobación del funcionamiento de las partes trabajando en conjunto.

-Regresión: Se comprueba que no se rompe alguna parte del sistema al agregar nuevas funciones o reparar errores.

-Funcionamiento del servidor web, base de datos, red y el sistema en su conjunto teniendo en cuenta la seguridad de la información.

4.1 Tipos de herramientas

Existen tres tipos de herramientas para el desarrollador o programador de pruebas que son:

-De validación HTML/CSS: permite validar que el código siga la norma W3C.

-De servicio: Pone a prueba la concurrencia del sistema y capacidad de clientes.

-De navegador: Mediante una extensión en el navegador se permite inspeccionar la respuesta entre el cliente y el servidor mostrando todo tipo de información sobre rendimiento, elementos y de red.

También las herramientas se clasifican de acuerdo al nivel lógico de la aplicación en el que operan; se enumera a continuación con algunas herramientas de ejemplo.

[3] Kent Beck, Mike Beedle y otros. 2001**.** agilemanifesto.org. [En línea] 2001. http://agilemanifesto.org/.

Capa	Herramienta
acceso a datos	DBUnit
servicios	EasyMock jMock
controladores	StrutsTestCase Spring Mocos Cactus
presentación	Canoo jWebUnit

Tabla 1. Herramientas de prueba para cada capa de software de la aplicación web

5. Bases de datos

Una base de datos es un repositorio de datos diseñado para almacenar, rescatar y mantener eficientemente la información. Una base de datos puede estar especializada en almacenar archivos binarios, documentos, imágenes, videos, datos relacionales, datos multidimensionales, datos transaccionales, datos analíticos o datos geográficos.

Los datos pueden almacenarse en varias formas, llámese tabular, jerárquica o gráfica. Si los datos están almacenados en formas tabulares entonces se le llama base de datos relacional. Cuando los datos están organizados en una estructura de árbol, entonces se le llama base de datos jerárquica. Los datos almacenados en gráficos que representan la relación entre objetos se refiere a una base de datos en red.

Existen cuatro propiedades fundamentales que deben cumplir las bases de datos y los sistemas de administración.

-Atomicidad (Atomicity)
-Consistencia (Consistency)
-Aislamiento (Isolation)
-Durabilidad (Durability)

5.2 Modelos de datos

La característica más importante de las bases de datos es que presente una abstracción de datos en diferentes niveles, con esto, los diferentes usuarios pueden percibir los datos en el nivel y detalle preferido. La abstracción de datos se refiere a la supresión de los detalles de almacenamiento, organización y conceptos significativos que hacen comprensible la lectura de los datos. Para lograr una abstracción de datos, es necesario utilizar un modelo que permita describir la estructura de los datos a través de una colección de conceptos. Muchos modelos incluyen operaciones básicas para la recuperación y actualización de los datos.

5.3 Niveles de abstracción

Se distinguen tres niveles de abstracción en las bases de datos los cuales son:

-Nivel interno: Tiene un esquema interno el cuál describe la estructura física de almacenamiento. El esquema interno utiliza el modelo físico de datos para describir la ruta de acceso a los datos y detalles de almacenamiento.

-Nivel conceptual: Tiene un esquema conceptual que describe la estructura y lógica del dominio de la base de datos para la comunidad de usuarios. El esquema conceptual se concentra en la descripción de entidades, relaciones, tipos de datos, operaciones sobre los datos y restricciones.

-Nivel externo: Incluye esquemas externos o vistas de usuario que describen parte de la base de datos para grupos en particular y ocultan datos que no son permitidos a estos.

5.4 Independencia en la base de datos

La independencia en una base de datos se refiere a la capacidad de cambiar un esquema en cualquier nivel de abstracción y que los cambios no afecten a los niveles restantes. Existen dos clasificaciones las cuáles son a nivel lógico y a nivel físico.

5.5 Fases de diseño

Al momento de implantar un sistema de información[4] con una base de datos se requiere de las siguientes fases de diseño:

[4] Los sistemas de bases de datos forman parte de los sistemas de información que, en sus inicios, eran implementados mediante archivos clasificados en papel, eventualmente fueron implantados mediante ambientes computarizados. El sistema de información incluye todos los recursos involucrados en la colección, administración, uso y diseminación de los recursos de información en la organización.

-Diseño conceptual: Se especifica la semántica de los requerimientos, las reglas de negocio y se representa mediante diagramas de entidad, atributos y relaciones; la representación más común en base de datos relacionales es el Modelo Entidad-Relación.

-Diseño lógico: Parte del diseño conceptual y emplea un modelo de datos para describir la estructura de la base de datos. En las bases de datos relacionales se emplea el Modelo Relacional

-Diseño físico: Describe la forma en se almacenan los datos y actualmente el Sistema Manejador de Base de Datos se encarga de esta tarea de acuerdo al diseño conceptual y lógico.

5.5.1 Modelo Entidad-Relación

Es un modelo conceptual de las entidades con atributos y sus interrelaciones propuesto por Peter Chenn, se representa de forma esquemática y generalmente mediante rectángulos (con una etiqueta en letras mayúsculas que indica que es una entidad), de elipses (con etiquetas en minúsculas que indican atributos de la entidad), y de rombos (indican la relación entre entidades).

La notación gráfica puede presentar cambios debido al tipo de concepto que se represente, por ejemplo: una llave primaria se indica con texto subrayado o una entidad débil presenta doble borde.

5.5.1.2 Modelo Relacional

Es el esquema mediante el cual se describen las tablas que conforman las bases de datos y las relaciones se indican mediante llaves foráneas a partir de llaves primarias. Se anotan las consideraciones semánticas y restricciones en los datos que presentarán las tuplas.

Una vez creado el modelo relacional se procede a adjuntar el diccionario de datos con la descripción de los tipos de datos, tipo de llaves y restricciones por cada tabla diseñada, con el fin de tener metadatos sobre la base para análisis posteriores y finalmente se conforman las sentencias de definición de datos en el DBMS (SMBD).

5.6.1 Sistema Manejador de Base de datos

Mejor conocido el acrónimo DBMS, el sistema manejador de base de datos es un conjunto de herramientas de software que permiten controlar el acceso, organizar, administrar, recuperar y mantener los datos en una base de datos. Es necesario contar con un DBMS porque debe de permitírsele a varios usuarios insertar, borrar y actualizar sobre el mismo repositorio sin que estas operaciones se encuentren estropeadas unas con otras.

5.6.2 Lenguajes del Sistema Manejador de Bases de Datos

El lenguaje de alto nivel empleado por los usuarios que permite manipular datos en el DBMS es SQL (Structure Query Languaje), este lenguaje fue creado como parte del proyecto System R de IBM que tenía en por objetivo en 1970 implantar de forma práctica el modelo relacional de Ted Codd.

SQL tiene tres variantes o categorías según la funcionalidad involucrada, las cuales son:

-DDL (Lenguaje de Definición de Datos): se utiliza para definir, cambiar o eliminar datos.
-DML (Lenguaje de Manipulación de Datos): se utiliza para leer o modificar datos y presenta dos tipos, a saber:

* Con procedimientos (Álgebra relacional).
* Sin procedimientos (Cálculo relacional).

-DCL (Lenguaje de Control de Datos): se utiliza para permitir o revocar autorizaciones.

También existen lenguajes asociados al DBMS para aplicaciones y pueden ser externos (como lenguaje C/C++, java) e internos como PL/SQL (Programing Languaje /SQL).

PL/SQL utiliza variantes de SQL en procedimientos o funciones almacenadas o en línea.

5.6.3 Álgebra relacional

Al manipular datos relacionales; se emplean, según Ted Codd, las operaciones tradicionales: unión, intersección, diferencia, producto cartesiano, y las operaciones especiales relacionales: selección, proyección, join y división.

5.6.4 Operaciones tradicionales

-Union: Cada tabla debe tener los mismos campos para “concatenar” las tuplas de la segunda tabla al final de las tuplas primera tabla.

-Intersección: Dadas dos tablas con igual definición, después de ejecutar la operación se obtienen las tuplas comunes en una nueva tabla.

-Diferencia: Se obtienen las tuplas de la primera tabla quitándole las tuplas comunes con la tabla segunda.

-Producto cartesiano: Se relaciona cada tupla de la primera tabla con cada tupla de la segunda tabla.

5.6.5 Operaciones especiales

-Selección: Indica que tupla o tuplas deben aparecer en la consulta ya que cumplen con el criterio de selección.

-Proyección: Se refiere a los campos que cada tupla seleccionada debe mostrar.

-Join: Existen varios tipos, pero la finalidad última es mezclar campos de tablas bajo una llave o campo, por ejemplo: obtener tablas desnormalizadas para el análisis de datos.

-División: Consiste en presentar todos los conjuntos de tuplas organizados como lo indica la tabla divisor.

6. Tendencias de los sistemas de bases de datos

La evolución de las aplicaciones de las bases de datos y los sistemas de administración permitieron integrar disciplinas analíticas para extraer información importante para las organizaciones. La figura 1 muestra la tendencia de los sistemas de bases de datos.

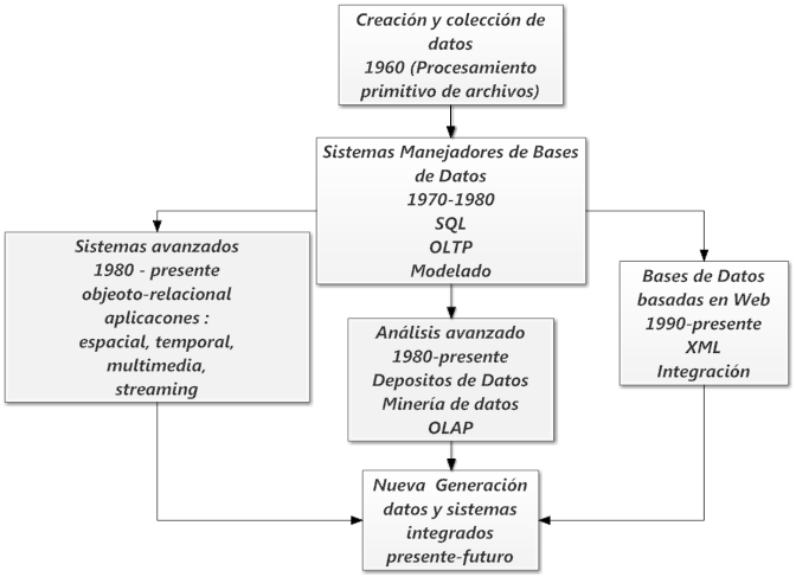


Figura 1. Evolución de las Bases de Datos
Fuente: "Data Mining: Concepts and Techniques" (2006)

6.1 Big Data (Macrodatos)

Actualmente el concepto de Big Data es más comprensible debido a un estudio de IBM en el que encuestó a varios perfiles de tecnología y de negocios para describir lo que entendían sobre Big Data y lograr encauzar la comprensión del término, del resultado se establecieron cuatro características para identificar el concepto.

-Volumen: La gran cantidad de datos aprovechables para la toma de decisiones de las organizaciones. Un volumen alto dependerá del sector y ubicación geográfica en el que se basan los datos pero los encuestados coindicen con el crecimiento eventual de los datos.

-Variabilidad: Tiene que ver con gestionar gran variedad y complejidad de los tipos de datos, incluyendo los datos estructurados, semi-estructurados y no estructurados. La integración de diversas fuentes de información internas y externas de la organización.

-Velocidad: La velocidad a la que se crean, procesan y analizan los datos continua aumentando, las herramientas tradicionales les resulta imposible capturar, almacenar y analizar los datos.

-Veracidad: se refiere al nivel de fiabilidad asociado a cierto tipo de datos. El reto actual de Big Data es obtener datos de alta calidad tomando en cuenta la imprevisibilidad de los datos.

En definitiva las cuatro características expresadas anteriormente definen Big Data y permiten a las organizaciones obtener ventajas competitivas en el mercado digitalizado.

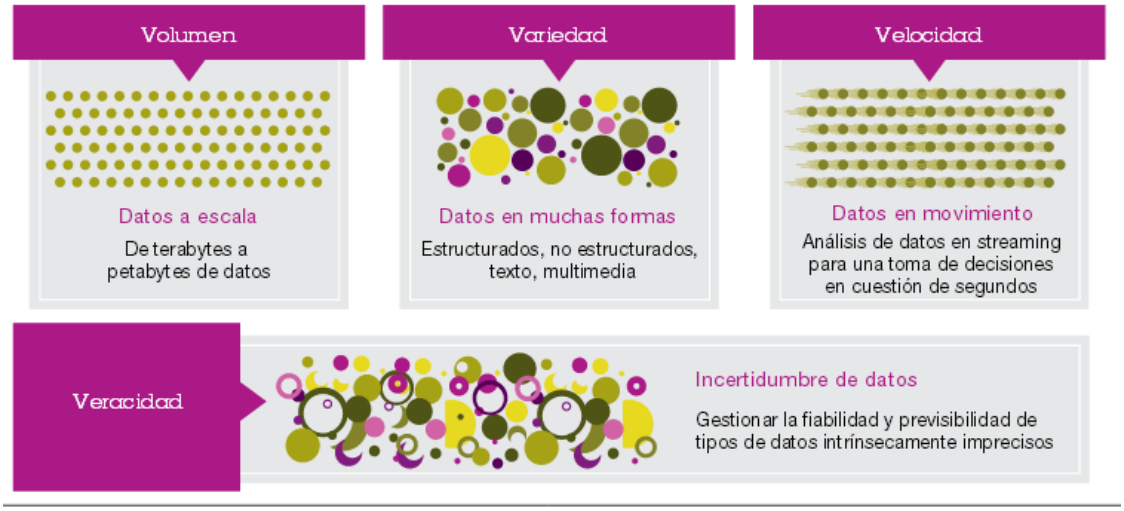


Figura 2. Dimensiones de Big Data
Fuente: IBM Corporation, El uso de Big Data en el mundo real (2012)

7. Minería de datos

La Minería de datos[5], según el Grupo Gartner, es:

"El proceso de buscar nuevas y significativas correlaciones, patrones y tendencias mediante la examinación de grandes cantidades de datos almacenados en repositorios, utilizando tecnologías de reconocimiento de patrones también conocidas como técnicas matemáticas y estadísticas".

Es común que se utilice el término KDD (Knowledge Discovery from Data) como sinónimo de Minería de datos. Alternativamente se le conoce como un paso más en el proceso de extracción de conocimiento.

La naturaleza multidisciplinaria entorno a la Minería de datos se muestra en la figura 3.

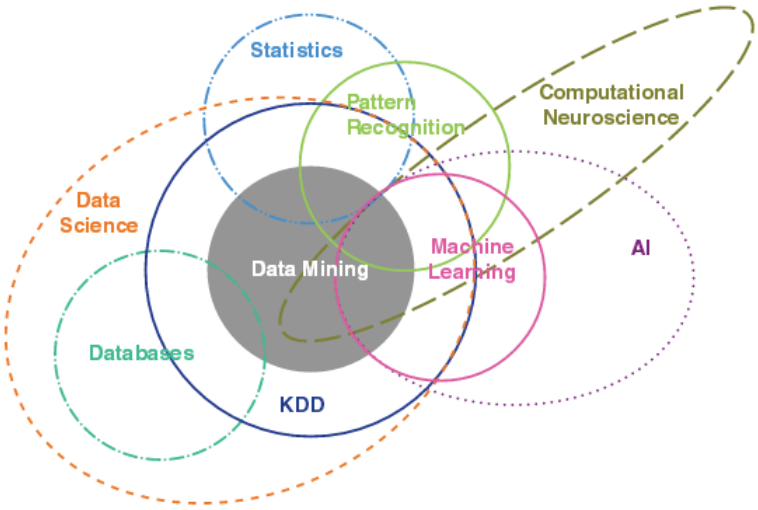


Figura 3. Naturaleza Interdisciplinarias de la Minería de datos
Fuente: SAS Enterprise, Material de entrenamiento en Minería (1998)

7.1 Minería de datos y Big Data

Actualmente las herramientas de análisis de datos son descritas como plataformas de descubrimiento de datos (Data Discovery Platforms), estas herramientas están diseñadas para simplificar el análisis de grandes volúmenes de datos.

[5] Términos similares pero con diferencias específicas son: Análisis de patrones/datos, arqueología de datos entre otros.

La administración, manejo y mantenimiento de grandes cantidades de datos no es tarea fácil, por lo que extraer muestras de calidad en repositorios muy grandes es la clave para la veracidad de los datos en Big Data, con esto se logra aportar valor a los detalles en el negocio de la organización para los clientes.

El término Minería de datos cada vez está siendo sustituido en los nuevos entornos de análisis en los sistemas de información, pero forma parte de los conceptos que se han forjado en el pasado sobre los sitemas de información a medida que ha cambiado el paradigma sobre el procesamiento de los datos. De hecho (Sumathi, y otros, 2006) hablan de una evolución de los algoritmos de Minería de datos para escalas más grandes de procesamiento.

7.2 Procesos de Minería de datos

Las etapas del proceso de extracción de conocimiento (KDD, Knowledge Data Discovery) se describen gráficamente en la figura 4.

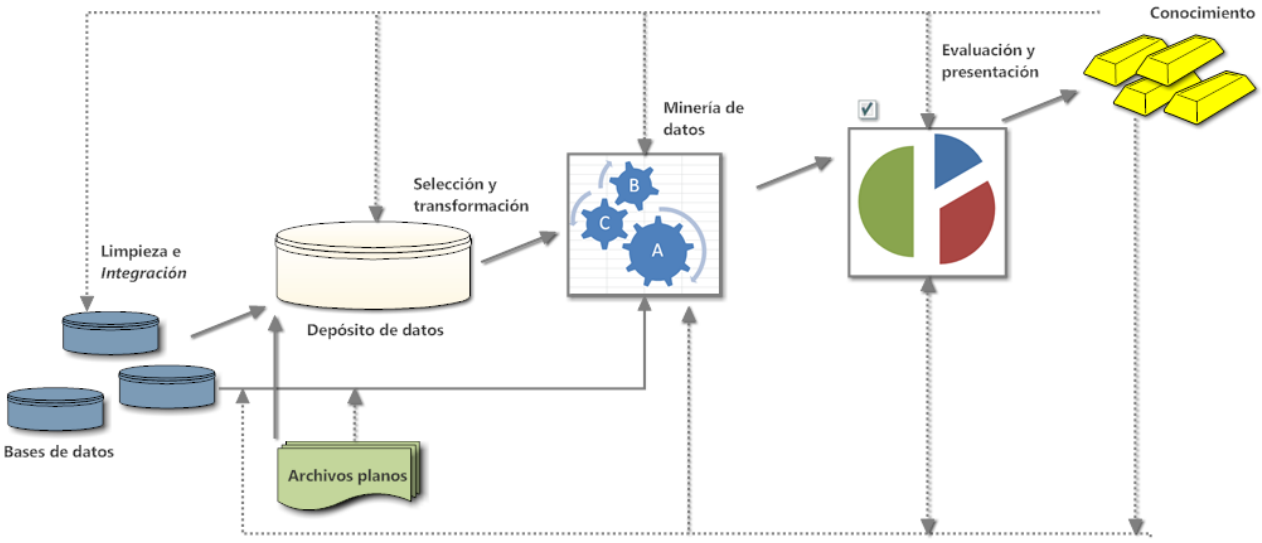


Figura 4. Proceso KDD

7.3 Técnicas

Gran parte de la bibliografía no marca exactamente la diferencia entre modelos descriptivos y modelos predictivos, pero algunos autores si utilizan esta clasificación y se refieren a estos conceptos como técnicas. A continuación se muestran los esquemas de la clasifican según dos autores que se complementan el uno con el otro.

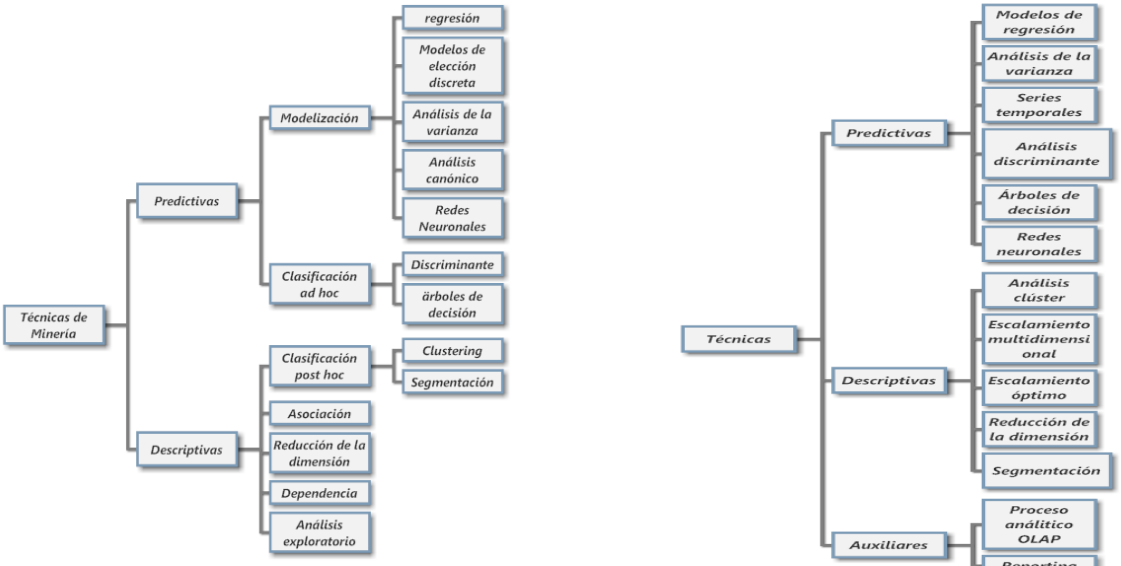


Figura 5. a) Clasificación según Pérez L.César,2008 b) Clasificación según Pérez M. Maria, 2014

Como se puede observar en la clasificación de las técnicas de las figura 5, las técnicas consisten en ponderar las variables de análisis de tipo cuantitativo y cualitativo para dar prioridades de dependencia en los modelos.

Dentro de las técnicas auxiliares emergen herramientas para reportar datos y visualizarlos con técnicas de diseño gráfico para hacer llegar el mensaje de los resultados de manera profesional[6]. También las técnicas de los depósitos de datos (ETL) forman parte de las técnicas auxiliares porque permiten preparar los datos en un sistema de información.

7.4 Modelo de proceso CRISP

Para una organización es importante entender las reglas del negocio y exploración de los datos, es por ello que se ha adoptado un proceso estándar que ha evolucionado con el tiempo, se trata del modelo CRISP.

El modelo de proceso CRISP-DM tiene por objetivo implantar la Minería de datos en la estrategia general para resolver un problema de negocio. En la figura 6 se muestra el modelo CRISP-DM que evoluciona a partir del proceso KDD y que toma lugar en la Minería de datos.

[6] La herramienta Tableau se sitúa como una técnica auxiliar del análisis de datos, esta herramienta se investigó durante la prestación del servicio social.

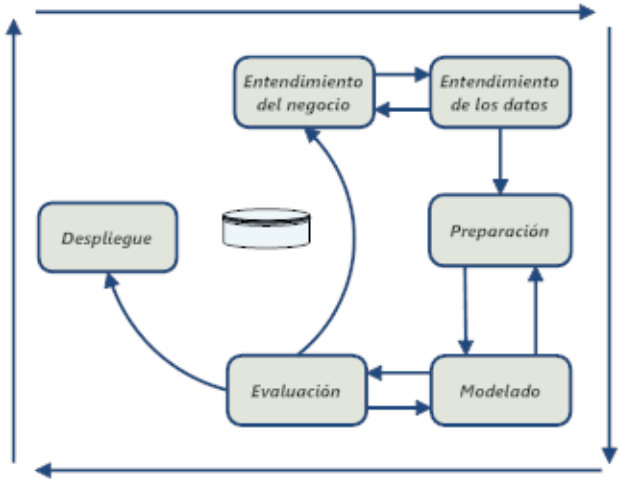


Figura 6. Modelo CRISP-DM

8. Análisis de datos

8.1 Estadística

Según (Wackerly, Dennis y otros, 2008) basándose en definiciones de varios autores, la estadística es:

"[…] una teoría de información, siendo la inferencia su objetivo."

En la estadística existen reglas de estimación expresadas en ecuaciones matemáticas e indican medidas de tendencia central como la media y medidas de dispersión como la varianza o la desviación estándar. Para diferenciar las medidas de las muestras y de la población existen notaciones especiales. Las siguientes ecuaciones indican las diferencias descritas para las medidas estadísticas.

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ ecuación (1)
Media de una muestra n respuestas $y_1, y_2, y_3 \ldots y_n$

$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i$ ecuación (2)
Media poblacional

$S^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ ecuación (3)
Varianza de una muestra de mediciones $y_1, y_2, y_3 \ldots y_n$

$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$ ecuación (4)
Varianza poblacional

$s = \sqrt{s^2}$ ecuación (5)
Desviación estándar de una muestra

$\sigma = \sqrt{\sigma^2}$ ecuación (6)
Desviación estándar poblacional

8.2 Estadística inferencial

Los objetivos de la estadística son realizar inferencias acerca de los parámetros poblacionales, estas inferencias se interpretan de dos formas: como estimaciones de los parámetros o como pruebas de hipótesis sobre los valores obtenidos.

Las pruebas de hipótesis son semejantes al método científico porque se observa la naturaleza, se formula una teoría y por último se confronta con los datos analizados, en este contexto, el científico de datos plantea una hipótesis respecto a uno o varios parámetros poblacionales: de que son iguales a parámetros especificados. Enseguida se toma una muestra de la población y se comprueba con la hipótesis. Si la hipótesis no concuerda con las observaciones entonces se rechaza. De lo contrario se concluye la validez de la hipótesis o que la muestra no detectó los valores reales de los hipotéticos.

8.2.1 Elementos de una prueba de hipótesis

Para probar que la teoría es cierta se tendría que comprobar que lo contrario a lo que se afirma no presenta las pruebas suficientes para ser verdadero, en consecuencia, se afirma que la teoría es válida.

Los elementos concretos de la prueba de hipótesis son los siguientes:

1. Hipótesis nula H_0 (Hipótesis a ser probada, por ejemplo p=0.5)
2. Hipótesis alternativa H_a (Hipótesis a ser aceptada en caso de que la H_0 sea rechazada, por ejemplo $H_a > 0.5$)
3. Estadístico de prueba (Es una función de las mediciones muéstrales en las que se basará la decisión estadística)
4. Región de rechazo (Especifica los valores del estadístico de prueba para el cual la H_0 ha de ser rechazada en favor de H_a)

Hay muchos estadísticos de prueba pero en esta ocasión se examinará el de tipo t para analizar dos muestras debido a que puede representar una ventaja para reducir los costos de las encuestas u obtención de datos.

Para comparar dos muestras poblacionales el siguiente cuadro es de utilidad

Supongase dos muestras pequeñas independientes tal que $\sigma_1^2=\sigma_2^2$

$H_0: \mu_1 - \mu_2 = R_0$

$$H_a: \begin{cases} \mu_1 - \mu_2 > R_0 \ (alternativa\ de\ cola\ superior) \\ \mu_1 - \mu_2 < R_0 \ (alternativa\ de\ cola\ inferior) \\ \mu_1 - \mu_2 \neq R_0 \ (alternativa\ de\ dos\ colas) \end{cases}$$

Estadístico de prueba: $T = \frac{\overline{Y_1} - \overline{Y_2} - R_0}{S_p\sqrt{\frac{1}{n_1} - \frac{1}{n_2}}}$; $ecuación\ (7)\ \ S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\ ecuación\ (8)$

Región de rechazo: $\begin{cases} t > t_\alpha \quad (Región\ de\ rechazo\ de\ cola\ superior) \\ t < -t_\alpha \quad (Región\ de\ rechazo\ de\ cola\ inferior) \\ |t| > t_{\frac{\alpha}{2}} \quad (Región\ de\ rechazo\ de\ dos\ colas) \end{cases}$

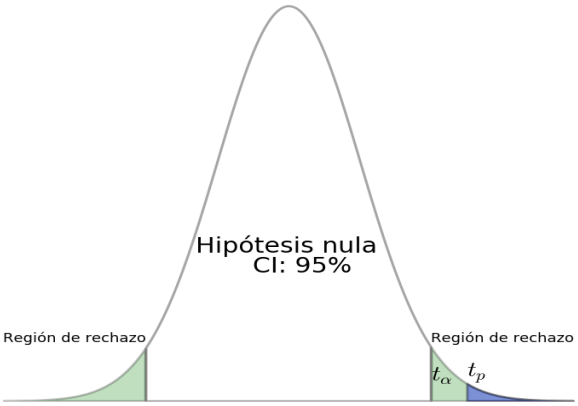


Figura 7. Ejemplo de la distribución t y regiones de rechazo[7]

Las áreas coloreadas en verde de la figura 7 representan las regiones de rechazo, sí las observaciones demuestran que la probabilidad de la variable analizada cae dentro de la región de rechazo entonces se rechaza la hipótesis nula porque no hay pruebas suficientes que demuestren lo contrario. La región en azul representa el área bajo la curva hasta donde indique el estadístico t de la prueba, a su vez representa la probabilidad de la prueba.

8.2.2 Errores

Las decisiones empresariales son muy importantes porque pueden beneficiar o perjudicar el negocio y el prestigio puede estar en juego. Por esta razón el contraste de hipótesis introduce los tipos de errores que se pueden cometer en el momento aplicar pruebas.

Se comete un error tipo I si H_0 es rechazada cuando H_0 es verdadera. La probabilidad de un error tipo I está denotada por α. El valor de α se denomina nivel de la prueba.

Se comete un error Tipo II si H_0 es aceptada cuando Ha es verdadera. La probabilidad de un error tipo II está denotada por β.

[7] Imagen generada en Matplotlib y LATEX

8.3 Probabilidad condicional

El teorema de Bayes se refiere a la probabilidad condicional y expresa la probabilidad de eventos bajo el dominio de un evento condicionante. La relación entre la probabilidad condicionada y la probabilidad total se expresa en la ecuación (9).

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)} \quad \text{ecuación (9)}$$

En otras palabras, la probabilidad total expresa la probabilidad de que ocurra un evento (A) condicionado por otro evento particular (B) dividido por todo el dominio del evento (A) dada cada instancia del dominio (B). De lo anterior se sigue: en el numerador se anotan los casos de probabilidad condicionada de un evento particular de (B) y, en el denominador, los casos de (A) bajo todos los casos de (B) con lo que se mantiene la simetría del número de casos de un evento comparado con el total mediante la razón de números propuesta en la teoría de la probabilidad.

9 Análisis de texto

La lingüística computacional (conformada por la Inteligencia Artificial y la lingüística) se encarga de buscar, comparar y manejar conocimientos expresados en lenguaje natural de forma escrita. El área encargada de este estudio es la Minería de texto, al igual que la Minería de datos, el término de "análisis de texto" está empezando a tomar más auge entre los analistas de datos en la actualidad.

Para resolver el problema de la lingüística computacional, se utiliza un procesador lingüístico conformado por tres módulos, a saber:

-Modulo morfológico: reconoce palabras, añade las marcas de tiempo, género y número.

-Modulo sintáctico: reconoce las relaciones explicitas en una oración, añade marcas gramaticales a las palabras.

-Modulo semántico: Reconoce la estructura completa del texto y crea una red semántica.

En la computación existe una disciplina o ciencia de la ingeniería que estudia el lenguaje formal involucrado en los sistemas de cómputo, por ello se introdujeron algunos conceptos de esta teoría generada por Noam Chomsky en este trabajo.

9.1 Lenguajes y operaciones

Definido sobre un alfabeto (conjunto finito de caracteres) es el conjunto de todas las palabras que se pueden construir con las letras de dicho alfabeto. Se denota por ω (Σ).

El lenguaje universal de cualquier alfabeto es infinito, y siempre pertenece a él la palabra vacía.

Ejemplo: si Σ = {a} entonces ω (Σ) = {λ, a, aa, aaa,...}

Las operaciones sobre los lenguajes son de unión (sigue las propiedades de la unión en teoría de conjuntos), concatenación, clausura (o cierre de Kleene) y clausura positiva.

9.2 Palabras y operaciones

Las palabras en este contexto significan una secuencia finita de caracteres de un alfabeto. El conjunto de palabras conforma un lenguaje y naturalmente tienen una longitud que indica el número de caracteres que forman cada palabra de un lenguaje. Al igual que los lenguajes, en las palabras se pueden realizar operaciones, las cuáles se describen a continuación:

-Concatenación: Es poner los caracteres de una palabra seguidos de los caracteres de otra palabra conservando su orden original; contiene las propiedades de operación cerrada, elemento neutro y no conmutativa.

-Potencia i-esima: Consiste en concatenar una palabra consigo misma i veces.

-Reflexión o inversa: Consiste en formar una palabra con los caracteres dispuestos en orden inverso.

9.3 Gramáticas

La gramática debe guiar la estructura en la que se ponen las palabras para describir la estructura de las frases de un lenguaje. Por ejemplo en la frase "*el perro corre deprisa*" se siguen las siguientes reglas:

Reglas gramaticales:
1. <sentencia> ::= <sujeto> <predicado>
2. <sujeto> ::= <articulo> <nombre>
3. <predicado> ::= <verbo> <complemento>
4. <predicado> ::= <verbo>
5. <articulo> ::= el
6. <nombre> ::= perro

7. <verbo> ::= corre
8. <verbo> ::= come
9. <complemento> ::= deprisa
10. <complemento> ::= mucho

Las reglas anteriores pueden considerarse como las reglas de producción y es habitual que se representen mediante un árbol de derivación.

De manera formal la definición de una gramática se realiza mediante conjuntos que representan el alfabeto, el conjunto de símbolos terminales, los no terminales y el conjunto de las producciones. Se clasifican en cuatro grupos según Noam Chomsky.

-Gramáticas tipo 0 (recursivamente numerables)
-Gramáticas tipo 1 (dependientes del contexto)
-Gramáticas tipo 2 (independientes del contexto)
-Gramáticas tipo 3 (regulares)

9.4 Expresiones Regulares

Una expresión regular es la notación normalizada para representar un lenguaje generado por una gramática regular o de tipo 3.

Para definir una expresión regular se utilizan los símbolos del alfabeto, los símbolos de cadena vacía (λ o ε), conjunto vacío y las operaciones de unión, concatenación, cierre de Kleene, cerradura positiva y los paréntesis que dan prioridad a los grupos de símbolos de los que se conforma la expresión regular.

	Descripción	Ejemplo	Lenguaje generado
.	Cualquier caracter	.	a, b, .
*****	Repite cero o más veces el grupo precendente	.*	a, ab, abab, ε
^	Inicio de cadena	^b.*	b, baaaaaaaa
$	Final de cadena	b.*b$	b, baaaab, bab
[ ]	Conicide con cualquier caracter del grupo	[a-cz]	a, b, c, z
[^]	Coincide excepto por cualquier caracter del grupo	[^ a]	b, c, 1, 2
( )	Subexpresión capturada	(a.*)	a, abb
{m, n}	Coincide almenos m veces y máximo n veces el grupo precedente	a{2,4}	aa, aaa, aaaa
|	Alternación or, uno u otro	a|b	a, b
+	Uno o más del grupo precedente	a+	a, aa, aaa,aaaaaaaa
?	Cero o una vez	a?	ε,a
\d	Digito	\d	1, 5, 0
\D	No digito	\D	a, b, (
\s	Espacio blanco		
\S	No espacio blanco		
\w	Palabra de caracteres		
\W	No palabra de caracteres		

Tabla 2. Notación de metacaracteres de la expresión regular

La Minería de texto tiene por objeto clasificar palabras, encontrar relaciones entre las categorías y describir estadísticamente el conjunto de palabras tanto de textos estructurados como de textos no estructurados. Para plantear un modelo de asociación entre las palabras es imprescindible tener como recurso un corpus; que es un repositorio de texto organizado en renglones, es decir, cada conjunto de palabras, documentos o categorías correspondientes de un documento, se almacenan en cada renglón de una tabla para su análisis.

9.5 Tareas del procesamiento de textos

Dentro de las tareas podemos indicar la búsqueda o extracción de la información, la preparación del texto y la Minería de texto. En la preparación del texto se busca corregir ortográficamente y gramaticalmente las palabras separadas. En la Minería de texto se busca categorizar, clasificar, agrupar, asociar, detectar asociaciones, detectar tendencias y generar resúmenes del texto.

9.6 Grafos conceptuales

Los grafos conceptuales (Montés y Gómez) citando a (Sowa, 1984): son un sistema de lógica orientado a la representación de la semántica del lenguaje natural. Básicamente, un grafo conceptual es un grafo bipartito que tiene dos tipos diferentes de nodos: conceptos y relaciones conceptuales.

Los conceptos representan sujetos, verbos o adjetivos, tienen tipo conceptual y referente. En cuanto a las relaciones se caracterizan por el tipo de relación conceptual y la valencia; la primera indica cual es la relación entre dos conceptos y la segunda indica cuantas relaciones existen.

10 Python

Es un lenguaje multiparadigma e interpretado, integra paquetes o bibliotecas de C/C++ o FORTRAN para realizar operaciones de álgebra lineal o transformaciones de Fourier entre otros algoritmos. Python es ideal para cómputo científico intensivo, no es ideal para aplicaciones altamente concurrentes o de multihilos. La razón por la que no es conveniente utilizar Python en el tipo de aplicaciones mencionadas es debido al mecanismo de bloqueo del interprete global (GIL, global interpreter lock) que no permite la ejecución "al mismo tiempo" de bytecode en el CPU.

Aunque es deseable realizar el procesamiento en paralelo para aplicaciones en Big Data con Python se tendría una limitante, para subsanar este problema existe un proyecto llamado Cython que integra OpenMPI, un framework para el cómputo paralelo.

De los paquetes actuales para uso científico se encuentran:

Numpy: Objetos de arreglos multidimensionales, transformadas de Fourier, integración con C, FORTRAN.

Pandas: Contiene estructuras de datos y funciones diseñadas para trabajar con datos rápidamente.
Matplotlib: Lenguaje procedural para crear imágenes estadísticas y científicas, integra LATEX.

IPython: Provee un robusto ambiente computacional interactivo y exploratorio diseñado para codificar, depurar y probar programas.

Scipy: Es una colección de paquetes que permiten trabajar con ecuaciones diferenciales, álgebra lineal, procesamiento de señales y uso de código C++ para acelerar el procesamiento de arreglos.

El ambiente interactivo de Jupyter Notebook permite codificar eficientemente mediante una página web ya que integra IPython y muchas otras funciones para administrar los artefactos de software.

Capítulo II Pruebas ágiles

1. Ambientes virtuales para desarrollo

La primera asignación en el servicio social fue investigar la herramienta Vagrant y Virtual Box para crear y gestionar ambientes de desarrollo. La creación de ambientes de desarrollo permite tener muy controlado el desarrollo de software, mediante las herramientas investigadas se entendió la importancia de las mismas y el beneficio para integrarlas en metodologías ágiles.

Vagrant es una herramienta de línea de comandos que permite configurar rápidamente las bases de datos, bibliotecas y servicios dentro del ambiente para el desarrollo de una aplicación. Para utilizar Vagrant se necesita de un ambiente que permita utilizar los recursos virtualizados como Virtual Box, también es necesaria una herramienta para emular comandos del sistema operativo Linux en el sistema operativo Windows como Cygwin que permite la ejecución de los comandos ls, pwd, ssh, g++[8], gcc etc. En el anexo A esta disponible la documentación (entregada al líder técnico) necesaria para la instalación y configuración de las herramientas mencionadas.

2. Aplicación blog en cup cake php

Con el fin de familiarizarse con el framework cup cake php se investigó documentación escrita en inglés por estar más completa, se desarrolló la aplicación "blog" mediante el patrón de diseño Modelo-Vista-Controlador en el que se basa el framework, en el anexo B se encuentra la documentación de instalación y configuración de cup cake php. Al desarrollar la aplicación se encontraron problemas técnicos como incompatibilidades entre clases de las versiones del framework cup cake php y configuraciones con la base de datos implementada en una máquina virtual Linux.

En la figura 8 se muestra una captura de pantalla en la que se muestran algunos errores en la instalación del framework, el trabajo sobre la máquina virtual y la generación de documentación técnica.

[8] Mediante esta herramienta la compilación de programas escritos en lenguaje C y C++ es muy útil cuando se trabaja con línea de comandos; cabe señalar que requiere instalar los paquetes gcc core y g++ con compatibilidad entre versiones.

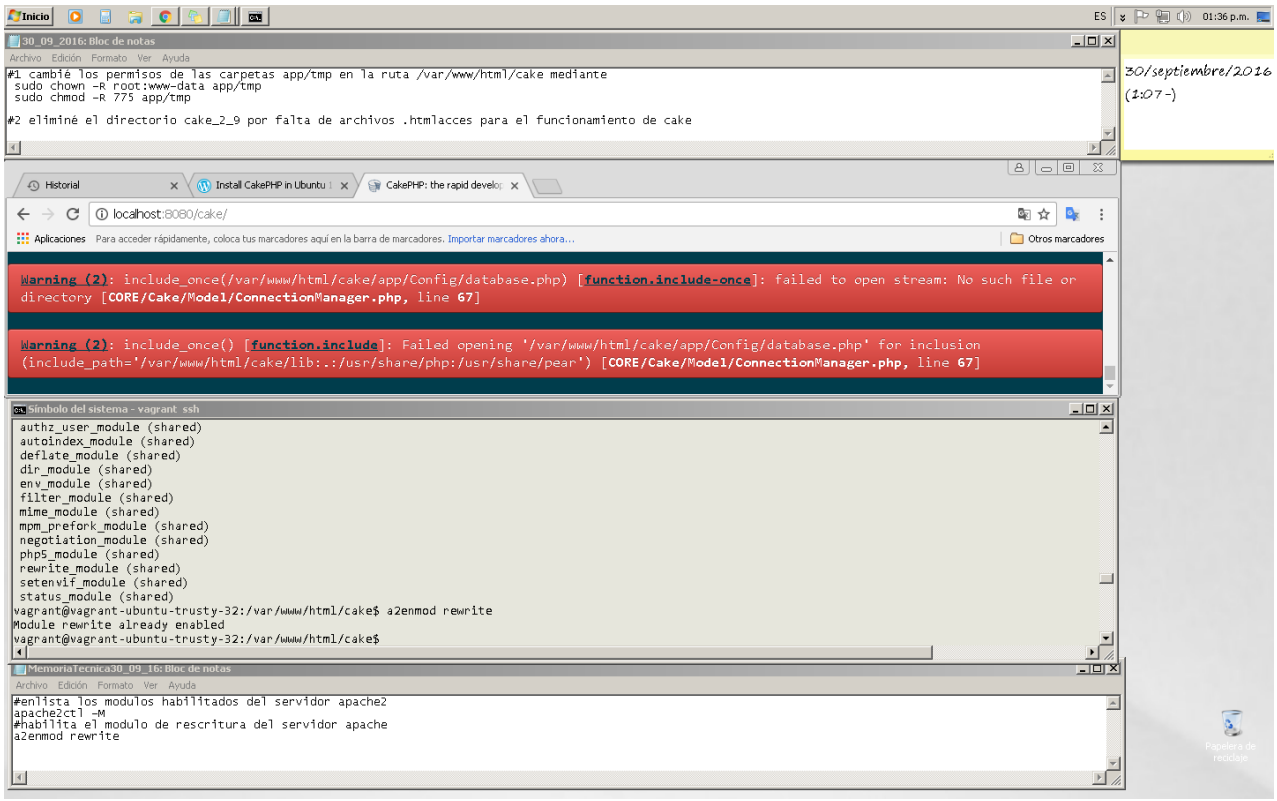


Figura 8. Ambiente de trabajo en la instalación de cup cake php

La aplicación “blog” es un ejemplo de desarrollo en la documentación de cup cake y consiste en la implementación de operaciones básicas sobre registros mediante vistas, a saber: agregar, leer, editar y eliminar tuplas. A continuación se presentan los pasos y problemas que se presentaron en la recreación de la aplicación blog.

1. Cronstruí el modelo Post en el directorio /cake_2_5_2/app/model que hereda de la clase appModel.

2. Construí el controlador Post en el directorio /cake_2_5_2/app/Controller que hereda de la clase AppController-

3. Configuré el controlador PostsController.php para visualizar el contenido de la tabla posts de la base de datos BlogCakePHP en forma HTML mediante navegador web

4.1. Construí la vista /var/www/html/cake_2_5_2/app/View/Posts/index.ctp con el código php que obtiene los datos a partir del controlador.

4.2. Construí la vista “view” en el directorio /var/www/html/cake_2_5_2/app/View/Posts/view.ctp para visualizar un post en partícular.

5. Agregué la funcionalidad de agregar, editar o borrar un post a la aplicación Posts
bajo las consideraciones en las funciones respectivas en el controlador PostsControler

La clase Flashccontroller cambió por SessionController porque la clase no está soportada en la versión 2.5 (version 2.7 incorpora FlashController)

La clase FlashHelper cambió por SessionHelper por la misma razón

6. Modifiqué la regla de restricción notBlank por notEmpty en la verificación de agregar un post

7. Agregué la vista /var/www/html/cake_2_5_2/app/View/Posts/add.ctp para mostrar el formulario de agregar un post mediante código php.

8. Edité la vista index.ctp para que mostrará las nuevas funcionalidades.

Finalmente, el aspecto de la aplicación se muestra en las figuras 9 y 10, se probó manualmente la funcionalidad de las operaciones (pruebas unitarias); sin embargo, cuando se trata de presentar un informe al equipo de desarrollo o a los clientes para validación, se requiere un reporte del resultado de la prueba, para ello se investigó la herramienta Webtest Canoo por las características de presentación de informes y por ser una herramienta ligera para reportar pruebas de rendimiento, de regresión y de carga en las aplicaciones web.

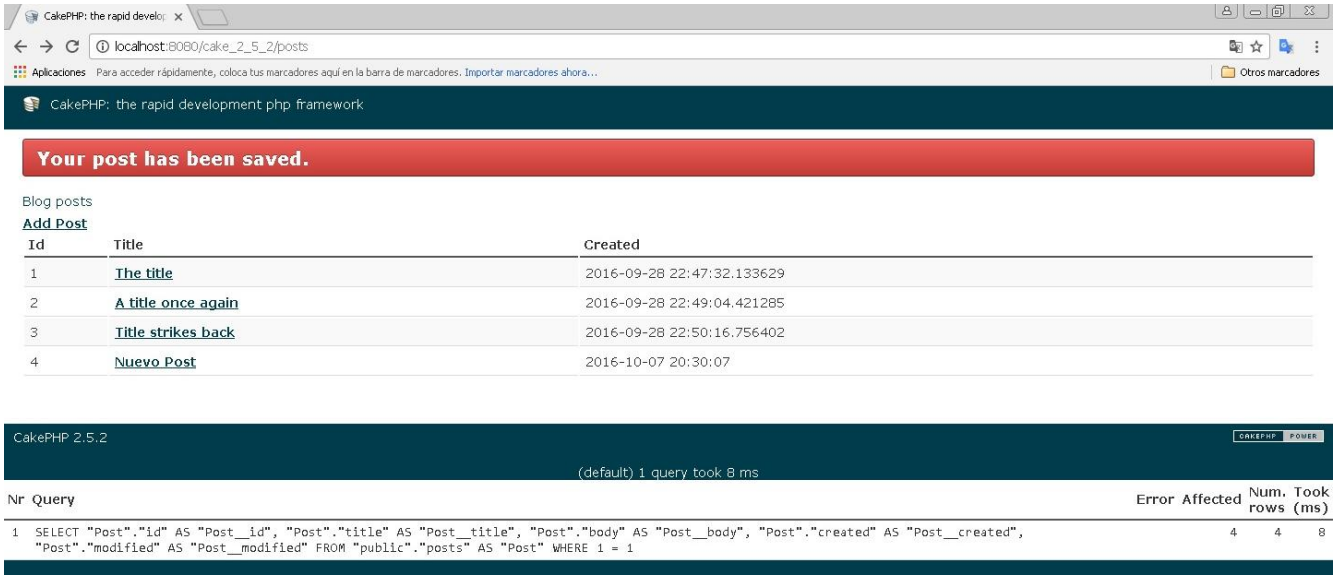


Figura 9. Aspecto de la aplicación “blog”

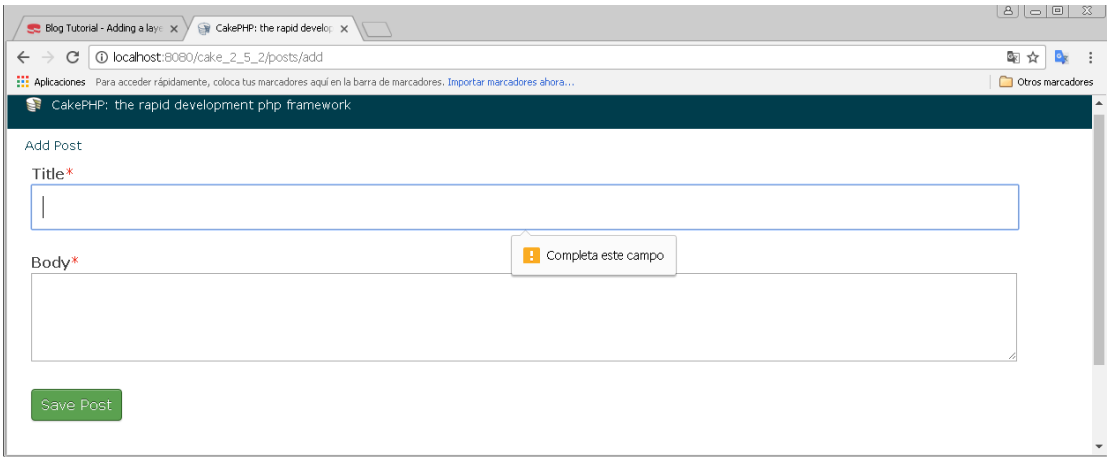


Figura 10. Aspecto de la vista add

Existen técnicas para dirigir el desarrollo de las aplicaciones, en este caso se siguió una metodología tradicional por ser una aplicación muy sencilla, esta investigación permitió descubrir el alcance de una herramienta de prueba en metodologías ágiles y como pueden integrarse en el proceso de desarrollo de software[9].

El primer acercamiento a la documentación de la herramienta Webtest Canoo permitió conocer el soporte técnico por parte de los distribuidores y por parte de la comunidad de usuarios. De lo anterior se concluyó que era muy poca la documentación y los ejemplos, pero por sus características, valía la pena investigar las funcionalidades de Webtest Canoo, además es una herramienta con poco soporte. La investigación en gran parte fue por ensayo y error.

3. Pruebas con Webtest Canoo

3.1 Diseño de la prueba

La prueba se diseñó para probar conjuntamente (pruebas de integración) las funciones de agregar, editar, borrar, y leer posts de la aplicación, por lo que se planteó qué elementos de la aplicación serían los objetivos para el programa de prueba, se realizaron las modificaciones necesarias en la aplicación para que la prueba operara correctamente (creación de identificadores en el código HTML entre otras). En la figura 11 se observa un esquema de la aplicación que permitió organizar la prueba e identificar los elementos que se modificaron de acuerdo a la documentación de Webtest Canoo. Cabe señalar que la herramienta de Webtest Canoo opera a nivel de presentación de la arquitectura de la aplicación para simular el comportamiento de un usuario, esta

[9] Existe la técnica de desarrollo dirigida por las pruebas, en la que se diseña la prueba y después se desarrolla el software.

investigación también tuvo la intención de plantear pruebas de integridad en los datos para sistemas concurrentes, de carga, regresión y rendimiento[10].

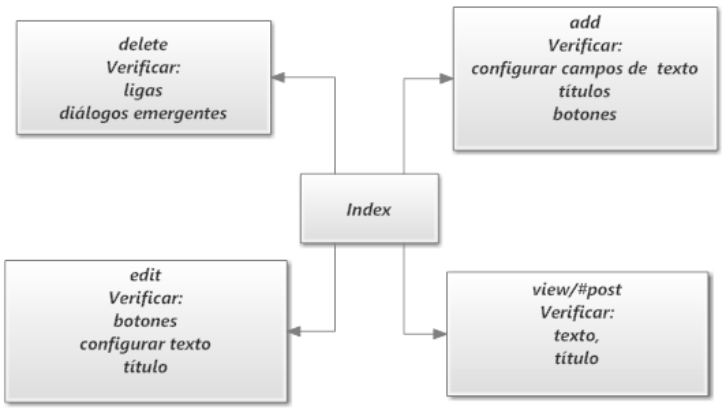


Figura 11. Elementos a probar en cada vista

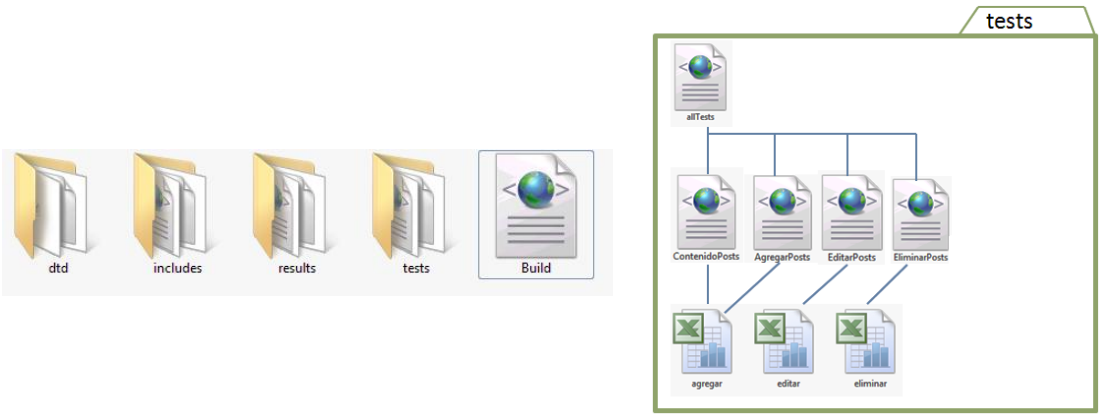
3.2 Ejecución

La prueba se estructuró en diferentes archivos XML sobre directorios de manera jerárquica; se hizo de esa forma para reutilizar código y no saturar la memoria en tiempo de ejecución, estos archivos XML toman como recursos de entrada archivos excel que contienen los registros de los posts; se realizaron pruebas dinámicas (operaciones de agregar, actualizar y eliminar) y estáticas (validar el contenido y presencia de etiquetas).

Debido a que Webtest Canoo se basa en HtmlUnit se puede emular un navegador web a través de la aplicación, esta herramienta también integra un constructor de proyectos dirigido con estructuras y archivos XML basado en ANT, este constructor se encarga de dirigir la ejecución de la prueba y generar el reporte de resultados en una página HTML.

A continuación se muestra en la figura 12 la estructura del proyecto de prueba desarrollado y en el apartado C parte de la codificación.

[10] Se tuvo la experiencia de realizar pruebas de software en una aplicación web para el INAH a nivel de la capa de presentación de forma manual con los compañeros en la División de Colaboración y Vinculación. Esta prueba tenía por objetivo verificar la integridad de los datos en procesos concurrentes.



a) b)

Figura 12. a) Directorio del proyecto de prueba “nuevo Test” b) Dependencia de archivos XML

En la carpeta “includes” se agregaron (en archivos XML) variables para dirigir la prueba, como cambiar entre las vistas de la aplicación.

3.3 Resultados

Al ejecutar el constructor de proyectos basado en ANT de Webtest Canoo se aplicó la prueba mediante el archivo Build con extensión XML y se generó automaticamente el reporte de resultados en páginas web, estos resultados estan disponibles en el directorio *results* del proyecto.

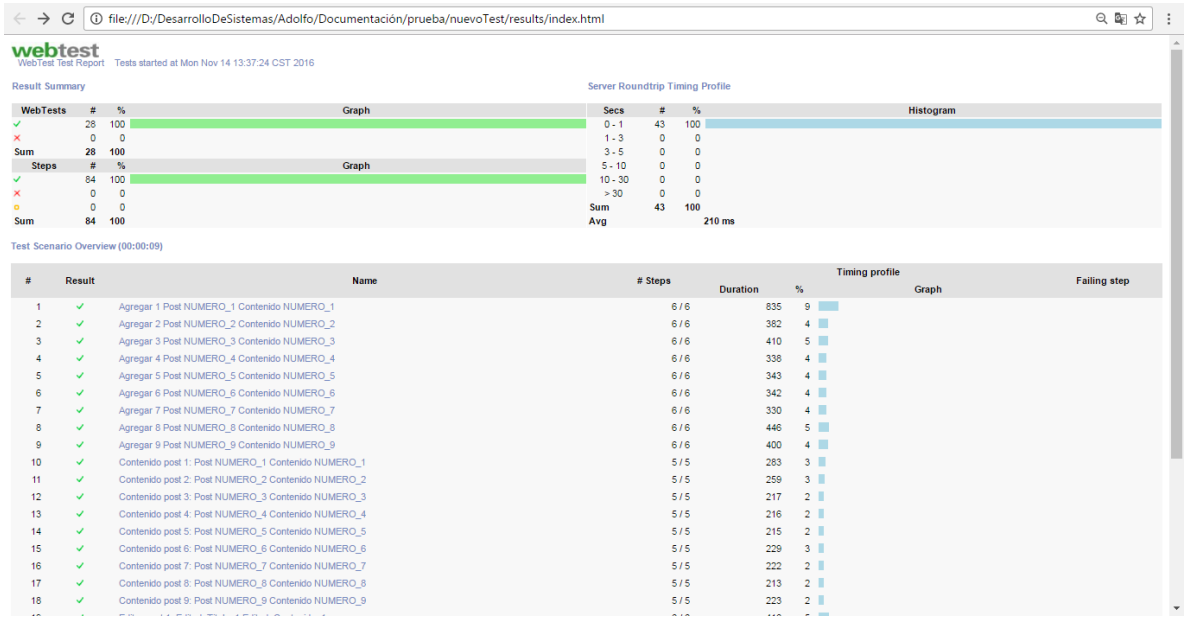


Figura 13. Estadísticas del reporte de las pruebas

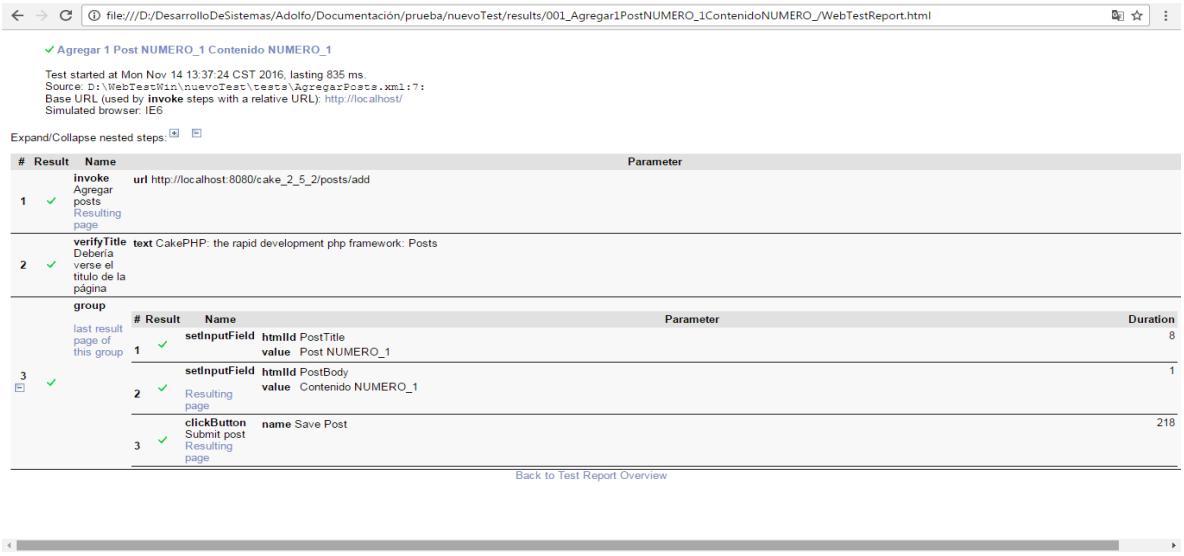


Figura 14. Detalle de algunos casos de prueba

Al ejecutar la prueba el aspecto de la aplicación cambió como se muestra en la siguiente ilustración.

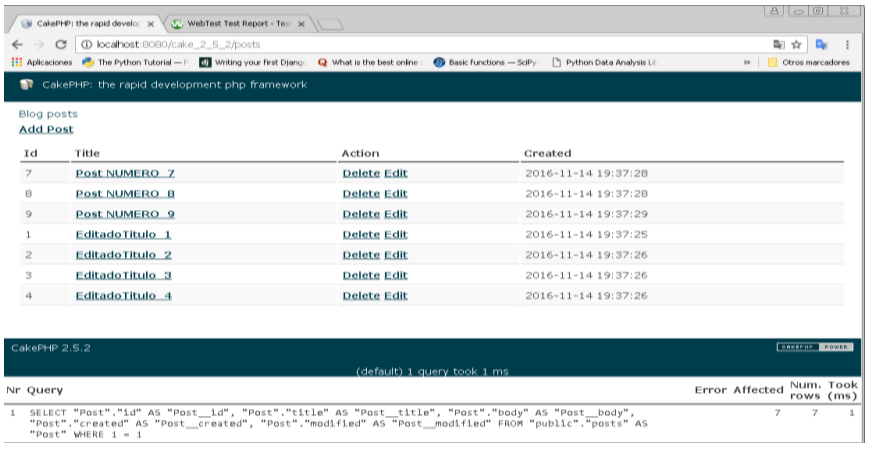


Figura 15. Aspecto de la vista posts después de la prueba

Por último, también se investigó el uso de esta herramienta a través del lenguaje Groovy pero resultó muy complicado utilizar recursos como archivos excel mediante un controlador de datos por lo que no se recomienda utilizar Webtest Canoo con Groovy ya que requiere una investigación más profunda con poca documentación[11].

[11] Sólo se entregaron ejemplos de pruebas estáticas mediante lenguaje Groovy

Capítulo III Minería de Datos

1. Análisis de datos

1.1 Tramas de datos

Como parte de la demostración de ejemplos desarrollados en Python, en internet se encontraron recursos para resolver y mostrar las técnicas de programación en este lenguaje[12], los recursos están orientados por el término ciencia de datos o análisis de datos en el que se trabajó con operaciones de álgebra relacional, estadística y probabilidad. Como se ha mencionado en el marco teórico, el término difiere un poco con la Minería de datos pero son disciplinas que van conjuntas para descubrir conocimiento, por esa razón en este trabajo se clasificó como parte de la Minería de datos.

En Python actualmente existen bibliotecas o paquetes orientados a la ciencia de datos, de hecho son muy populares entre los pocos científicos de datos en la actualidad por la simplicidad y potencia del lenguaje.

Conforme se realizó la solución de los problemas planteados el nivel de dificultad aumentó y se mejoró en la técnica de programación.

Los ejemplos requirieron de la habilidad para manipular tramas de datos (tablas) para agrupar, organizar, operar y obtener parámetros estadísticos sobre los archivos con extensión "csv" ubicados en el repositorio local y externos (ubicados en repositorios en internet). Estos ejemplos también exigen un manejo de flujos de datos para extraer los datos de las fuentes, además de habilidad para preparar los datos y obtener los resultados deseados.

Mediante las funciones de lectura de la biblioteca Pandas, se extrajeron tablas de datos de diferentes fuentes como archivos "csv", "excel", bases de datos y de texto en dataframes (tramas de datos) sobre las que se puede operar con álgebra relacional con sentencias de Python.

En la tabla 3 se muestra un ejemplo de la tabla extraída por esta herramienta.

[12] **Brooks, Christopher.** *Introduction to Data Science in Python.* s.l., Michigan, EUA : Coursera.

Out[1]:

	Rank	Documents	Citable documents	Citations	Self-citations	Citations per document	H index	Energy Supply	Energy Supply per Capita	% Renewable	1996
Country											
China	1	127050	126767	597237	411683	4.70	138	1.271910e+11	93.0	19.754910	1.617630e+12
Japan	3	30504	30287	223024	61554	7.31	134	1.898400e+10	149.0	10.232820	5.012733e+12
Russian Federation	5	18534	18301	34266	12422	1.85	57	3.070900e+10	214.0	17.288680	8.466689e+11
Canada	6	17899	17620	215003	40930	12.01	149	1.043100e+10	296.0	61.945430	1.120541e+12
Germany	7	17027	16831	140566	27426	8.26	126	1.326100e+10	165.0	17.901530	2.864379e+12
India	8	15005	14841	128763	37209	8.58	115	3.319500e+10	26.0	14.969080	6.702362e+11
France	9	13153	12973	130632	28601	9.93	114	1.059700e+10	166.0	17.020280	2.061442e+12
South Korea	10	11983	11923	114675	22595	9.57	104	1.100700e+10	221.0	2.279353	5.908287e+11
Italy	11	10964	10794	111850	26661	10.20	106	6.530000e+09	109.0	33.667230	1.890310e+12
Spain	12	9428	9330	123336	23964	13.08	115	4.923000e+09	106.0	37.968590	9.661629e+11
Australia	14	8831	8725	90765	15606	10.28	107	5.386000e+09	231.0	11.810810	7.145788e+11
Brazil	15	8668	8596	60702	14396	7.00	86	1.214900e+10	59.0	69.648030	1.414180e+12

Tabla 3. Trama de datos sobre suministro energético y documentos en la población de EUA

Se realizaron cálculos, preparación de datos y transformaciones sobre las tablas. En la figura 16 se muestra el enunciado solicitado, el código de una función que calcula el promedio en forma horizontal (sobre campos de tuplas) en la tabla de la figura 16 y despliega los quince primeros lugares ordenados en forma descendente.

Question 3 (6.6%)

What is the average GDP over the last 10 years for each country? *This function should return a Series named avgGDP with 15 countries and their average GDP sorted in descending order.*

```
import numpy as np

def answer_three():
    Top15,lose = answer_one()
    columns = Top15.columns[-10:]                          #proyecta últimas 10 columnas
    df=Top15.apply(lambda x: np.average(x[columns]), axis=1)  # promedio en forma horizontal (columnas )
                                                           #por país con una función anónima
    return df.sort_values(ascending=False)                 # ordena los promedios

answer_three()
```

```
Country
Japan          5.136055e+12
Germany          3.082915e+12
China          2.421490e+12
France          2.326660e+12
Italy         2.041992e+12
Brazil         1.564053e+12
Canada           1.333080e+12
Spain          1.161654e+12
…
dtype: float64
```

Figura 16. Codificación de una función en Python

1.2 Regresión lineal

Además, se demostró la utilidad de la biblioteca Matplotlib para visualizar gráficas de datos; en este caso se muestra la correlación de dos variables y el coeficiente de Pearson calculado.

```
def plot9():
    import matplotlib as plt
    %matplotlib inline

    Top15,lose = answer_one()
    Top15['PopEst'] = Top15['Energy Supply'] / Top15['Energy Supply per Capita']
    Top15['Citable docs per Capita'] = Top15['Citable documents'] / Top15['PopEst']
    Top15.plot(label='rxy(pearson)='+str(answer_nine()),x='Citable docs per Capita',
    y='Energy Supply per Capita', kind='scatter', xlim=[0, 0.0006])
```

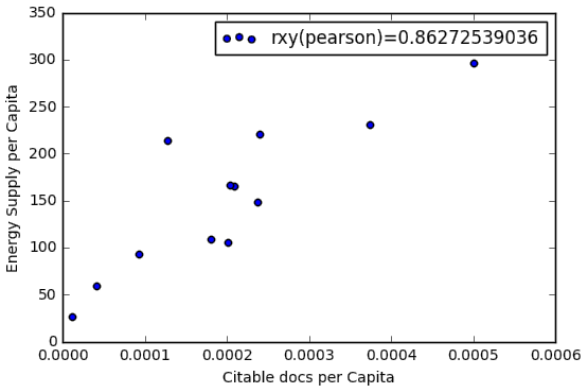


Figura 17. Regresión lineal con coeficiente de Pearson

Como se puede observar, se indagó mediante una regresión lineal la correlación entre el suministro energético per cápita y los documentos citables per cápita de la población estadounidense.

1.3 Discretización de dominio

Se realizaron ejemplos de transformación de datos mediante operaciones de álgebra relacional, de cálculo y discretización de dominios como muestra la figura 18, en la que se mezcla una tabla con información de continentes con otra sobre países que contiene datos discretizados, finalmente se agrupan por continente y por grupos agregando la cuenta de países por cada índice.

Question 12 (6.6%)

Cut % Renewable into 5 bins. Group Top15 by the Continent, as well as this new % Renewable (Renovable) bins. How many countries are in each of these groups?
*This function should return a **Series** with a MultiIndex of Continent, then the bins for % Renewable. Do not include groups with no countries.*

```
def answer_twelve():
    df  = continent_DF()
    #df.set_index(['Continent','Country'],inplace=True)
    cut = pd.DataFrame.from_dict(pd.cut(df['% Renewable'],5).to_dict(),orient='index')
    cut.rename(index=str, columns={0:'% Renewable Categorical'},inplace=True)
    df  = df[['Continent','Country','% Renewable']]
    df.reset_index(inplace=True)
    cut.reset_index(inplace=True)
    df2 = pd.merge(df,cut,how='inner',left_index=True, right_index=True)
    df2 = df2[['Continent','Country','% Renewable Categorical']]
    return df2.groupby(['Continent','% Renewable Categorical']).count()

answer_twelve()
```

Out[15]:

		Country
Continent	**% Renewable Categorical**	
Asia	**(15.753, 29.227]**	1
	(2.212, 15.753]	3
Australia	**(2.212, 15.753]**	1
Europe	**(15.753, 29.227]**	3
	(29.227, 42.701]	2
North America	**(56.174, 69.648]**	1
South America	**(56.174, 69.648]**	1

Figura 18. Discretización, mezcla y agrupamiento sobre tramas de datos

1.4 Prueba de Hipótesis

El análisis más sofisticado realizado mediante programación en Python fue el contraste de hipótesis con distribuciones de probabilidad "t". La distribución t en el método se empleó porque la población se partió en dos muestras y una muestra tenía cardinalidad más pequeña en comparación con la otra, era necesario comparar las dos muestras para saber si tenían la misma propiedad descrita en la asignación y establecida en la hipótesis nula.

1.4.1 Planteamiento del problema

Definiciones
Un cuarto especifica un periodo de tres meses, Q1 de Enero A Marzo, Q2 de Abril a Junio, Q3 de Julio a Septiembre y Q4 de Octubre a Diciembre.

Una Recesión se define en su comienzo con la declinación de dos cuartos consecutivos del GDP[13] y finaliza con el crecimiento de dos cuartos consecutivos de GDP.

Un pivote de Recesión es el cuarto con el valor más bajo dentro de la recesión.

Una ciudad universitaria es la ciudad que presenta un alto porcentaje de estudiantes universitarios comparado con el resto de la población de la ciudad.

Hipótesis: Las ciudades universitarias tienen, en promedio, el precio de sus casas menos afectadas por la recesión.
Ejecuta la prueba–t para comparar la razón de los precios un cuarto antes de la recesión con el pivote de la recesión.
(price_ratio=quarter_before_recession/recession_bottom)

1.4.1 Desarrollo

El diagrama de la figura 19 indica la serie de pasos seguidos y marcados con las etapas de pre-procesamiento de la Minería de datos para realizar la prueba-t. Las etapas de limpieza, transformación e integración se agruparon en la notación porque son etapas no excluyentes, es decir, dado que para realizar limpieza, pueden transformarse los datos mediante promedios o incluir nuevas columnas de otras fuentes de datos.

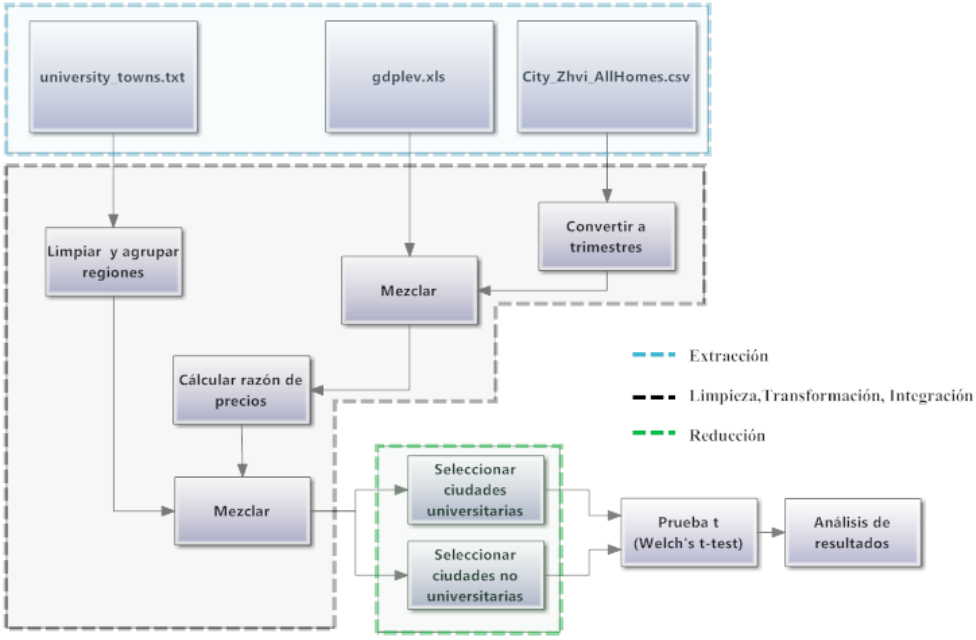


Figura 19. Proceso de trabajo para el ejemplo de contraste de hipótesis

[13] Gross Domestic Product o similar a Producto Interno Bruto

En las etapas de pre-procesamiento se tomaron en cuenta los valores nulos porque en la etapa de análisis no afectan estos valores; debido a la propiedad de la función ttest_ind para el manejo de valores nulos. En la figura 20 se presenta una porción del código que llama a otras funciones encargadas del pre-procesamiento de las tablas y la etapa en la que se aplica la prueba-t. Este fragmento de código representa una función en Python que devuelve los parámetros de interés para el análisis de resultados.

```
def run_ttest():
#1 muestra aumento o declinación en los precios de casas durante la recesión
   df  = convert_housing_data_to_quarters()      #Dataframe con los precios de casas agrupados State,RegionName y
                                                 #cuartos de año
   indice =  df.columns.get_loc(get_recession_start())-1    #obtiene la columna que representa el trimestre anterior al
                                                            #inicio de la recesión
   df_forRatio =  df.iloc[:,indice]
   df                =  df.loc[:,get_recession_start():get_recession_bottom()]    #selección de columnas de interés
   df['price_ratio'] = df_forRatio/df.iloc[:,-1]                                  #calcula la razón en los precios

  #2 Universidades más pobladas vs Universidades menos pobladas
   dfU = get_list_of_university_towns()

   df3=pd.merge(dfU,df,left_index=True,right_index=True,how='inner')

   dftown      =  df[df.index.isin(df3.index)]       #Conjunto de ciudades universitarias con trimestres
                                                     #[inicio recesión hasta pivote recesión]
   dfuntown  = df[~df.index.isin(df3.index)]     #Conjunto de ciudades no universitarias con trimestres
                                                 # [inicio recesión hasta pivote recesión]

   Ttest   = ttest_ind(dftown['price_ratio'], dfuntown['price_ratio'],nan_policy ='omit',axis=0, equal_var=True)
   Ttest2 = ttest_ind(dftown['price_ratio'], dfuntown['price_ratio'],nan_policy ='omit',axis=0, equal_var=False)

   return Ttest, Ttest2, dftown['price_ratio'].std(), dfuntown['price_ratio'].std(), dftown['price_ratio'].mean(), dfuntown['price_ratio'].mean()
```

Figura 20. Prueba de hipótesis en Python

1.4.3 Resultados

Al ejecutas la función se obtienen los siguientes parámetros:

```
(Ttest_indResult(statistic=-2.2380580605001423,
                 pvalue=0.025239520749924828,
 Ttest_indResult(statistic=-2.7650724862877167,
                 pvalue=0.0061666659307329013),
 0.10860372721298293,
 0.13548511961054835,
 1.0757025525810295,
 1.0967731949132407)
```

Puesto los datos proceden de la misma tabla se considera que la varianza es igual y se sigue el siguiente razonamiento:

Se refuta el planteamiento de la hipótesis por la naturaleza de la función ttest_ind[14] que indaga la igualdad de las dos muestras. Se plantea lo siguiente:

H_0: Promedio de precios en ciudades universitarias = Promedio de precios en ciudades no universi -tarias.

H_a: Promedio de precios en ciudades universitarias < Promedio de precios en ciudades no universitarias.

Si p=0.025 y α=0.010 entonces H_0 = Verdadera, dftown.mean = dfuntown.mean

Si p=0.025 y α=0.05 entonces (No hay suficientes pruebas para afirmar que son iguales las razones de precios promedios) y en el mejor de los casos: dftown.mean < dfuntown.mean

Se concluyó que los datos presentan la suficiente evidencia para afirmar que las ciudades universitarias son igualmente afectadas por la recesión con un intervalo de confianza del 99% dados los requerimientos pero en el mejor de los casos apoyándose en la estimación de la media y la dispersión de la desviación estándar se concluyó que son menos afectadas las ciudades universitarias (lo que equivale a una reducción en la pérdida del mercado). Sin embargo, para un valor p igual a 0.05 la hipótesis nula no presenta suficiente evidencia puesto que el intervalo de confianza se redujo.

La prueba fue de cola inferior al presentarse una desigualdad (ver figura 21).

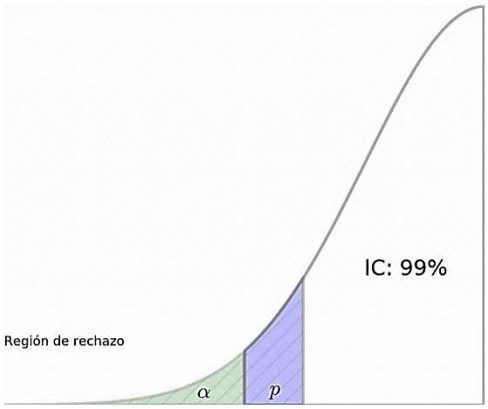


Figura 21. Prueba de cola inferior

El mejor caso se presenta en las ciudades universitarias donde el precio medio de la razón calculada es ligeramente menor.

[14] En la documentación de Scipy se indica que si p tiende a "uno", entonces las muestras son significativamente iguales. En caso contrario, son desiguales dependiendo del nivel de prueba, siendo los más comunes: "0.01" y "0.05".

Este ejemplo permitió interpretar el análisis estadístico y los intervalos de confianza, aunque existen dudas por la afirmación realizada debido al cambio del intervalo de confianza que es sensible a la presencia de valores atípicos, se quedó con el mejor caso.

2. Diagnóstico de la base de datos RUA

2.1 Fase de entendimiento del negocio y de los datos

Para analizar la base de datos de la RUA (Red Universitaria de Aprendizaje) de la UNAM primero se inició la etapa de entendimiento del negocio y el entendimiento de los datos, esto se logró mediante el diccionario de datos, el Modelo Entidad-Relación y de programas de computadora.

Para reportar el entendimiento de los datos se utilizaron elementos de estadística descriptiva como histogramas, estimaciones y rangos. En la figura 22 se presenta el Modelo Entidad-Relación en el que se indican las tablas descartadas (no contienen tuplas) en color naranja.

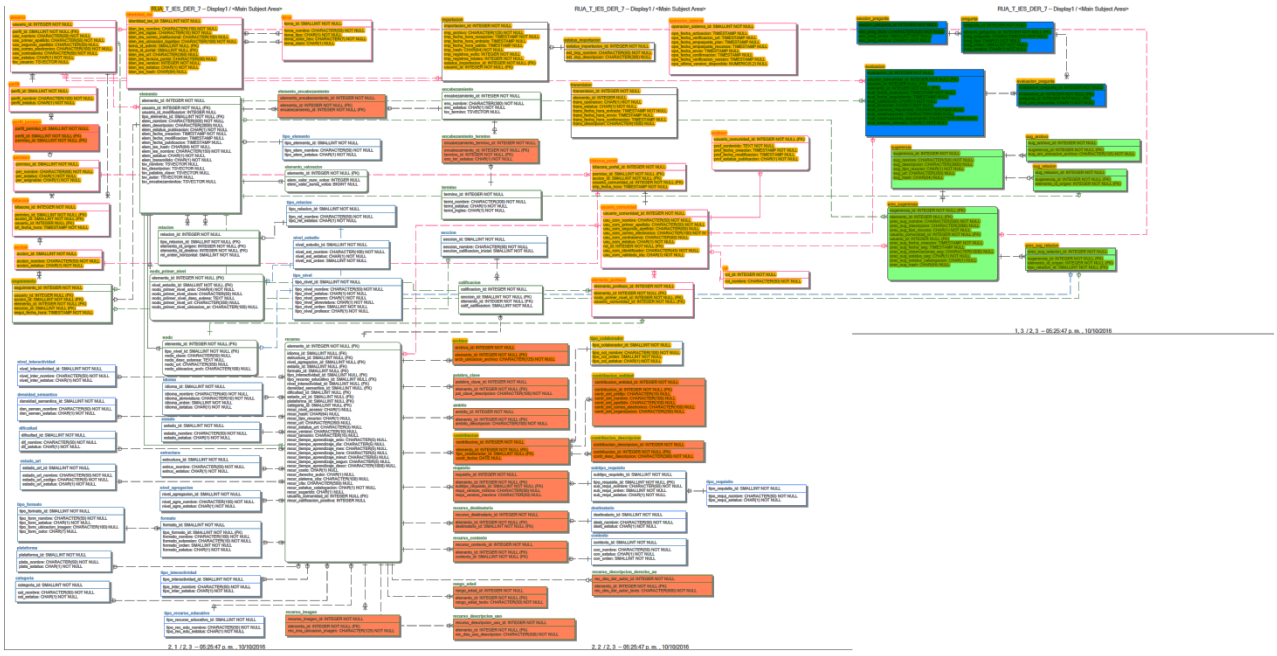


Figura 22. Diagrama Entidad-Relación RUA

Mediante el diccionario de datos también se indagó sobre las reglas de negocio y de los datos, en la figura 23 se muestra un fragmento del diccionario de datos en el que se observan los tipos de datos, si son únicos, el tipo de llave y algunos ejemplos que pueden almacenarce.

Diccionario de Datos
Sistema Web Red Universitaria de aprendizaje RUA IES

densidad_semantica		Contiene el listado de posibles densidades semánticas de un recurso educativo. La densidad semántica se refiere al grado de precisión de un recurso educativo en función de su tamaño, ámbito o en el caso de recursos autoregulados tales como audio y vídeo, la duración.					
N°	ATRIBUTO	DESCRIPCIÓN	TIPO DE DATO	LLAVE (PK / FK)	NULL / NOT NULL	ÚNICO (SÍ / NO)	EJEMPLO(S)
1	densidad_semantica_id	Identificador de *densidad_semantica*	SMALLINT	PK	Not Null	Sí	1, 2, 3 ...
2	den_seman_nombre	Nombre de la densidad semántica	VARCHAR(50)		Not Null	No	Muy baja, Baja, Media, Alta, Muy Alta
3	den_seman_estatus	Estatus de la densidad semántica	CHAR(1)		Not Null	No	**0** - Inhabilitado **1** - Habilitado

destinatario		Contiene el listado de posibles destinatarios de un recurso educativo. El destinatario se refiere al tipo de usuario a quien va dirigido el recurso educativo. Este valor se encuentra relacionado con la finalidad del objeto, es decir si el objeto es útil para enseñar, aprender, evaluar, etc.					
N°	ATRIBUTO	DESCRIPCIÓN	TIPO DE DATO	LLAVE (PK / FK)	NULL / NOT NULL	ÚNICO (SÍ / NO)	EJEMPLO(S)
1	destinatario_id	Identificador de *destinatario*	SMALLINT	PK	Not Null	Sí	1, 2, 3 ...
2	desti_nombre	Nombre que identifica al destinatario	VARCHAR(50)		Not Null	No	Administrador, Autor, Estudiante, Profesor
3	desti_estatus	Estatus del destinatario	CHAR(1)		Not Null	No	**0** - Inhabilitado **1** - Habilitado

Figura 23. Fragmento del Diccionario de Datos de la base de datos RUA

Mediante un programa desarrollado en Python se construyeron los histogramas correspondientes a cada columna o campo de todas las tablas de la base de datos RUA. El programa desarrollado se conecta a la base de datos mediante un motor de manipulación de bases de datos, llamado sqlAlchemy. El programa convierte el dominio de la columna a cadena de texto para facilitar el conteo de valores nulos, debido a que las funciones en Python descartan los valores nulos se ejecutó la técnica descrita. El proceso de obtención de resultados es considerablemente lento. En la figura 24 se muestra un histograma sobre el catálogo de densidad semántica.

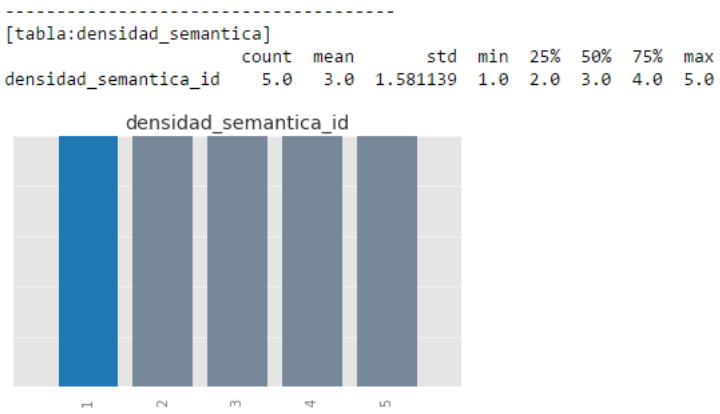


Figura 24. Histograma de la columna densidad semántica

Las gráficas obtenidas y las descripciones permitieron obtener un primer acercamiento para entender la densidad de los datos y los catálogos existentes. Con apoyo del Diagrama Entidad-Relación se diseñaron las vistas que contienen los datos candidatos a ser analizados, después se volvió a ejecutar el programa desarrollado para obtener las estadísticas de los campos de la tabla.

Al observar los gráficos, se concluyó que el dominio de las columnas son nominales; es decir son cadenas de texto que corresponden a categorías que pueden ser manejadas por los algoritmos de

clúster de Minería de datos, estos datos requerían de limpieza para que las categorías quedarán mejor tipificadas. En la figura 25 se muestra un histograma de la columna densidad_semántica _nombre en el que se ilustran las características de los datos.

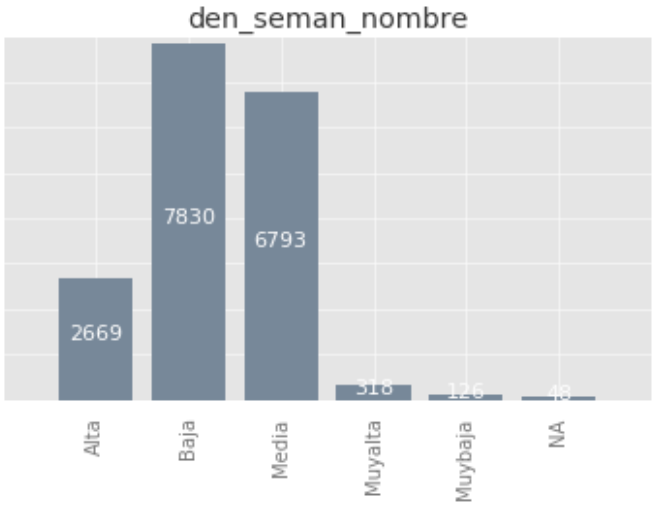


Figura 25. Histograma de la columna densidad semántica nombre

2.2 Fase de preparación

En la fase de preparación de los datos se tomó el primer diagnóstico y se limpiaron los datos. Cabe señalar que la limpieza de los datos puede ser de diferentes maneras, de las principales, se destaca el de eliminar las tuplas que contengan valores nulos en sus campos; otra, es de sustituir el valor nulo mediante el promedio o con cero, finalmente se puede realizar la sustitución mediante el valor más probable. Mediante el software de Knime es muy sencillo realizar las sustituciones descritas y no es necesario programar funciones para completar esta tarea.

También fue necesario reducir las categorías de los datos para un mejor manejo del algoritmo, por ejemplo, para las categorías: muy bajo, bajo, medio, alto y muy alto, se redujeron a tres: bajo medio y alto. La tarea de reducir el dominio de las columnas se identificó en el momento de pasar de la fase de preparación a la de modelado.

Los dominios numéricos requirieron de transformación; los datos numéricos se discretizaron, es decir se separaron en conjuntos de igual rango de números o en su caso se transformaron a nominales. Fue muy poco el tratamiento sobre valores numéricos porque la base en general contenía conceptos categóricos.

2.3 Fase de modelado

En la fase de modelado se eligieron los modelos de análisis que mejor se ajustan a los datos con el objetivo de obtener conocimiento acerca de la valoración de los recursos educativos del sistema RUA.

Los modelos elegidos fueron de clasificación por clúster, reglas de asociación, clasificación por árboles y también clasificación basada en probabilidad bayesiana.

2.4 Fase de evaluación

La siguiente etapa, la evaluación, se analizaron los resultados apoyándose en la Minería de datos y de las técnicas auxiliares de reportes a través de gráficas en el que se comparan los datos entre campos de la vista minable, se analizaron los errores de clasificación y los coeficientes de confianza para las reglas de asociación, se obtuvieron los siguientes resultados:

-Clasificación j48

```
Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       14462               81.3203 %
Incorrectly Classified Instances      3322               18.6797 %
Kappa statistic                          0.1379
Mean absolute error                      0.2838
Root mean squared error                  0.3776
Relative absolute error                 89.502  %
Root relative squared error             94.8425 %
Total Number of Instances            17784

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.107    0.013    0.670      0.107    0.185      0.214    0.687     0.375     Positivas
               0.987    0.893    0.818      0.987    0.895      0.214    0.687     0.881     Negativas
Weighted Avg.  0.813    0.719    0.789      0.813    0.754      0.214    0.687     0.781

=== Confusion Matrix ===

     a     b   <-- classified as
   377  3136 |     a = Positivas
   186 14085 |     b = Negativas
```

Figura 26. Métricas del algoritmo de clasificación j48

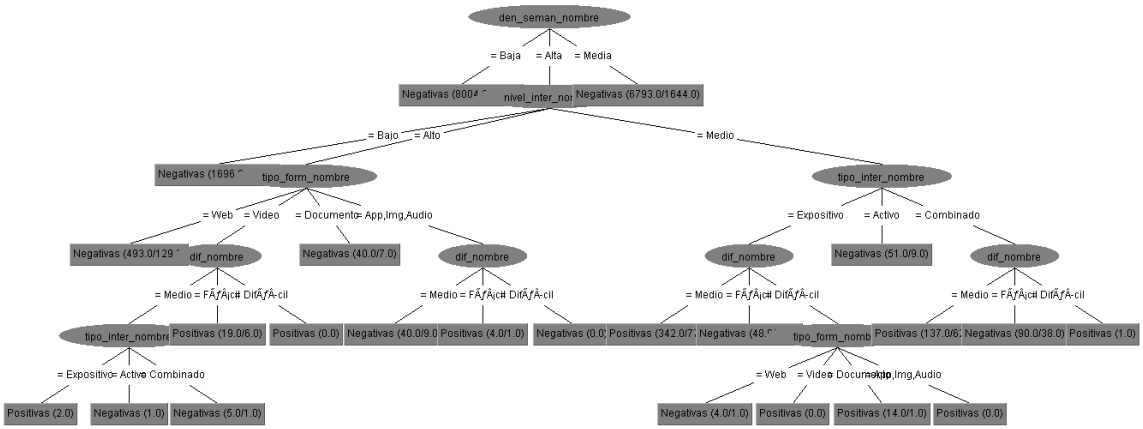


Figura 27. Árbol de decisión del algoritmo j48

Se observa en los resultados que las instancias incorrectamente clasificadas es de aproximadamente del 20%, este error se redujo de 30% a 20% después de aplicar limpieza, aun así es un error considerable, debido al sesgo en los datos y a la cantidad de los mismos; sin embargo, se puede apreciar la dependencia entre las categorías del árbol generado, la preferencia del usuario a valorar más los recursos de video con dificultad fácil y una alta interacción, por otra parte los recursos de documentos con interacción media y dificultad difícil presentan valoraciones positivas. Por otro lado los recursos web a pesar de tener mayor proporción hay poca participación en la valoración de los mismos.

-Farthest First

Para obtener más evidencias sobre los datos se aplicó una clasificación por clúster FarthestFirst (Envoltura de metaclúster que obtiene la distribución de probabilidad y densidad); se propusieron seis grupos para obtener más posibilidad de tener tipos diferentes de recurso, se descartaron los derechos de autor y los costos, se obtuvieron los siguientes resultados.

```
Cluster centroids:

Cluster 0
         [0-3] Bajo Baja Medio Web
Cluster 1
         [4-7] Medio Media FÃƒÂ¡cil Audio
Cluster 2
         [0-3] Alto Alta FÃƒÂ¡cil Video
Cluster 3
         [7-] Bajo Alta DifÃƒÂ-cil Documento
Cluster 4
         [4-7] Bajo Media Medio Video
Cluster 5
         [0-3] Alto Media Medio AplicaciÃƒÂ³n mÃƒÂ³vil


Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      14133 ( 79%)
1        831 (  5%)
2        985 (  6%)
3        207 (  1%)
4       1288 (  7%)
```

Figura 28. Reporte de clúster Farthest First

Los datos que se muestran en la figura 28 se agrupan los campos de: recurso calificación positiva, nivel de interacción, densidad semántica, dificultad y tipo de formato. Se obtiene mayor desglose para los recursos en video, pues son mejor valorados los videos con interactividad media, densidad semántica media y fácil dificultad; en contraparte, se observan contenidos con alta densidad semántica y alta interactividad que son menos valorados aunque la dificultad sea fácil.

También se observa que los documentos con baja interactividad y dificultad alta son más valorados, el número de usuarios que realizaron estas valoraciones es reducido.

En general hay poca participación para valorar los recursos educativos de RUA cuando se trata de contenidos con poca densidad semántica, nivel de interacción baja y de tipo web. En cuanto a los videos o audios son mejor valorados si la densidad de los contenidos y la interacción sobre los medios permanece en un nivel medio.

-Filtered Cluster

Este modelo de clasificación clúster agrupa equitativamente la cantidad de instancias en cada clúster conforme se define un numero mayor de grupos en el algoritmo entorno a un atributo. Se agrupan los atributos en base con otro clúster, en este caso se basó en el modelo simple Kmeans.

```
Final cluster centroids:
                                        Cluster#
Attribute                    Full Data          0          1          2          3          4          5
                             (17784.0)  (3998.0)  (2468.0)   (933.0)  (4957.0)  (3812.0)  (1616.0)
=============================================================================================================
recur_calificacion_positiva      [0-3]      [0-3]      [0-3]      [0-3]      [0-3]      [0-3]      [0-3]
nivel_inter_nombre                Bajo       Bajo      Medio       Alto       Bajo       Bajo       Bajo
den_seman_nombre                  Baja      Media      Media       Alta       Baja       Baja      Media
dif_nombre                      Medio      Medio      Medio      Medio  FÃƒÂ¡cil      Medio  FÃƒÂ¡cil
tipo_form_nombre                  Web      Video        Web        Web        Web Documento        Web



Time taken to build model (full training data) : 0.28 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       3998 ( 22%)
1       2468 ( 14%)
2        933 (  5%)
3       4957 ( 28%)
4       3812 ( 21%)
5       1616 (  9%)
```

Figura 29. Reporte de Filtered Cluster

-MakeDensityBased Clusterer

La clasificación de este algoritmo se basa en la densidad de un atributo de una clase en particular, pero presenta un alto porcencentaje de instancias no correctamente clasificadas ya que se basa en la desviación estándar, que en este caso no sería tan preciso debido a los atributos nominales. En la figura 30 se muestra la salida con los resultados de este algoritmo.

```
Missing values globally replaced with mean/mode

Final cluster centroids:
                                    Cluster#
Attribute                Full Data          0          1
                         (17784.0)  (12472.0)   (5312.0)
============================================================
recur_calificacion_positiva  [0-3]      [0-3]      [0-3]
nivel_inter_nombre            Bajo       Bajo      Medio
den_seman_nombre              Baja       Baja      Media
dif_nombre                  Medio      Medio      Medio
tipo_form_nombre              Web Documento        Web

Class attribute: dif_nombre
Classes to Clusters:

    0    1  <-- assigned to cluster
 8374 3654 | Medio
 3809 1816 | FÃƒÂ¡cil
  106   25 | DifÃƒÂ-cil

Cluster 0 <-- Medio
Cluster 1 <-- FÃƒÂ¡cil

Incorrectly clustered instances :       7594.0   42.7013 %
```

Figura 30. Reporte de clúster Make Density Clusterer

En los clústeres anteriores se constata la poca participación para valorar los recursos y los tipos de formatos más frecuentes son documento y web.

-Naive Bayes

Con este modelo se logró clasificar el mayor número de instancias, se puede observar el resultado con mayor detalle y comparar los valores por categorías de acuerdo a las probabilidades con respecto a un valor de clase. Los resultados se muestran a continuación.

Class counts for recur_calificacion_positiva

Class:	0-3	4-7	8-11
Count:	16513	1170	101

Total count: 17784

Threshold to used for zero probabilities: 0.0

P(den_seman_nombre | class=?)

Class/den_seman_nombre	Alta	Baja	Media
0-3	2527	7762	6224
4-7	429	229	512
8-11	31	13	57
Rate:	17%	45%	38%

P(dif_nombre | class=?)

Class/dif_nombre	Difícil	Fácil	Medio
0-3	114	5103	11296
4-7	16	481	673
8-11	1	41	59
Rate:	1%	32%	68%

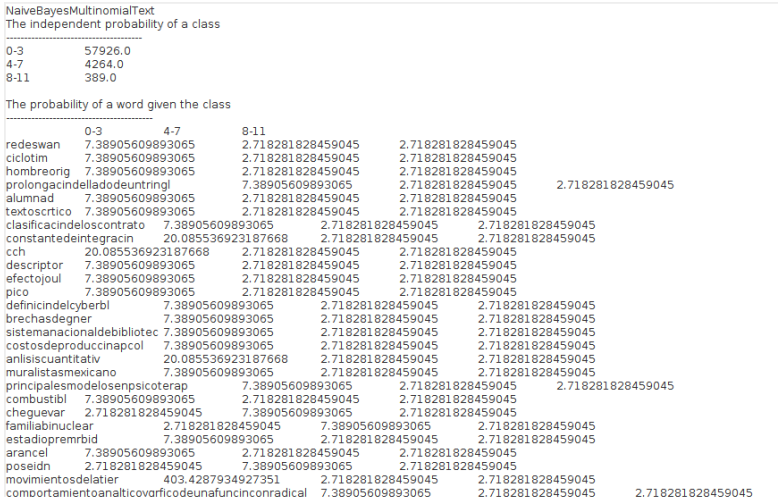
Figura 31. Reporte de clasificación Naive Bayes

En los resultados se observa que las categorías tipificadas con nivel medio tienen mayor proporción en comparación con niveles difíciles y densidades semánticas altas.

El error de la clasificación bayesiana para este ejemplo fue del 10%.

-Naive Bayes Multinomial Text

Este modelo permite obtener la probabilidad de una palabra dado el rango de calificaciones positivas, para este caso la palabra es la referencia conceptual del recurso y engloba el contenido del recurso en pocos conceptos. Después de aplicar el algoritmo se obtuvieron las proporciones de los datos. Los datos aunque no estén normalizados y tengan poca legibilidad, pueden servir para investigar más sobre el contenido semántico de los recursos de aprendizaje.



NaiveBayesMultinomialText
The independent probability of a class

0-3	57926.0
4-7	4264.0
8-11	389.0

The probability of a word given the class

	0-3	4-7	8-11		
redeswan	7.38905609893065	2.718281828459045	2.718281828459045		
ciclotim	7.38905609893065	2.718281828459045	2.718281828459045		
hombreorig	7.38905609893065	2.718281828459045	2.718281828459045		
prolongacindelladodeuntringl		7.38905609893065	2.718281828459045	2.718281828459045	
alumnad	7.38905609893065	2.718281828459045	2.718281828459045		
textoscrtico	7.38905609893065	2.718281828459045	2.718281828459045		
clasificacindeloscontrato	7.38905609893065	2.718281828459045	2.718281828459045		
constantedeintegracin	20.085536923187668	2.718281828459045	2.718281828459045		
cch	20.085536923187668	2.718281828459045	2.718281828459045		
descriptor	7.38905609893065	2.718281828459045	2.718281828459045		
efectojoul	7.38905609893065	2.718281828459045	2.718281828459045		
pico	7.38905609893065	2.718281828459045	2.718281828459045		
definicindelcyberbl	7.38905609893065	2.718281828459045	2.718281828459045		
brechasdegner	7.38905609893065	2.718281828459045	2.718281828459045		
sistemanacionaldebibliotec	7.38905609893065	2.718281828459045	2.718281828459045		
costosdeproduccinapcol	7.38905609893065	2.718281828459045	2.718281828459045		
anlisiscuantitativ	20.085536923187668	2.718281828459045	2.718281828459045		
muralistasmexicano	7.38905609893065	2.718281828459045	2.718281828459045		
principalesmodelosenpsicoterap		7.38905609893065	2.718281828459045	2.718281828459045	
combustibl	7.38905609893065	2.718281828459045	2.718281828459045		
cheguevar	2.718281828459045	7.38905609893065	2.718281828459045		
familiabinuclear	2.718281828459045	7.38905609893065	2.718281828459045		
estadiopremrbid	7.38905609893065	2.718281828459045	2.718281828459045		
arancel	7.38905609893065	2.718281828459045	2.718281828459045		
poseidn	2.718281828459045	7.38905609893065	2.718281828459045		
movimientosdelatier	403.4287934927351	2.718281828459045	2.718281828459045		
comportamientoanalticovarficodeunafuncinconradical	7.38905609893065	2.718281828459045	2.718281828459045		

Figura 32. Reporte de clasificación multinominal de texto

Por último, las reglas de asociación obtenidas con un nivel de confianza mayor al 75% (ver figura 33), nos muestra que los recursos educativos, en general no tienen costo, la dificultad es media o baja y tienen derechos de autor.

Antecedent		Consequent
recur_derecho_autor=S, den_seman_nombre=Baja, tipo_inter_nombre=Expositivo	→	recur_costo=N, nivel_inter_nombre=Bajo
recur_derecho_autor=S, nivel_inter_nombre=Bajo, den_seman_nombre=Baja	→	recur_costo=N, tipo_inter_nombre=Expositivo
recur_derecho_autor=S, nivel_inter_nombre=Bajo, dif_nombre=Medio	→	recur_costo=N, tipo_inter_nombre=Expositivo
recur_derecho_autor=S, dif_nombre=Medio, tipo_inter_nombre=Expositivo	→	recur_costo=N, nivel_inter_nombre=Bajo
recur_costo=N, recur_derecho_autor=S, den_seman_nombre=Baja	→	nivel_inter_nombre=Bajo, tipo_inter_nombre=Expositivo
nivel_inter_nombre=Bajo, dif_nombre=Medio, tipo_inter_nombre=Expositivo	→	recur_costo=N, recur_derecho_autor=S
nivel_inter_nombre=Bajo, den_seman_nombre=Baja, tipo_inter_nombre=Expositivo	→	recur_costo=N, recur_derecho_autor=S
recur_costo=N, den_seman_nombre=Baja, tipo_inter_nombre=Expositivo	→	recur_derecho_autor=S, nivel_inter_nombre=Bajo

Figura 33. Reglas de asociación

Con el resultado obtenido se validan los objetivos para el tipo de usuarios a los que está dirigida la plataforma, se reconoce la oportunidad de planificar los recursos de enseñanza entre la extensión del contenido, la dificultad y el nivel de interacción.

El análisis puede plantear las métricas para estudios pedagógicos sobre la creación de contenidos didácticos en los formatos requeridos y la medición de la valoración o aprovechamiento de los recursos ya sea para los usuarios de la plataforma RUA o para otros si es conveniente.

2.5 Fase de despliegue

Por último en la fase de despliegue o difusión, se reportaron los datos y se propusó la presentación de los resultados considerando el tipo de formato y el software para el monitoreo en tiempo real

de la base de datos para obtener reportes estadísticos en forma periódica. También, la información obtenida puede servir de apoyo para los diseñadores de depósitos de datos para orientar el proceso ETL y creación de cubos OLAP mediante el diseño de constelaciones de tablas.

A medida que se pasa por las etapas del modelo CRISP se van descubriendo áreas de oportunidad para analizar los datos y se va refinando más el proceso de obtención de conocimiento. En el análisis desarrollado se enfocó más en la valoración de los recursos para obtener un parámetro numérico a partir de la relación de categorías conceptuales con el fin de extrapolar el análisis al aprovechamiento escolar mediante un parámetro de valoración, uno de calificación (promedio académico) o ambos.

La base de datos tiene otras áreas de oportunidad para el análisis pero a medida que se recorre el ciclo del modelo CRISP se podrá identificar claramente otros recursos útiles para monitoreo o para ofrecer información sobre los recursos educativos a partir de la fuente de la provengan y de la materia que traten, para lograr una distinción entre las valoraciones de materiales de ciencias básicas, artes, ciencias sociales o ciencias de la salud.

Mediante la herramienta Tableau se puede implantar estrategias de monitoreo sobre los datos en tiempo real, a continuación se proponen las dimensiones de las gráficas para obtener información sobre el sistema RUA.

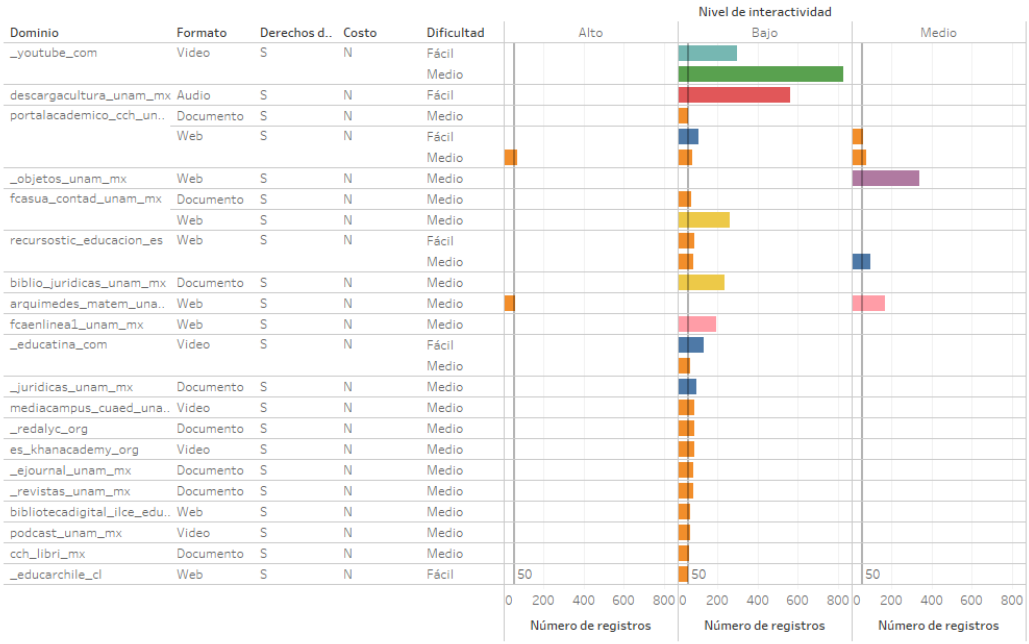


Figura 34. Gráfica de recursos por procedencia, despliegue por categorías

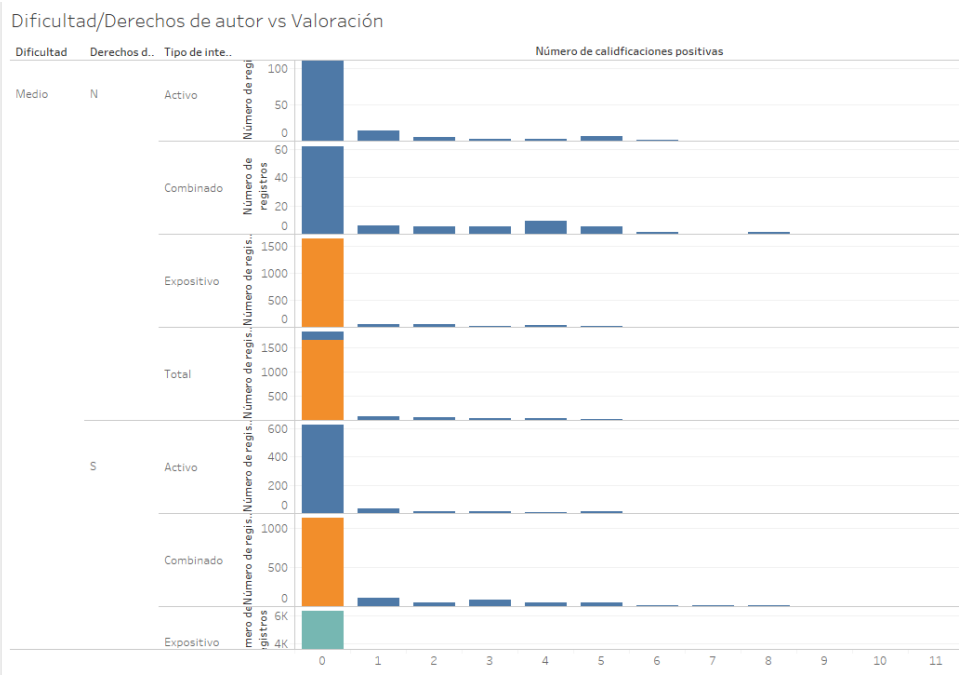


Figura 35. Histogramas de valoración por categorías

Como parte del reporte de experiencia, se documentó la dificultad de limpiar los datos que presentaban problemas de codificación y la dificultad de carga de datos en Weka por el mismo problema. A partir de la experiencia se propone una estrategia para prevenir errores (a causa de la codificación de caracteres) que puedan incidir en los reportes finales.

Se desarrolló la habilidad de tener bien definidas las estrategias de limpieza de datos y transformación de dominios.

Los resultados permiten establecer las estrategias para desarrollar módulos en Python para analizar los datos de acuerdo al dominio de la base de datos con conexiones a Big Data.

3. Análisis de texto

Como parte de las obligaciones de la UNAM que requiere el INAI, se analizó la Ley General de Transparencia y Acceso a la Información Pública para sintetizar la información y la relación de conceptos entre fracciones de los artículos 70, 75 y 77 para su difusión en páginas web. Esta actividad también tenía la intensión de demostrar el uso de las herramientas de análisis de texto para su posible implementación en el proyecto “Sistema de Información Universitaria”.

3.1 Extracción y pre-procesamiento

Para analizar texto es necesario tener una estructura o corpus, el corpus es el conjunto de documentos almacenados en una tabla; cada renglón representa un documento y, en las columnas, se almacenan los metadatos del documento como la referencia o el texto a analizar.

La extracción del corpus se realizó mediante programación en Python para estructurar la ley a partir de un archivo pdf, este archivo contiene la redacción de la ley. El conteo de palabras de n-gramaticas [15] se realizó en Python y se comparó con otras herramientas, la finalidad de la comparación era reducir el tiempo de preparación y análisis.

El corpus debe tener un formato "tab" para realizar el análisis desde Orange 3, por lo que es necesario pasar de un archivo en formato "csv" a un archivo en formato "tab". Por otra parte, en Knime la conversión se puede realizar en el momento de programar visualmente por bloques.

En la siguiente etapa, el pre-procesamiento, se limpia el corpus con palabras no significativas del español (stopwords); signos de puntuación; etiquetado POS (Position On Speech) y otras características de interés para el análisis de datos. En la figura 36 se muestra el corpus y las palabras etiquetadas.

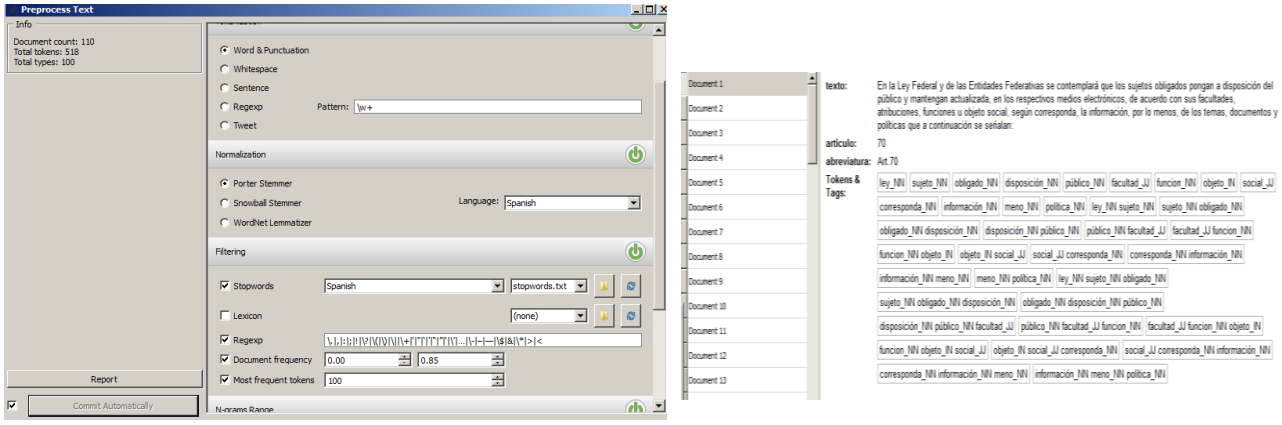


Figura 36. Exploración del corpus después del pre-procesamiento

Mediante el software Knime se obtuvó el etiquetado POS basado en el procesador de lenguaje natural para el español de la Universidad de Stanford; las palabras formadas con un máximo de 3-gramáticas; parámetros de frecuencia relativa; entropía ; distancia relativa y distancia absoluta de la palabra en el corpus. Este software ofrece una mayor gama de nodos para el procesamiento de texto y oportunidad de exportar los datos para realizar la visualización que representa el procesamiento de texto.

[15] N-gramáticas es una técnica ejecutada de pre-procesamiento en la que se obtienen conceptos de n palabras

3.2 Grafo conceptual

Gephi es una herramienta que permitió crear grafos conceptuales. Los grafos conceptuales indicaban los conceptos relacionados mediante líneas dirigidas a las categorias del texto, por requerimientos se precisó por artículo y fracción, por ejemplo: Art. 70 XVII. Cabe señalar que la ley no presentó artículos con apartados por lo que se habría caído en una confusión, pero esencialmente son conceptos pertenecientes a un mismo artículo. Se realizó investigación sobre la estructura de una ley mexicana[16] para la conversión del texto semi-estructurado a estructurado.

El grafo conceptual contiene información sobre la frecuencia de las palabras mediante el tamaño de los nodos, las etiquetas de los nodos pueden representar las palabras o las categorías de las palabras o referencias. La magnitud de los nodos que representan las categorías de palabras indican el porcentaje de palabras utilizadas de la oración o el grado de salida del nodo, es decir, cuántas palabras estan conectadas a dicho nodo.

Todas los datos posibles se obtuvieron para las métricas del grafo, los datos se plasmaron mediante una aplicación web interactiva para discriminar los grupos de nodos, la relación entre los conceptos y lograr simplificar el análisis de los requerimientos legales. En la figura 37 se muestra el gráfo producido (generado con la extensión sigma.js que genera una aplicación web) y la capacidad interactiva que posee, también se muestra la modificación que se efectuó sobre archivos de configuración JSON para el renderizado del grafo conceptual, desafortudamente no se plasmaron las distancias de las palabras sobre las relaciones, pero sobre el ambiente de gephi se puede ver esta información.

[16] **Senado de la República, IJJ UNAM.2010**. División Estructural de la Ley. III

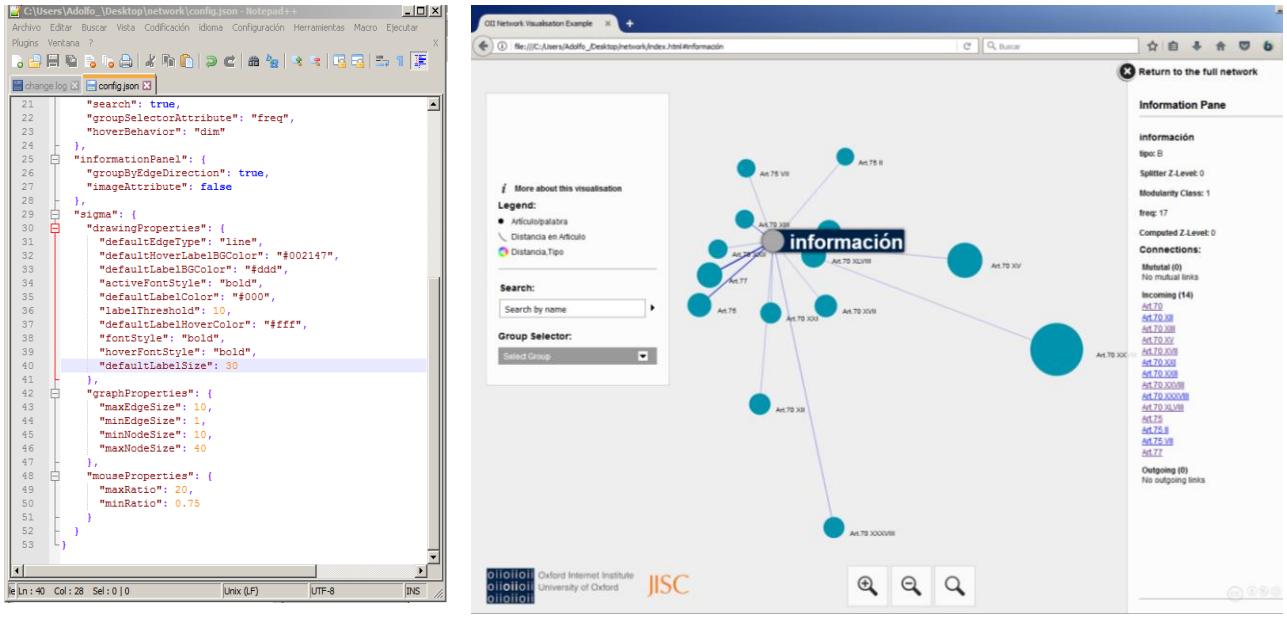


Figura 37. a) Edición de la condiguración JSON, b)Vista de la aplicación interactiva

Por otra parte se generó un grafo conceptual con información de la Escuela Nacional Preparatoria para identificar la relación entre los conceptos y las materias de acuerdo al contenido de los planes de estudio, en la figura 38 se muestra el grafo generado.

Figura 38. Nodo termodinámica y su relación con materias de la ENP

El grafo de conceptos y sus relaciones por materia permite identificar que conceptos pueden ser redundantes entre los planes de estudio o como pueden afectar al aprendizaje bajo cierta configuración de relaciones entre conceptos, una buena propuesta es agregar información de

aprovechamiento escolar mediante el tamaño de la arista o del nodo para identificar las áreas de mejora para la planificación de materiales de aprendizaje.

También se realizó un grafo conceptual con información del Colegio de Ciencias y Humanidades en el que se observa la relación de los conceptos entre los semestres de los planes de estudio del 2003 y los planes del 2016 con el fin de indagar en los cambios generados por niveles. Como se puede identificar en la figura 39, el concepto de energía dejo de estar presente en el quinto y primer semestre del plan 2016.

Este tipo de conocimiento es de gran utilidad para indagar en los cambios entorno a un concepto en particular para la planeación de los planes de estudio y de que manera puede influir en el aprendizaje.

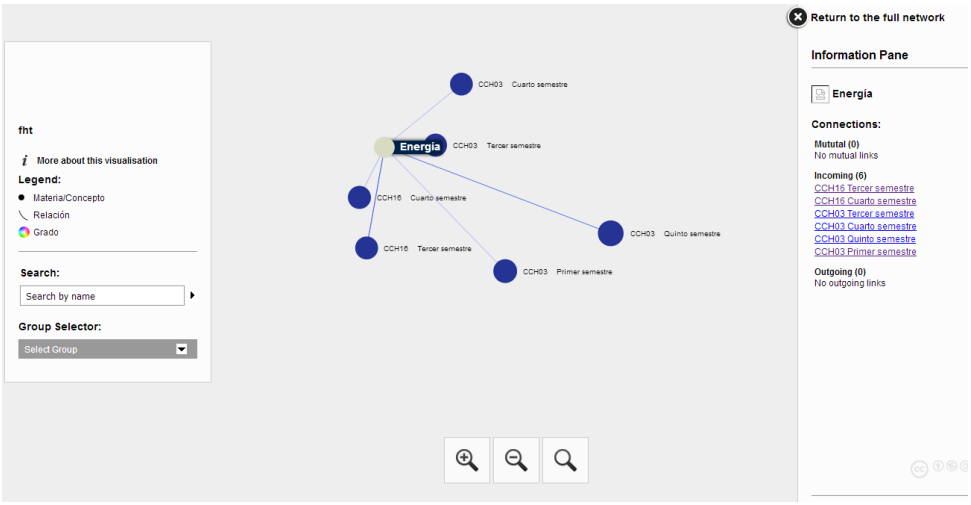


Figura 39. Nodo energía, sus relaciones por grado y plan de estudio

3.3 Similaridad de documentos

Por otra parte, de las capacidades de Knime se logró obtener vectores de documentos. El documento vectorizado consiste en enlistar las clasificaciones o referencias del documento en los renglones, las palabras, en las columnas; en cada celda se indica si la palabra pertece al documento mediante un "uno" o la frecuencia de la palabra, y en caso contrario se indica mediante un "cero"; de lo anterior se conforma una tabla de n-dimensiones correspondientes a las dimensiones del espacio vectorial de palabras, algo similar al espacio vectorial de dimensión "n" en el álgebra lineal.

or	D dotadas	D auton...	D actual...	D respo...	D solicit...	# neare...	...	D similar...	S Category	S Catego...
	0	0	0	0	0	1	...	0.722	Art.70 XXVIII	Art.70 XXVIII
	0	0	0	0	0	1	...	0.707	Art.70 III	Art.70 XV
	0	0	0	0	0	1	...	0.845	Art.70 XXVIII	Art.70 XXVIII
	0	0	0	0	0	1	...	0.845	Art.70 XXVIII	Art.70 XXVIII
	0	0	0	0	0	1	...	0.722	Art.70 XXVIII	Art.70 XXVIII
	0	0	0	0	0	1	...	0.566	Art.70 XXVIII	Art.77 I
	0	0	0	0	0	1	...	0.566	Art.77 I	Art.70 XXVIII
	0	0	0	0	0	1	...	0.555	Art.70 XXX...	Art.70 XX
	0	0	0	0	0	1	...	0.555	Art.70 XX	Art.70 XXX...
	0	0	0	2	0	1	...	0.671	Art.77 II	Art.70 XXVIII
	0	0	0	2	2	1	...	0.671	Art.70 XXVIII	Art.77 II
	0	0	0	2	2	2	...	0.671	Art.70 XXVIII	Art.70 XXVIII
	0	0	0	0	0	1	...	0.707	Art.70 XV	Art.70 III
	0	0	0	2	2	1	...	0.671	Art.70 XXVIII	Art.70 XXVIII

Tabla 4. Similaridad por distancia coseno entre segmentos de la ley

De la tabla 4 se muestran las categorias de la ley de transparencia de mayor similitud, puede entenderse como porcentajes, por ejemplo, el Art.77 I y Art.70 XXVIII tienen un porcentaje de similitud del 56.6% .

Los espacios de palabras son muy utiles para encontrar la similitud entre dos categorias o documentos, ya que se puede aplicar la distancia coseno entre los vectores. La distancia coseno consiste en el producto interno de dos vectores que pertenecen al espacio vectorial de palabras.

Desde otra perspectiva el espacio vectorial puede cambiarse por un espacio de documentos mediante una operación de transposición. Los nuevos vectores permiten aplicar operaciones de suma o resta para obtener relaciones entre palabras y clasificaciones.

3.4 Proyección del espacio de palabras

También se realizó la proyección del espacio de “n” dimensiones de palabras a un espacio de dimensión “dos” para visualizar la distancia entre los conceptos, en el gráfico generado no es posible que muestre las etiquetas que indiquen el concepto, sólo seleccionando el nodo; una estrategia para resolver este problema de visualización es la obtención de las coordenadas de las palabras y trabajarlas con gephi. En la figura 40 se muestra la dispersión de conceptos en el ambiente Knime, se puede apreciar la densidad en un espacio de dimensión tres, pero esencialmente, es la proyección de dos dimensiones; ya que sólo se repitió una dimensión como la tercera.

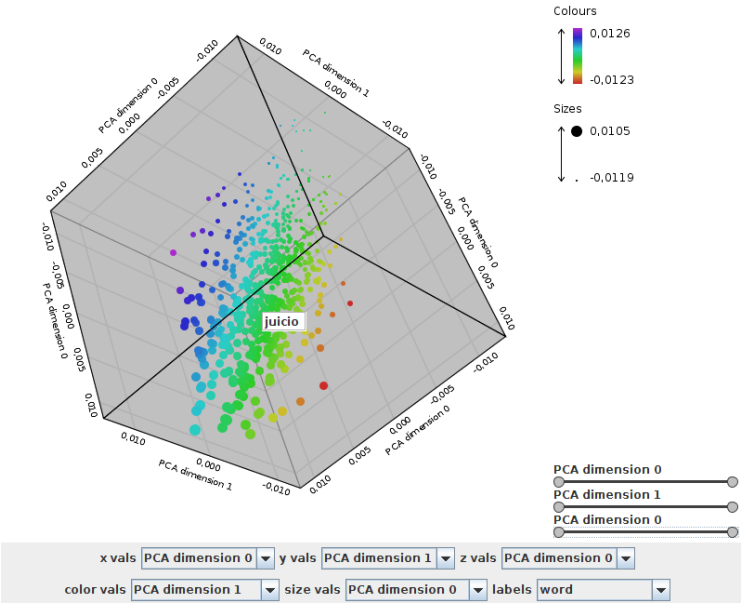


Figura 40. Proyeccion PCA del espacio vectorial de palabras

3.5 Dendograma

Aunque el software de Orange 3 presente dificultades para importar o exportar datos, dispone de widgets muy prácticos para analizar datos, uno de ellos es el widget hierarquical cluster para el análisis de texto, mediante la clasificación jerárquica de los conceptos se genera como salida un dendograma que relaciona los conceptos por la distancia a la que se encuentran entre sí, en la figura 41 se muestra el dendograma generado al aplicar el algoritmo a la columna de texto en el corpus y ajustar el umbral para la mejor apreciación posible y en la figura 42 se presenta una vista con mayor detalle sobre las leyes de transparencia y acceso a la información pública.

El análisis correspondiente permite sintetizar el contenido de la información, se expone la siguiente síntesis.

“La ley general de transparencia y acceso a la información pública en sus artículos 70,75 y 77 contiene las normas aplicables a sujetos obligados sobre información por áreas de la dependencia en las que se especifiquen los informes de contratos, fondos, sueldos por nivel, convocatorias, concursos, programas y demás sobre servicios públicos en formatos de listado especificados por el INAI.”

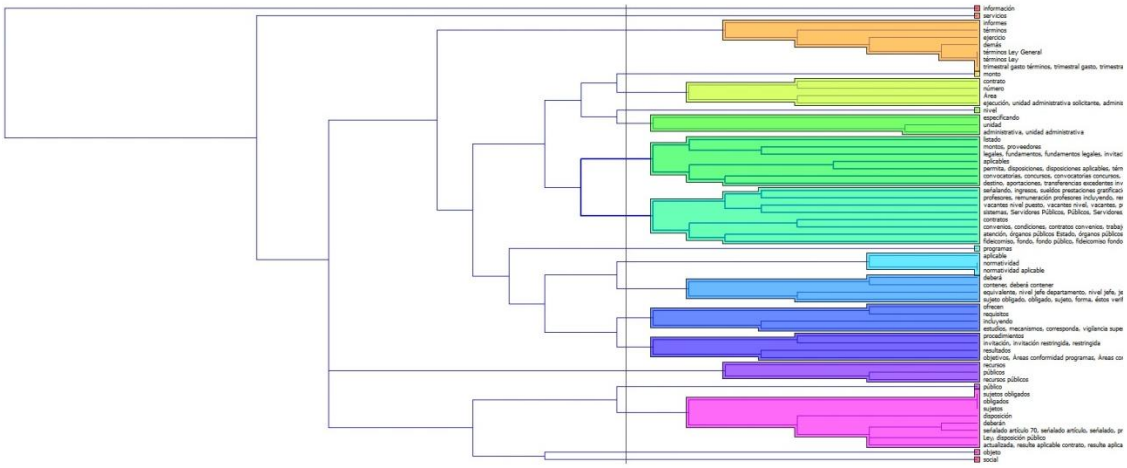


Figura 41. Dendograma INAI

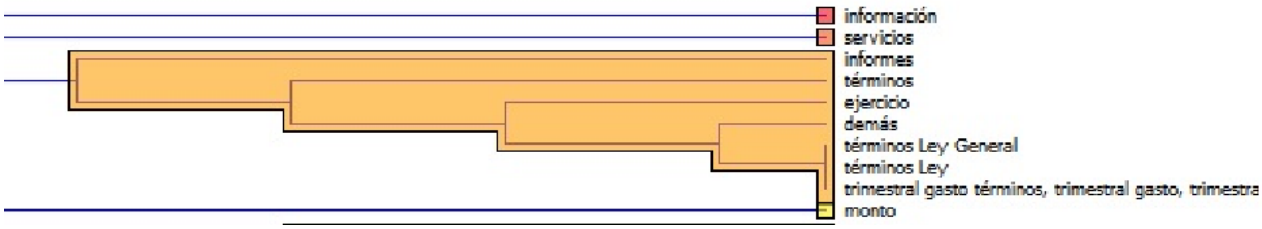


Figura 42. Dendograma INAI detallado

Los ejemplos desarrollados permitieron tener una base de análisis para aplicarlo a datos de texto y complementar las capacidades de los diferentes productos de software.

Capítulo V Problemas técnicos y soluciones

Se presentaron varios problemas de instalación y usabilidad en las herramientas de software, estos problemas de no tenerlos presentes pueden perjudicar el desarrollo de la estrategia de análisis. En la documentación generada, y entregada al líder técnico del proyecto, se expresan los problemas y las posibles rutas de solución que pueden formar parte de la lista de riesgos del proyecto.

A manera de tabla se presenta la información documentada.

Software	Errores	Soluciones
WinPython (3.5.3.0)	Falló en actualización de paquetes como scipy (compilación lenguaje c)	Utilizar un compilador MVBC Utilizar Conda
Conda	Compatibilidad de aplicaciones con Qt5	Utilizar Conda (v. 4.1.1 64 bits) en Windows o utilizar versión 3.6 en Linux que integra Qt4
Instalación Orange 3 **Distribución 2UDA (sólo 64 bits)**	Renderizado del entorno gráfico	Instalar la distribución de Conda
Weka	Ubicación de controladores	Instalación con conexión a internet del paquete JDBCDriversDummyPackage que permite ubicar los controladores de bases de datos sobre los directorios del paquete y de weka.
Weka	Importación de tablas consultadas con SQL	Mapeo de tipos de datos nativos a la base de datos a tipos de datos nativos de java en el archivo DatabaseUtils.props
Orange 3	Requiere de la función quantile escrita en lenguaje C con la API de Linux para PostgreSQL (>=9.5).	-Instalar un ambiente virtual en Linux con PostgreSQL >9.5 e instalar los paquetes de desarrollo para compilar las funciones. -Utilizar la solución cross-platform 2UDA que distribuye de PostqreSQL con las funciones requeridas - Compilar desde Windows con MVBC y la API de Linux Compilar mediante herramientas Cygwin o MinGW que provean del wrapper de POSIX - Compilar desde Linux con Windows SDK
	Requiere de la función tsm_system_time	Instalar una versión mayor a la 9.5
Tableau	-Requiere compatibilidad entre versiones de Tableau desktop y Tableau server -Cambia los tipos de datos enteros a flotantes (posible causa: la presencia de valores nulos)	Actualizar la versión del servidor
Knime	Ejecución del entorno con lag o latencia grande	Modificar la capacidad del heap de memoria de knime
PostgreSQL	Restaurar una base de datos mediante el archivo con extensión backup o sql	Verificar la codificación y en su caso cambiarla a UTF8 bajo un template diferente

Tabla 5. Problemas Técnicos

En la información entregada se indicó el proceso de instalación para evitar los errores encontrados. Debido a que es software libre se puede justificar los problemas de usabilidad e instabilidad de las aplicaciones del software Orange 3, ya que la comunidad de desarrollo implementó los cambios actuales. La documentación de instalación se adjuntó a este reporte en el Anexo D.

Finalmente, haciendo énfasis en el error de conexión a base de datos en el entorno de Orange 3, la integración de funciones escritas en lenguaje C para PostgreSQL se justifica por la velocidad de conexión a la base de datos para extraer cuantiles y tener mayor velocidad de procesamiento de los datos. La figura 43 muestra una comparativa entre los tiempos de respuesta de las funciones de bases de datos escritas en diferentes lenguajes.

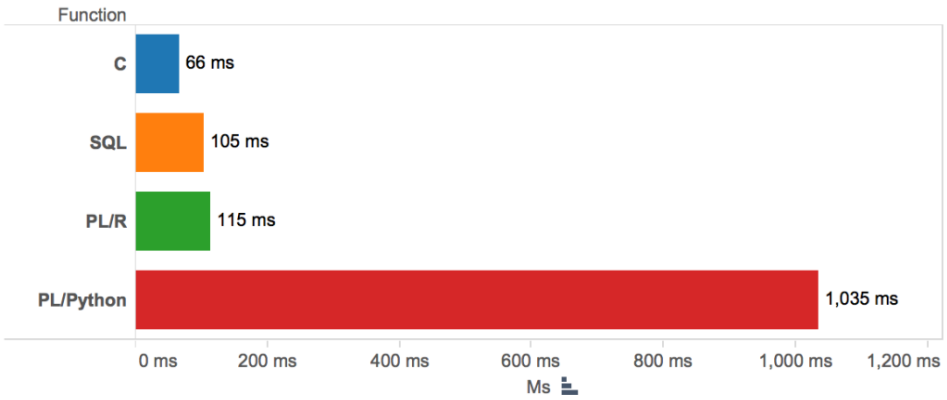


Figura 43. Tiempo de respuesta de diferentes tipos de funciones [17]

Como se puede observar en la figura 43 las funciones escritas en lenguaje C tienen un menor tiempo de respuesta por lo que es muy conveniente integrarlas a Python, de hecho la librería Scipy logró esta integración y es la más popular en la actualidad para el análisis científico y educativo. Sin embargo, los componentes adicionales de software pueden retardar el procesamiento de los datos, este hecho es factor importante en el momento de integrar las herramientas de análisis a los procesos del sistema.

Por otra parte se presentaron muchas incompatibilidades entre versiones y en la realización de joins debido a que los datos presentaban espacios en blanco por lo que es conveniente eliminar los mismos para obtener un resultado de la operación.

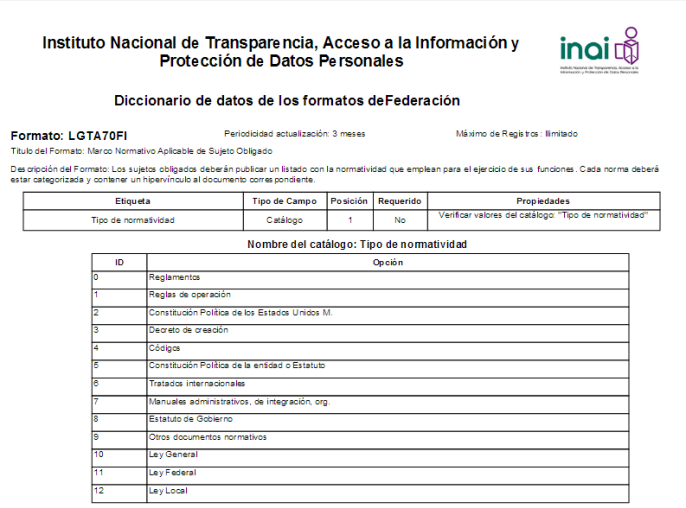
[17] **Kris(2017).**[En línea][Citado el: 2 Abril de 2017].https://www.datascienceriot.com/644/kris/

Capítulo IV Actividades adicionales

Durante el periodo de prestación de servicio social se realizaron actividades de diseño y programación que apoyaron la implementación de sistemas y de requerimientos para la División de Colaboración y Vinculación.

Actividad 1

Se analizó documentación[18] para extraer, del diccionario de datos, los catálogos que se utilizarán en el proyecto del INAI, la información analizada se entregó en archivos excel con la especificación y contenido de cada formato. A continuación se incluye un fragmento de la documentación:



Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales

Diccionario de datos de los formatos deFederación

Formato: LGTA70FI Periodicidad actualización: 3 meses Máximo de Registros : Ilimitado

Título del Formato: Marco Normativo Aplicable de Sujeto Obligado

Descripción del Formato: Los sujetos obligados deberán publicar un listado con la normatividad que emplean para el ejercicio de sus funciones. Cada norma deberá estar categorizada y contener un hipervínculo al documento correspondiente.

Etiqueta	Tipo de Campo	Posición	Requerido	Propiedades
Tipo de normatividad	Catálogo	1	No	Verificar valores del catálogo: "Tipo de normatividad"

Nombre del catálogo: Tipo de normatividad

ID	Opción
0	Reglamentos
1	Reglas de operación
2	Constitución Política de los Estados Unidos M.
3	Decreto de creación
4	Códigos
5	Constitución Política de la entidad o Estatuto
6	Tratados internacionales
7	Manuales administrativos, de integración, org.
8	Estatuto de Gobierno
9	Otros documentos normativos
10	Ley General
11	Ley Federal
12	Ley Local

Figura 44. Documentación de requerimientos INAI

Los conceptos clasificados en el archivo excel se entregaron al líder de proyecto para el diseño de la base de datos. Los conceptos se clasificaron con mayor detalle en directorios y se realizó la extracción de datos para cargarlos a la base de datos, estos datos sirvieron para la validación con el cliente.

[18] La documentación presentaba información confusa debido a que se utilizaron versiones diferentes de renderizado entre las herramientas del sistema de origen a pdf bajo la especificaciones de Acrobat

Mediante la programación sobre directorios en Python se logró generar los archivos con las inserciones de valores en tablas (DML) en formato “sql”. La figura 45 muestra parte del código que extrajo y dio forma al archivo sql.

```
rootDir='C:\\Users\\boromir\\Desktop\\INAI'
for dirName, subdirList, fileList in os.walk(rootDir):
    #print('Directorio encontrado: %s' % dirName)
    for fname in fileList:
        if "xls"  in fname:
            insertVector(dirName +"\\"+ fname)
```

```
--C:\Users\boromir\Desktop\INAI\ART 70\01__LGTA70FI\LGTA70FI-Formato Marco Normativo Aplicable de Sujeto Obligado.xls
('LGTA70FI','10020','Tipo de normatividad','9'),
('LGTA70FI','10021','Denominación de la norma','2'),
('LGTA70FI','10022','Fecha de publicación en DOF u otro medio','4'),
('LGTA70FI','10026','Fecha de última modificación','4'),
('LGTA70FI','10023','Hipervínculo al documento de la norma','7'),
('LGTA70FI','10024','Fecha de validación','4'),
('LGTA70FI','10025','Área responsable de la información','1'),
('LGTA70FI','10017','Año','12'),
('LGTA70FI','10018','Fecha de Actualización','13'),
('LGTA70FI','10019','Nota','14'),
--C:\Users\boromir\Desktop\INAI\ART 70\02__LGTA70FII\LGTA70FII-Formato Estructura Orgánica.xls
('LGTA70FII','10033','Denominación del área.','2'),
('LGTA70FII','10032','Denominación del puesto.','2'),
('LGTA70FII','10034','Denominación del cargo','1'),
('LGTA70FII','10035','Clave o nivel del puesto','1'),
('LGTA70FII','10028','Tipo de integrante','9'),
('LGTA70FII','10036','Área de adscripción','1'),
('LGTA70FII','10037','Denominación de la norma','2'),
('LGTA70FII','10038','Fundamento Legal','1'),
('LGTA70FII','10039','Atribuciones, responsabilidades y/o funciones','2'),
('LGTA70FII','10040','Hipervínculo al perfil','7'),
('LGTA70FII','10041','Prestadores de servicios','2'),
('LGTA70FII','10042','Hipervínculo al Organigrama','7'),
('LGTA70FII','10027','Leyenda respecto de los prestadores de servicios','1'),
('LGTA70FII','10043','Fecha de validación','4'),
('LGTA70FII','10044','Área responsable de la información','1'),
('LGTA70FII','10031','Año','12'),
('LGTA70FII','10029','Fecha de actualización','13'),
('LGTA70FII','10030','Nota','14'),
--C:\Users\boromir\Desktop\INAI\ART 70\03__LGTA70FIII\LGTA70FIII-Formato Las facultades de cada Área.xls
('LGTA70FIII','10049','Denominación del Área.','2'),
```

Figura 45. Programación sobre directorios

Finalmente se comprobó la carga del archivo que contiene las tuplas de la tabla, se comprobaron permisos de ejecución, de codificación y se entregó la documentación correspondiente al líder de proyecto.

Actividad 2

Se realizó un diseño basado en el Modelo Entidad-Relación con las dependencias de la UNAM para catalogar a las mismas, este modelo sirvió para el cumplimiento de los requerimientos por parte del INAI para con la UNAM en base a la Ley General de Transparencia y Acceso a la Información Pública (artículos 70 75 77). La figura 46 muestra el diagrama entidad-relación generado, también se incluye el modelo relacional y la definición de datos.

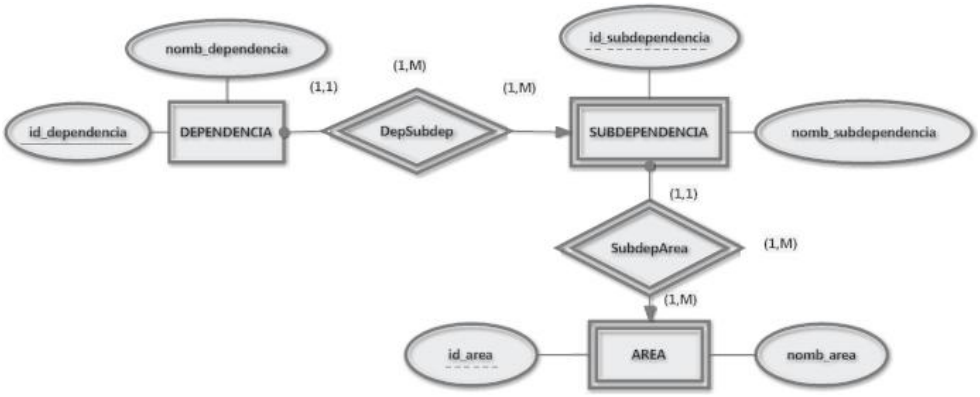


Figura 46. Diagrama Entidad-Relación dependencias UNAM

El modelo Relacional generado:

DEPENDENCIA(id_dependencia, nomb_dependencia)

SUBDEPENDENCIA(id_dependencia(FK),id_subdependencia(D),nomb_subdependencia)

AREA((id_dependencia(FK),id_subdependencia(D))(FK), id_area(D),nomb_area)

Y la definición de datos mediante DDL:

```
CREATE TABLE "DEPENDENCIA"
(
  id_dependencia numeric NOT NULL,
  nomb_dependencia character varying(255),
  CONSTRAINT "DependenciaPK" PRIMARY KEY (id_dependencia)
)
WITH (OIDS=FALSE);

ALTER TABLE "DEPENDENCIA"
  OWNER TO postgres;

CREATE TABLE "SUBDEPENDENCIA"
( id_dependencia numeric NOT NULL,
  id_subdependencia numeric NOT NULL,
  nomb_subdependencia character varying(255),
  CONSTRAINT "SubdepPK" PRIMARY KEY (id_dependencia, id_subdependencia),
  CONSTRAINT "idDependenciaFK" FOREIGN KEY (id_dependencia)
      REFERENCES "DEPENDENCIA" (id_dependencia) MATCH SIMPLE
      ON UPDATE CASCADE ON DELETE CASCADE
)
WITH (OIDS=FALSE);
ALTER TABLE "SUBDEPENDENCIA"
 OWNER TO postgres;

CREATE TABLE "AREA"
(
  id_dependencia numeric NOT NULL,
  id_subdependencia numeric NOT NULL,
  id_area numeric NOT NULL,
  nomb_area character varying(255),
  CONSTRAINT "areaPK" PRIMARY KEY (id_dependencia, id_subdependencia, id_area),
  CONSTRAINT "IdSubdependenciaFK" FOREIGN KEY (id_dependencia, id_subdependencia)
      REFERENCES "SUBDEPENDENCIA" (id_dependencia, id_subdependencia) MATCH SIMPLE
      ON UPDATE NO ACTION ON DELETE SET NULL
)
WITH (OIDS=FALSE);

ALTER TABLE "AREA"
  OWNER TO postgres;
```

Figura 47. Definición de datos

En la figura 47 se observa la definición de datos con las restricciones de dominio y la integridad referencial sobre las operaciones de eliminación y actualización mediante las llaves foráneas. Asimismo, se indican pequeños fragmentos de lenguaje DCL para el control de usuarios.

Actividad 3

Como parte de la generación de material de aprendizaje en la organización se investigaron las características de la herramienta Tableau como herramienta de técnica auxiliar de la Minería de datos (Reporting).

La interfaz permite colocar campos o dimensiones de los datos en el eje de los renglones o de las columnas e ir incrustando más dimensiones para formar sub-categorías. Por otra parte, se colocan medidas o datos numéricos en el eje complementario o en el mismo eje en el que se colocaron las dimensiones. Finalmente se obtiene un gráfico sobre la información del negocio con un gran atractivo visual, detalle e interactividad para generar reportes o historias sobre los datos.

Esta herramienta presenta problemas de compatibilidad entre versiones de diferentes tipos de productos. También está limitado por los permisos de licencia para su implementación en una aplicación web; empero, hay una versión gratuita que permite incrustar las visualizaciones en páginas web. La versión gratuita es muy útil pero presenta limitantes en la confidencialidad de los datos. La figura 48 muestra la vista incrustada en una página web.

Además, se investigó sobre la gama de productos de Tableau y sus limitantes para su implementación en el sistema.

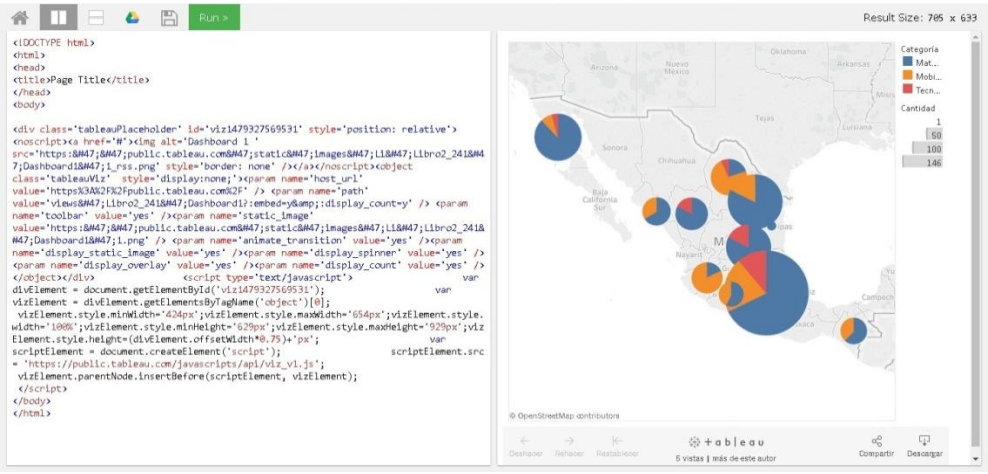


Figura 48. Vista incrustada en página web

Se investigó software alternativo para la visualización de datos en las aplicaciones web y se destacan bibliotecas basadas en Javascript, de hecho Tableau se basa en JavaScript. Pero estas herramientas requieren de aprendizaje, tiempo de desarrollo porque son lenguajes declarativos de visualización de datos, se sugirió que sean utilizadas en conjunto con las herramientas de visualización de los entornos Knime y Orange 3 para agilizar la programación y reducir el tiempo de desarrollo.

Por otra parte se encontró un producto de licencia libre similar a Tableau, se trata de Weave[19]. Weave presenta la misma arquitectura que Tableau porque requiere de un servidor para ver las gráficas generadas en una página web con diseño responsivo.

[19] http://www.iweave.com/

Capítulo VI Conclusiones

Las actividades desarrolladas durante el servicio social permitieron aportar ideas de análisis para el proyecto "Sistema de Información Universitaria" (SIU) en el que se utilizaran herramientas de Big Data que permitirán tener un control sobre el cúmulo de datos académicos sobre el aprendizaje y analizarlos conjuntamente con otros factores sociales que puedan tener efecto en el aprendizaje o el rendimiento de los alumnos. Sin embargo, los ejemplos generados parten de la teoría clásica de Minería de datos y de hecho las conexiones se realizaron a sistemas tipo OLTP con bases de datos relacionales, este hecho no influye en la etapa de análisis porque se puede cambiar la fuente de datos y las técnicas de extracción por las presentes en Big Data.

En cuanto a herramientas, la herramienta más recomendable es la de Knime por presentar menos errores y muchas capacidades de pre-procesamiento o de análisis. La herramienta Knime incorpora nodos certificados en Big Data por lo que es un punto a favor para esta herramienta. En comparativa con Python, Knime puede complementar muy bien recursos de análisis como el procesamiento de lenguaje natural de la universidad de Stanford, en Python, también puede aportar mucho al desarrollo de componentes por la programación orientada a objetos ya sea mediante scripts en Knime, en Orange 3 o de módulos con extensión "py".

Los problemas técnicos documentados permitieron tener una base para la creación de la lista de riesgos.

El diseño de bases de datos para capturar requerimientos permitió desarrollar habilidades técnicas de programación para la normalización de tablas, así mismo el ejemplo de estructuración de la ley de transparencia permitió el desarrollo de estas habilidades.

Por otra parte, se aportaron ejemplos del uso de herramientas para la síntesis de información para la planificación de contenidos en diversos formatos de acuerdo a la experiencia de aprendizaje plasmada en las métricas analizadas del sistema RUA.

Además, se presentó información útil para tomar decisiones, por ejemplo: de tener un formato de plan de estudios general para todas las dependencias de la UNAM con el fin de realizar Minería de texto y obtener relaciones para el aprovechamiento escolar o la aplicación de análisis de texto después de un proceso de digitalización de documentos. Asimismo, se obtuvieron análisis de ayuda para la planeación o justificación en la creación de otros recursos o proyectos educativos,

por ejemplo la creación de videotecas con contenido disponible para recién egresados o público en general o alumnos con el fin de tener el conocimiento sintetizado para su estudio como complemento a las clases ordinarias en las que se profundiza más sobre los contenidos y de complemento a la enseñanza tradicional.

Se logró participar en la implementación de estrategias para el desarrollo y seguimiento de sistemas avanzados mediante la documentación entregada sobre la caracterización del software designado. Aunque la designación en principio fue sobre los entornos de Tableau y Python, se me permitió la flexibilidad indagar sobre las herramientas conocidas ya de antemano por mi formación en la Facultad de Ingeniería, esto me permitió enriquecer más el conocimiento sobre las herramientas y las estrategias para complementar los instrumentos de análisis de datos.

Para la formación profesional, fue una experiencia muy enriquecedora porque se desarrollaron habilidades, estrategias de investigación sobre software en otros idiomas, de generación de documentación técnica, de documentación para la transmisión del conocimiento (al emplear estrategias para superar la curva de aprendizaje para el equipo de análisis de datos en el proyecto SIU). También se consolidó el aprendizaje teórico de las ciencias básicas, ciencias de la ingeniería y ciencias aplicadas de la carrera de Ingeniería en Computación al identificar el marco conceptual en los procesos de desarrollo de software en el ambiente de trabajo en la División de Colaboración y Vinculación. Además, se logró identificar las áreas de mejora para el desarrollo de un proyecto de software desde la infraestructura hasta la logística.

El servicio social aportó a mi formación profesional la creación de un portafolio de análisis de datos, presentación de informes estadísticos, organización de la información en faces de proceso, desarrollo de habilidades cognitivas, de técnicas de programación y de identificación de áreas de oportunidad para el diseño en arquitecturas de inteligencia de negocios. Así también, para el desarrollo de pruebas y de aplicaciones involucradas con sistemas de análisis y de negocios. Además, se identificaron los problemas técnicos y el posible efecto de aquellos en el ciclo de vida del software.

La participación en el servicio social tuvo impacto indirecto por actuar sobre la línea de aprendizaje de la organización y de apoyo en el procesamiento de requerimientos mediante herramientas de software, siendo los grupos principalmente beneficiados: niños, jóvenes y adultos mayores. Asimismo, este documento tiene la intención de compartir mi experiencia con compañeros y profesores para enriquecer la crítica en la senda del conocimiento.

De lo anterior se destaca la actitud de servicio, de compromiso y responsabilidad desarrollada e inculcada en mí formación por la Universidad Nacional Autónoma de México y la Facultad de Ingeniería.

Bibliografía

Antonio, Moreno y Teófilo, Redondo. *Text Analytics: the convergence of Big Data and Artificial Intelligence.* Madrid : s.n. IEEE.

Arup Nielsen, Finn. 2015. *Data Mining with Python (Working draft).* 2015.

Biolab. 2017. *Orange3 Text Mining Documentation.* 2017. págs. 86.

Bird, Steven y otros. 2008. *Natural Language Processing in Python.* s.l. : O'Reilly, 2008. págs. 358.

Cake Software Fundation. 2016. *CakePHP Cookbook Documentation.* Release 2.x. 2016. págs. 859.

Ciocca, Damian. 2008. Arquitectura de Software y Entornos de Trabajo (Frameworks) y Contenedores Ligeros. [ed.] Javier Blanqué. Buenos Aires : Universidad Nacional De Luján, 2008, 1.

Crispin, Lisa. 2015. *A look at Canoo WebTest.* [pdf] s.l. : Software, Better, Noviembre de 2015.

Crispin, Lisa y Gregory, Janet. 2008. *Agile Testing: A Practical Guide for Testers and agile teams.* Crawfordsville, Indiana : Pearson Education, 2008. ISBN-13: 978-0-321-53446-0.

Downey, Allen B. 2012. *Think Bayes.* 1. Needham, Massachusetts : Green Tea Press, 2012. págs. 3-10, 126-128.

Downey, B. Allen. 2014. *Think Stats, Exploratory data analysis in Python.* 2. Needham, Massachusetts : Green Tea Press, 2014. págs. 15,16,29,58.

Elmasri, Ramez y B. Navathe, Shamkant. 2010. *Fundamentals of database systems.* Sexta. s.l. : Addison-Wesley, 2010. ISBN-13: 978-0-136-08620-8.

Gephi: an open source software for exploring and manipulating networks. **M., Bastian y S., Heymann. 2009.** International AAAI Conference on Weblogs and Social Media.

Goutam, Chakraborty y Krishna, Pagolu M. 2014. *Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining.* 2014.

Hashimoto, Mitchell. 2013. *Vagrant: Up and running.* s.l. : O'Reilly, 2013. págs. 155.

IBM Corporation y Saïd Business School. 2012. *Analytics: el uso de big data en el mundo real.* Global Bussiness Services, Business Values. Madrid, España : s.n., 2012. pág. 22, Ejecutivo.

Jiawei, Han y Micheline, Kamber. 2006. *Data Mining: Concepts and Techniques.* Segunda. San Fancisco, CA : Elsevier Inc, 2006. ISBN 10: 1-55860-901-6.

John, Hunter y otros. 2017. *Matplotlib.* 2017. pág. 3120. Release 2.0.0.

Larose, Daniel T. 2005. *Discovering knowledge in data. An introduction to Data Mining.* New Jersey : John Wiley & Sons, 2005. págs. 6-24.

Mausam. *Document Similarity in Information Retrieval.* [pdf]

Meta, Brown S. 2014. *Data mining For Dummies.* New Jersey : Jhon Wiley and Sons, 2014. ISBN 978-1-118-89316-6.

Montes y Gómez, Manuel. *Minería de texto: Un nuevo reto computacional.* México : CIC, IPN. Investigación.

Mora, Sergio Luján. 2002. *Programación de aplicaciones web: historia, principios básicos y clientes web.* [pdf] SanVicente, Alicante, España : Club Universitario, 2002.

Neeraj, Sharma y otros. 2010. *Database fundamentals.* Markham,ON : IBM Cánada , 2010.

Novak, David. 2016. Knjiznica za analizo besedilnih dokumentov v programskem okolju. s.l. : Univerza v Ljubljani, 2016, págs. 41-65.

Oscar, Marbán y otros. 2009. *A Data Mining & Knowledge Discovery Process Model, Data Mining and Knowledge Discovery in Real.* [ed.] Julio Poce y Karahoca Adem. s.l. : InTech, 2009. ISBN: 978-3-902613-53-0.

Pérez L., César. *Minería de datos: técnicas y herramientas.* Paraninfo, 2008

Pérez M., María.2014. *Minería de datos a trávez de ejemplos.* Madrid : RC Libros, 2014. ISBN: 978-84-941801-4-9.

Rincón, Luis. 2014. *Introducción a la probabilidad.* s.l. : Facultad de Ciencias UNAM, 2014.

Sumathi, S. y Sivanandam, S.N. 2006. *Introduction to Data Mining and its Applications.* [ed.] Janusz Kacprzyk. s.l. : Springer, 2006. págs. 151-163. Vol. 29. ISBN-10 3-540-34350-4.

Thiel, Killian y Berthold, Michael. 2012. *The KNIME Text Processing Feature: An Introduction.* s.l. : Knime AG, 2012. pág. 15, Reporte Técnico. Revision: 120403F.

Universidad de Alicante. 2006. rua.es. [En línea] 2006. [Citado el: 7 de Diciembre de 2016.]

Varoquax, Gaël y otros. 2015. *Scipy. Lecture Notes.* [pdf] [ed.] Gaël et al. Varoquaux. 2015.

Vincent, Rainardi. 2008. *Building a Data Warehouse: With Examples in SQL Server.* New York : Apress, 2008. ISBN-13: 978-1-4302-0527-2.

Wackerly, Dennis y otros. 2008. *Estadística matemática con aplicaciones.* Séptima. s.l. : Cengage Learning Editores, 2008. págs. 1-79,488-556. ISBN-13: 978-607-481-399-9.

Wes McKinney & PyData Development Team. 2016. *pandas: powerful Python data analysis toolkit.* [pdf] 2016. págs. 1937. Release 0.19.0.

Zacharski, Ron. *A Programmer's Guide to Data Mining: The Ancient Art of the Numerati.*

Mesografía

Brooks, Christopher. *Introduction to Data Science in Python.* s.l., Michigan, USA : Coursera.

Javier, Gutiérrez J. ¿Qué es un framework web? [En línea] [Citado el: 14 de Enero de 2017.] http://www.lsi.us.es/~javierj/investigacion_ficheros/Framework.pdf.

nebul4ck. virtualizacion gnulinux viviii virtualizar con cygwin. [En línea] [Citado el: 6 de Febrero de 2017.] https://nebul4ck.wordpress.com/2015/12/04/virtualizacion-gnulinux-viviii-virtualizar-con-cygwin/.

Kent Beck, Mike Beedle y otros.2001. agilemanifesto. [En línea] 2001. http://agilemanifesto.org/.

Koninklijke Bibliotheek. 2017. What is emulation? [En línea] 2017. [Citado el: 5 de Febreo de 2017.] https://www.kb.nl/en/organisation/research-expertise/research-on-digitisation-and-digital-preservation/emulation/what-is-emulation.

Rekha, Sree. 2017. talend. [En línea] 10 de Marzo de 2017. [Citado el: 30 de Marzo de 2017.] https://www.talend.com/blog/2017/03/10/unlocking-data-preparation-business-intelligence-bi/?utm_medium=socialpost&utm_source=linkedin&utm_campaign=blog.

Samper, J. Javier. 2009. Informatica.uv.es. [En línea] 2009. [Citado el: 14 de Noviembre de 2016.] http://informatica.uv.es/iiguia/IST/Tema1.pdf.

Victor, Carceler. Institute Puig Castellar. [En línea] [Citado el: 16 de Noviembre de 2016.] https://elpuig.xeill.net/Members/vcarceler/asix-m09/uf1/nf1/a5. A5.

Vmware. 2017. Virtualización de vmware. [En línea] 2017. [Citado el: 16 de Febrero de 2017.] http://www.vmware.com/latam/solutions/virtualization.html.

Anexos

A Instalación y configuración de herramientas para implantar ambientes virtuales para el desarrollo de software

#1 Descargué el archivo ejecutable de instalación vagrant_1.8.5.exe desde la url: https://www.vagrantup.com/downloads.html en el directorio C:\Users\boromir\Downloads

#2 Descargué el archivo ejecutable de instalación VirtualBox-5.1.6-110634-Win.exe desde la página: https://www.virtualbox.org/wiki/Downloads en el directorio C:\Users\boromir\Downloads

#3.Instalé Virtual Box 5.1.6 para ambiente Windows (32 bits) en la ruta:
C:\Program Files\Oracle\Virtual Box, mediante el procedimiento estándar de instalación en Windows

#4.Instalé Vagrant 1.8.5 (32 bits) en la ruta:
C:\HashiCorp\Vagrant, mediante el procedimiento estándar de instalación en Windows y reinicié el equipo

#5 Agregué la ruta C:\Program Files\Oracle\Virtual Box y la ruta C:\HashiCorp\Vagrant\bin a la variable de entorno PATH del sistema en Windows

---

Actual:
PATH= $PATH;
C:\HashiCorp\Vagrant\bin;
C:\Program Files\Oracle\Virtual Box

#6 Descargué de la URL: https://cygwin.com/install.html el instalador de cygwin

#7 instalé el paquete openssh de cygwin del sitio http://cygwin.mirror.constant.com en la ruta C:\cygwin para utilizar comandos de shell Linux en Windows para acceder a la máquina virtual
#8 Agregué la ruta C:\cygwin\bin a la variable de entorno PATH

Actual
PATH= $PATH;
C:\cygwin\bin;

CREACIÓN DEL AMBIENTE VIRTUAL

Modo estándar

#9 Construí el directorio C:\Users\boromir\Documents\Cajas Vagrant\Ubuntu para alojar el proyecto Vagrant
#10 Ejecuté el comando - vagrant init ubuntu/trusty32 - en el directorio creado y se alojaron los archivos

del proyecto automáticamente
#11 Ejecuté el comando - vagrant up - para descargar la caja base desde la url: https://atlas.hashicorp.com/ubuntu/trusty32 e iniciar la máquina virtual
Linux.

Modo manual

También es posible implantar cajas vagrant desde máquinas virtuales, para esto, es necesario agregar al usuario vagrant con contraseña vagrant y agregarlo al grupo de superusuario mediante el comando visudo y verificar el uso de llaves de seguridad pública y privada para la conexión mediante ssh

CONFIGURACIONES

#15 ------ Especificación de red ----------

ip host Windows: 132.248.63.39/24 , 192.168.56.1/24 , 127.0.0.1/32

En NAT
ip host windows: 10.0.2.2/24
ip guest ubuntu: 10.0.2.15/24

CONFIGURACIÓN DE PUERTOS

Configuré el reenvió de datos por medio de puertos entre el huésped (host) y el anfitrión (guest) para acceder a servicios de ubuntu desde aplicaciones en Windows mediante puertos especificados en las instrucciones añadidas en el archivo Vagrantfile:

config.vm.network "forwarded_port", guest: 80, host: 8080 //Servicios apache
config.vm.network "forwarded_port", guest: 5432, host: 65432 //Servicios Postgresql
config.vm.network "forwarded_port", guest: 22, host: 2222 //SSH

INICIO DE SESIÓN REMOTA

#12 Ejecuté el comando - vagrant ssh - para acceder a la máquina virtual mediante el usuario vagrant //túnel host 10.0.2.2/24 puerto: 2222 a guest 10.0.2.15/24 puerto:22
#13 Finalicé la sesión en ubuntu mediante la combinación de teclas ctrl-d
#14 Ejecuté el comando - vagrant halt - para apagar la máquina virtual

B Instalación y configuración de cup cake php

INSTALACIÓN DE SOFTWARE NECESARIO PARA CUPCAKEPHP

#1 Instalé apache, php y postgresql mediante los comandos en la máquina virtual ubuntu/trusty32:

```
sudo apt-get install apache2
sudo apt-get install php5
sudo apt-get install postgresql
```

#2 Conecté PostgreSQL con Apache y PHP mediante los comandos

```
sudo apt-get install php5-pgsql
sudo apt-get install libapache2-mod-auth-pgsql
```

COMPONENTES NECESARIOS PARA ADMINISTRACIÓN DE BASES DE DATOS

#3 Instalé postgresql en la ruta "C:\Program Files\PostgreSQL\9.3" del host Windows
usuario : postgresql
contraseña: vapp
puerto:5432
configuración regional: spanish,Mexico

Configuración de conexión de postgresql y apache entre Windows y ubuntu
---

#4 Configuré el reenvió de paquetes a través de los puertos de escucha de las aplicaciones pgadminIII y navegador web en el archivo VagrantFile

```
config.vm.network "forwarded_port", guest: 80, host: 8080
```

//Conexión a servidor Apache puerto 80 (guest ubuntu) 10.0.2.15 a través de puerto 8080 (host Windows) //10.0.0.2 nat ~ localhost

```
config.vm.network "forwarded_port", guest: 5432, host: 65432
```

#5 Cambié el password del usuario postgres creado en la instalación de postgresql mediante
sudo passwd postgres
password: postgres

#6 Configuré las conexiones remotas aceptadas para el servicio postgresl mediante las instrucciones correspondientes a los archivos de configuración siguientes:
/etc/postgresql/9.3/main

--> postgresaql.conf /*Agregar o des comentar las siguientes dos líneas*/

```
listen_addresses = '*'
password_encryption = on
```

-->pg_hba.conf

```
host    all         all         127.0.0.1/32          md5
host    all         all          10.0.2.2/24          md5    //filtro de ips que acceden a postgresql
```

#7 Configuré la conexión a la base de datos desde pgAdminIII en Windows con los siguientes parámetros:
Name: Ubuntu
host: localhost

port: 65432
Username: postgres
password: postgres

INSTALACIÓN DEL FRAMEWORK CUPCAKE MODO DESARROLLO

#8 En el directorio /var/www/html/ ejecuté los comandos

#paso 1: descarga del framework

sudo wget https://codeload.github.com/cakephp/cakephp/legacy.zip/2.5.2

#paso 2: descomprimir el directorio mediante la aplicación unzip (sudo apt.get install unzip)
sudo unzip 2.5.2

#paso 3:renombrar el directorio
/var/www/html/

sudo mv cake* cake_2_5_2

#paso 4 ubicarse en el nuevo folder y cambiar permisos/propietarios
/var/www/html/cake_2_5_2

sudo chown -R root:www-data app/tmp
sudo chmod -R 775 app/tmp

#paso 5 permitir el modo overwrite para el archivo .htacces desde la configuración en apache2

#5.1 permitir módulo rewrite
sudo a2rnmod rewrite

#5.2 alojar overwrite al directorio /var/www
/etc/apache2/apache2.conf

```
<Directory /var/www>
        Option Indexes FollowSymlinks
        AllowOverride All
        Required all granted
</Directory>
```

#Cargar modulo desde archivo de configuración para asegurarse de que el modulo esté cargado
LoadModule rewrite_module libexec/apache2/mod_rewrite.so

#y en el archivo /etc/apache2/sites-available/000-default.conf

```
<Directory />
   Options FollowSymLinks
   AllowOverride All
</Directory>
<Directory /var/www>
   Options Indexes FollowSymLinks MultiViews
   AllowOverride All
   Order Allow,Deny
   Allow from all
</Directory>
```

#5.2 reinicié el servicio apache2
sudo service apache2 restart

#5.3 modifiqué los método de seguridad hashing en el directorio
/var/www/html/cake_2_5_2/app/Config/core.php

```
 //Configure::write('Security.salt', 'DYhG93b0qyJfIxfs2guVoUubWwvniR2G0FgaC9mi'); anterior
   Configure::write('Security.salt', 'GGhG93b0qyJfIxfs2guVo22bWwvniR2G0FgaC9mi');

 //Configure::write('Security.cipherSeed', '76859309657453542496749683645');      anterior
   Configure::write('Security.cipherSeed', '998593096574535400967496836 45');
```

#5.4 modifiqué los parámetros para el uso de la base de datos postgresql en cakephp en el archivo
#Construí la base de datos BlogCakePHP con la tabla posts para la aplicación ejemplo blogposts

/var/www/html/cake_2_5_2/app/Config/database.php

```
class DATABASE_CONFIG {

	public $default = array(
		'datasource' => 'Database/Postgres',
		'persistent' => false,
		'host' => 'localhost',
		'login' => 'postgres',
		'password' => 'postgres',
		'database' => 'BlogCakePHP',
		'prefix' => '',
		'encoding' => 'utf8',
	);
```

#5.5 verifiqué el contenido de los archivos .htacces de cakePHP 2.5.2

```
vagrant@vagrant-ubuntu-trusty-32:/var/www/html/cake_2_5_2/app$ cat .hta*
<IfModule mod_rewrite.c>
   RewriteEngine on
   RewriteRule    ^$    webroot/    [L]
   RewriteRule    (.*) webroot/$1    [L]
</IfModule>

vagrant@vagrant-ubuntu-trusty-32:/var/www/html/cake_2_5_2/app$ cat webroot/.hta*
<IfModule mod_rewrite.c>
   RewriteEngine On
   RewriteCond %{REQUEST_FILENAME} !-d
   RewriteCond %{REQUEST_FILENAME} !-f
   RewriteRule ^(.*)$ index.php [QSA,L]
  # RewriteRule ^ index.php [L]        ---Anterior
</IfModule>


vagrant@vagrant-ubuntu-trusty-32:/var/www/html/cake_2_5_2$ cat .hta*
<IfModule mod_rewrite.c>
  RewriteEngine on
  RewriteRule    ^$ app/webroot/    [L]
  RewriteRule    (.*) app/webroot/$1 [L]
</IfModule>vagrant@vagrant-ubuntu-trusty-32:/var/www/html/cake_2_5_2$
```

#6 Establecí conexión con la aplicación de inicio del framework mediante el navegador con url:
http://localhost:8080/cake_2_5_2

INSTALACIÓN DEL PLUGIN DEBUG

#7 Descargué el paquete de depuración en el directorio /var/www/html/cake/app/Plugin mediante
sudo wget https://github.com/cakephp/debug_kit/zipball/2.2

#8 Habilité el plug-in de depuración 2.2 en el archivo app/Config/bootstrap.php linea 70

#9 Visualicé el contenido de la página principal del framework cupcakePHP en el navegador host en la url:
localhost:8080/cake_2_5_2

C Codificación XML de la prueba con Web test Canoo

Build.xml | allTests.xml | EliminarPosts.xml | AgregarPosts.xml | ContenidoPosts.xml | EditarPosts.xml

```xml
<project default="wt.full">
    <property name="webtest.home" location="C:/WebTest"/><!-- ubicación base de instalación webtest-->
    <import file="${webtest.home}/webtest.xml"/><!-- importación de archivo de construcción general-->

    <target name = "wt.testInWork"><!-- Contenedor de archivos a refrencia de pasos de prueba-->
        <ant dir = "tests" antfile = "allTests.xml" />
    </target>
</project>
```

Build.xml | allTests.xml | EliminarPosts.xml | AgregarPosts.xml

```xml
<project default = "all">
    <target name = "all">

    <dataDriven tableContainer="agregar.xls">
        <ant antfile = "AgregarPosts.xml"/>
    </dataDriven>

    <dataDriven tableContainer="agregar.xls">
        <ant antfile = "ContenidoPosts.xml"/>
    </dataDriven>

    <dataDriven tableContainer="editar.xls">
        <ant antfile = "EditarPosts.xml"/>
    </dataDriven>

    <dataDriven tableContainer="editar.xls">
        <ant antfile = "ContenidoPosts.xml"/>
    </dataDriven>

    <dataDriven tableContainer="eliminar.xls">
        <ant antfile = "EliminarPosts.xml"/>
    </dataDriven>

    </target>
</project>
```

Build.xml | allTests.xml | AgregarPosts.xml | ContenidoPosts.xml | EditarPosts.xml | Elimina

```
<!DOCTYPE project SYSTEM "../dtd/Project.dtd">

<project default = "test">
  <target name = "test">


  <webtest name="Agregar ${num} ${Titulo} ${Contenido}">
    <steps>
    &goToAdd;
            <verifyTitle
             text = "CakePHP: the rapid development php framework: Posts"
             description = "Debería verse el titulo de la página"
            />
<group>
    <setInputField htmlId="PostTitle" value="${Titulo}"/>
    <setInputField htmlId="PostBody" value="${Contenido}"/>

    <clickButton description="Submit post" name="Save Post"/>
</group>

  </steps>
  </webtest>

  </target>
</project>
```

Build.xml | allTests.xml | AgregarPosts.xml | ContenidoPosts.xml | EditarPosts.xml | Elimina

```
<!DOCTYPE project SYSTEM "../dtd/Project.dtd">

<project default = "test">
  <target name = "test">


  <webtest name="Contenido post ${num}: ${Titulo} ${Contenido}">
    <steps>
 <invoke
             url = "http://localhost:8080/cake_2_5_2/posts/view/${num}"
             description = "inicio posts"
 />

            <verifyTitle
             text = "CakePHP: the rapid development php framework: Posts"
             description = "Debería verse el titulo de la página"
            />
<group>
    <verifyElementText htmlId="PostTitle" text="${Titulo}"/>
    <verifyElementText htmlId="PostBody" text="${Contenido}"/>
</group>

  </steps>
  </webtest>

  </target>
</project>
```

Build.xml | allTests.xml | AgregarPosts.xml | ContenidoPosts.xml | EditarPosts.xml | EliminarPo

```
1  <!DOCTYPE project SYSTEM "../dtd/Project.dtd">
2
3  <project default = "test">
4    <target name = "test">
5      <webtest name = "Eliminar ${num}">
6        <steps>
7          &goToInicio;
8
9
10         <clickLink description="Delete Post" href="#" />
11
12         <expectDialog description="Eliminar Post numero x"
13         dialogType="confirm" response="true" saveProperty="DeleteDialog" />
14
15
16
17       </steps>
18     </webtest>
19
20   </target>
21 </project>
```

D Instrucciones de instalación de Software analítico

WEKA

#1 descargué el software de instalación de WEKA de la url:
http://prdownloads.sourceforge.net/weka/weka-3-8-1.exe

#2 instalé WEKA

#3 Descargué el driver : JDBC3 Postgresql Driver, Versión 9.3-1103
URL:https://jdbc.postgresql.org/download/postgresql-9.3-1103.jdbc3.jar

#4 Instalé desde WEKA el paquete: JDBCDriversDummyPackage
Tools->Package Manager->JDBCDriversDummyPackage... install

#5 Extraje el archivo DatabaseUtils.props.postgresql del directorio C:\Program Files\Weka-3-8\weka.jar\weka\experiment
en el directorio
C:\Users\boromir\wekafiles\props

#6 Cambié el nombre del archivo DatabaseUtils.props.postgresql por DatabaseUtils.props

#7 coloqué el driver postgresql-9.3-1103.jdbc3 en el directorio
C:\Users\boromir\wekafiles\packages\JDBCDriversDummyPackage\lib
en caso de error ubicar el controlador en el directorio de instalación de weka

#8 Conecté con la base de datos en postgresql
mediante la instrucción
jdbc:postgresql://localhost:5432/RUA?user=postgres&password=postgres

Referencias
.....
https://weka.wikispaces.com/Databases

ANACONDA

#1Anaconda

paquete: https://repo.continuum.io/archive/Anaconda3-4.1.1-Windows-x86_64.exe
//versión recomendada para evitar problemas de compatibilidad con QT en la instalación de librerías
Ruta instalación: D:/adolfo/anaconda3411

#1
conda install pip

#2 Crea Entorno virtual python
cmd > conda create -n mineria

#3 Activación entorno virtual
cmd >activate mineria

#4instalación orange3 #librerías disponibles pandas,scipy,numpy,scikitlearn,nltk,
cmd (mineria) > pip install orange3

#5instalación driver postgres para Orange
cmd (mineria) > pip install psycopg2

#6 Correr orange
cmd (mineria) > orange-canvas ## || python -m Orange.canvas

#desactivar entorno virtual
cmd(mineria) > deactivate Minería

2UDA

#3 2UDA (OJO instalar sólo postgres por problemas con orange-instalado desde conda-)
Paquete: http://packages.2ndquadrant.com/2UDA/2UDA-9.6-windows-installer.exe

##Opciones de instalación

Dir.instalación : D:/Adolfo/2UDA
Dir. postgres : D:/Adolfo/Postgres96
puerto :5432

##Varibles de entorno
D:\Adolfo\2UDA\PostgreSQL-9.6\bin

##psql -h localhost -U postgres
##

leeme
""
""
Welcome to 2UDA - 2ndQuadrant's Unified Data Analytics platform.

The following lists defaults used as part of the installation process. If you did not change any values during 2UDA's install, the following list should hold true on your computer.

===============================
Connecting Orange to PostgreSQL
===============================

You can connect Orange to PostgreSQL by following these simple steps:

1) Drag and drop the 'SQL Table' widget on the canvas
2) Double click on this widget
3) Provide database credentials in the following order:
   i) Hostname (default: localhost)
   ii) Database name (default: sample)
   iii) User name (default: postgres)
   iV) Password
4) Click the refresh button next to the table dropdown menu
5) Select a table from the list or write a custom SQL, then close this dialog

6) Drag and drop any of the data analysis widgets provided in the left pane
7) Connect 'SQL table' to the analysis widget
8) Double click the analysis widget
9) Don't forget to have fun!

****
IMPORTANT: Some of the widgets in Orange require the 'tsm_system_time' and 'quantile' extensions to be available in the PostgreSQL database. The 'sample' database created at install time already has these extensions. If you create a new database, however, you will need to create these extensions to be able to use all Orange widgets. The extensions can be created by running the following simple commands on the database you have created:
****

CREATE EXTENSION tsm_system_time;
CREATE EXTENSION quantile;
"""
"""

E Extracción, Limpieza y Transformación de texto mediante Python

#Adolfo Herrera Arias, 27/Marzo/2017 , versión.1, (Servicio Social)

#descripción: Este programa carga un archivo pdf de la ley general de transparencia con vigencia a la fecha actual; pasa la estructura de la Ley a una tabla no normalizada para obtener frecuencia, distancia de las palabras más significativas del texto

```
import pandas as pd                #biblioteca para tramas de datos
import PyPDF2 as pyPdf             #'' para procesamiento de pdf
import re                          #'' para manejo de expresiones regulares
```

In [2]:

```
def getPDFContent(x):              #fuente stack overflow
    content = ""
    # Load PDF into pyPDF
    pdf = pyPdf.PdfFileReader(open(x, "rb"))
    # Iterate pages
    for i in range(0, pdf.getNumPages()):
        # Extract text from page and add to content
        content += pdf.getPage(i).extractText() + "\n"
    # Collapse whitespace
    #content = " ".join(content.replace("\xa0", " ").strip().split())
    return content

#limpieza de caracteres
txt=getPDFContent('LGTAIP__Coloreado.pdf')  #Paso de pdf a cadena en bytes
txt=txt.encode('latin-1').replace(b'\n',b'') #Codificación latin-1, sustituye '\n'
txt=txt.decode('latin-1','ignore')         #Decodificación latin-1
txt=re.sub(r'\s+',r' ',txt)              #Normalización de espacios en blanco
```

In [3]:

```
txt=re.sub(r'(Artículo\s[0-9]+o?\.)',r'%\1',txt)
#Caracter de separación como infijo para procesamiento en tablas
txt=re.sub(r'([I V X L]+\.)',r'%\1',txt)                    #Idem
txt=re.sub(r'(TÍTULO\s*\w*)',r'%\1\n',txt)                  #Idem
txt=re.sub(r'(\.\s|[A-Z]\s+)(Capítulo\s+[I V X]+)',r'\1%#\2',txt) #Idem
txt=re.sub(r'(Transitorios)',r'%\1',txt)                    #...
txt=re.sub(r'(Sección\s*\w*)',r'%\1',txt)
txt=re.sub(r'([a-z]\))',r'%\1',txt)
txt=re.sub(r'([.;\n:]|,\sy|,\so)(\s)([0-9]+\.)',r'\1\2%\3',txt)
txt=re.sub(r'\t',r' ',txt)          #Normalización de caracteres \s del texto
txt=re.sub(r'\n',r'',txt)
txt=re.sub(r'%',r'\n',txt)
```

```
encabezado='"LEY GENERAL DE TRANSPARENCIA Y ACCESO A LA INFORMACIÓN PÚBLICA CÁMARA DE DIPUTADOS DEL H. C
ONGRESO DE LA UNIÓN Secretaría General Secretaría de Servicios Parlamentarios Nueva Ley DOF 04-05-2015 [0-9]+ de [0-9]+"'
encabezado=re.compile(encabezado)       #Compila Expresión regular con texto no significativo
txt=re.sub(encabezado,r'',txt)          #Sustitución del texto no significativo
txt='texto'+'\n'+txt
```

In [4]:

```
#print(txt)
```

In [5]:

```
import sys,csv
from io import StringIO  #from StringIO import StringIO version<3
TESTDATA=StringIO(txt)   #DataFrame <- Flujo de datos tipo String

df = pd.read_csv(TESTDATA, sep="\n",encoding='utf-8',quoting=csv.QUOTE_NONE)
```

In [6]:

```
df=df.iloc[1:,:]       #Filtrado
#df
```

In [7]:

```
#Separación por columnas, tabla no normalizada

df['titulo']=[str(x) if 'TÍTULO' in x else '' for x in df['texto']]
df['capitulo']=[str(x) if '#Capítulo' in x else '' for x in df['texto']]
df['seccion']=[str(x) if 'Sección' in x else '' for x in df['texto']]
df['articulo']=[re.search(r'Artículo\s[0-9]+\.',x).group(0) if bool(re.search('Artículo\s[0-9]+\.', x)) else '' for x in df['texto']]
df['apartado']=[re.search(r'[A-Z]\.\s',x).group(0) if bool(re.search(r'[A-Z]+\s\.',x)) else '' for x in df['texto']]
df['fraccion']=[re.search(r'[I V X L C M]+\.\s',x).group(0) if bool(re.search('[I V X L C M]+\.\s', x)) else '' for x in df['texto']]
df['inciso']  =[re.search(r'[a-z]\)\s',x).group(0) if bool(re.search('[a-z]\)\s', x)) else '' for x in df['texto']]
df['numeral'] =[re.search(r'^\s*[0-9]+\.',x).group(0) if bool(re.search(r'^\s*[0-9]+\.', x)) else '' for x in df['texto']]
df['texto' ]  =[re.sub(r'Artículo\s[0-9]+\.|[I V X L C M]+\.\s|[a-z]\)\s|^[0-9]+\.','',x)if bool(re.search(r'Artículo\s[0-9]+\.|[I V X L C M]+\.|[a-z]\)|^[0-9
]+\.', x)) else x for x in df['texto']]

#Limpieza de palabras reservadas
df['capitulo']=[str(x) if 'Capítulo' in x else '' for x in df['capitulo']]
df['seccion']=[str(x) if 'Sección' in x else '' for x in df['seccion']]
df['apartado']=[re.search(r'[A-Z]+',x).group(0) if bool(re.search(r'[A-Z]+\s\.',x)) else '' for x in df['apartado']]
df['articulo']=[re.search(r'[0-9]+',x).group(0) if bool(re.search('Artículo\s[0-9]+\.', x)) else '' for x in df['articulo']]
df['fraccion']=[re.search(r'[I V X L C M]+',x).group(0) if bool(re.search('[I V X L C M]+\.\s', x)) else '' for x in df['fraccion']]
df['inciso']  =[re.search(r'[a-z]',x).group(0) if bool(re.search('[a-z]\)\s', x)) else '' for x in df['inciso']]
df['numeral'] =[re.search(r'^\s*[0-9]+\.',x).group(0) if bool(re.search(r'^\s*[0-9]+\.', x)) else '' for x in df['numeral']]
df['texto'] =[x.strip('\s') for x in df['texto']]
```

```
def asigna(x):
    global tmp
    tmp = x
    return x
#relleno de celdas agrupamiento de párrafos según categorías del texto y dependencias
tmp=''
df['titulo']=[tmp if (not str(x)) else asigna(x) for x in df['titulo']]
tmp=''
df['capitulo']=[tmp if (not str(x)) else asigna(x) for x in df['capitulo']]
tmp=''
df['seccion']=[tmp if (not str(x)) else asigna(x) for x in df['seccion']]
tmp=''
df['articulo']=[tmp if (not str(x)) else asigna(x) for x in df['articulo']]

df=df[~((df['texto']==df['titulo'])|(df['texto']==df['capitulo']))]    #Filtro de datos
```

In [8]:

```
#Relleno de valores según dependencia de categorías
def rellenainciso(x,y):
    for i in range(0,len(y)):
        if y.iloc[i] is not '':
            x.iloc[i]=x.iloc[i-1]
    return x
```

In [9]:

```
x=rellenainciso(df['inciso'].copy(),df['numeral'])
df['inciso']=x.copy()

y=rellenainciso(df['fraccion'],df['inciso'])
df['fraccion']=y.copy()
#Limpieza de cadenas para DML
df.replace('^\s+', '', regex=True, inplace=True)
df.replace('\s+$', '', regex=True, inplace=True)
df.replace('^#+', '', regex=True, inplace=True)
df['abreviatura']='Art.'+df.loc[:,'articulo']+' '+df.loc[:,'fraccion']
```

In [10]:

```
#df
```

In [11]:

```
df=df[~df['texto'].str.contains("Transitorios")]   #Filtro
```

In [12]:

```
df.replace('^\s+', '', regex=True, inplace=True)    #Limpieza, normalización de texto
```

```
df.replace('\s+$', '', regex=True, inplace=True)
```

In [13]:

```
df=df[['titulo','capitulo','seccion','articulo','apartado','fraccion','inciso','numeral','abreviatura','texto']]
```

In [14]:

```
df.to_csv('inai_categorizado.csv',header=True) # Guarda tabla en archivo csv
```

In [15]:

```
import nltk                    #Biblioteca para procesamiento de NPL
#nltk.download('punkt')         #Descarga paquetes necesarios
#nltk.download('stopwords')     #Idem
```

In [16]:

```
df['texto']                #Control de valores
df["unigrams"] = df["texto"].apply(nltk.word_tokenize)   #Separación del texto por palabras
```

In [17]:

```
#df.iloc[:,-1]
```

In [18]:

```
##Por renglon
```

In [19]:

```
Muestra707577=df[df['articulo'].isin(['70','75','77'])]    #Filtro - artículos de interés Muestra
Muestra707577=Muestra707577.iloc[:,-3:]                    #Filtro
#Muestra707577
```

In [20]:

```
from collections import Counter                  #Bibliotecas para conteo de palabras
from nltk.tokenize import word_tokenize
from nltk.util import ngrams
from nltk import FreqDist
from nltk.corpus import stopwords
from math import trunc
```

In [21]:

```
def remove_stopwords(sentence, language):   #Eliminación de caracteres no significativos
    from nltk.corpus import stopwords
    from nltk.tokenize import wordpunct_tokenize
    stop_words = set(stopwords.words(language))
    stop_words.update(['.', ',', "", '"', '?', '!', ':', ';', '(', ')', '[', ']', '{', '}']) # remove it if you need punctuation

    return [i for i in wordpunct_tokenize(sentence) if i.lower() not in stop_words]
```

In [22]:

```
#Buffer -->Procesamiento por renglon
i            = 0
desplazamiento = 0
```

```
#Tramas vacias para acumulación de datos
EspacioMuestral = pd.DataFrame([('A','B',0)],columns=['abreviatura','token','distanciaR'],index=[0])
#Ngrams  = pd.DataFrame([('A','B',0)],columns=['abreviatura','token','distanciaR'],index=[0])
Ngrams=pd.DataFrame()
Art_aprov  = pd.DataFrame([('A','B',0)],columns=['abreviatura','token','Freq_abrev'],index=[0])

#Recorrido por renglón y extracción de información
for i in range(0,len(Muestra707577)):
    registro_actual              = pd.DataFrame(Muestra707577.iloc[i,-1])     # Formato
    registro_actual['abreviatura']= Muestra707577.iloc[i,-3]                  # ...
    registro_actual.columns      = ['token','abreviatura']
    registro_actual              = registro_actual[['abreviatura','token']]
    registro_actual['distanciaR'] = registro_actual.index
    registro_actual['distanciaT'] = registro_actual.index+desplazamiento
    desplazamiento=len(registro_actual['distanciaR'])-1               #distancia de palabra

    unigramas_actual = remove_stopwords(Muestra707577.iloc[i,1],'spanish')  #Limpieza

    bigramas_actual  = list(nltk.bigrams(unigramas_actual))         #Enlistado
    trigramas_actual = list(nltk.trigrams(unigramas_actual))        #Enlistado

    unigramas_actual
    bigramas_actual  = [' '.join(grams) for grams in bigramas_actual]
    trigramas_actual = [' '.join(grams) for grams in trigramas_actual]

    b1=pd.DataFrame(unigramas_actual, columns=['token'])
    b2=pd.DataFrame(bigramas_actual, columns=['token'])
    b3=pd.DataFrame(trigramas_actual, columns=['token'])

    bactual=b1.append(b2)
    bactual=bactual.append(b3)

    bactual['abreviatura']=Muestra707577.iloc[i,-3]
    bactual=pd.merge(bactual,registro_actual,on='token',how='left')

    Ngrams =Ngrams.append(bactual,ignore_index=True)                #Concatenación de gramáticas 1,2,3
    Ngrams
    #bactual
    if Muestra707577.iloc[i,-3] != Muestra707577.iloc[i-1,-3]:
       Art_aprov=Art_aprov.append(pd.DataFrame([(Muestra707577.iloc[i,-3],
                                trunc(((len(unigramas_actual)*100)/len(Muestra707577.iloc[i,-2])))
                               )],columns=['abreviatura','Freq_abrev'])
                  ,ignore_index=True)   # Métrica de palabras tomadas contra palabras desechas
```

In [23]:

```
Art_aprov= Art_aprov.iloc[1:,:-1]  # Filtro
Art_aprov=Art_aprov[['abreviatura','Freq_abrev']]  # Formato
```

In [24]:

```
Ngrams=Ngrams.iloc[:,[0,1,3,4]]     #Filtro
Ngrams.columns=['token','abreviatura','distanciaR','distanciaT'] #Renombre de columnas
```

In [25]:

```
Ngrams=Ngrams.loc[:,['abreviatura','token','distanciaR','distanciaT']] # Filtro
```

In [26]:

```
#Cuenta total de unigrmas, bigramas, trigramas de toda la muestra
from sklearn.feature_extraction.text import CountVectorizer
word_vectorizer = CountVectorizer(ngram_range=(1,3), analyzer='word',lowercase=False)
sparse_matrix = word_vectorizer.fit_transform( Muestra707577.iloc[:,-2])
frequencies = sum(sparse_matrix).toarray()[0]

TotalCount=pd.DataFrame(frequencies, index=word_vectorizer.get_feature_names(), columns=['frequency'])
TotalCount.reset_index(inplace=True)
TotalCount.columns=['token','freq']
```

In [27]:

```
#TotalCount
```

In [28]:

```
Ngrams=pd.merge(Ngrams,TotalCount,on='token')    #reunión de Ngramaticas Frecuencia,distancia ...
```

In [29]:

```
Ngrams=Ngrams[['abreviatura','token','distanciaR','distanciaT']] # Formato de columnas
Ngrams.columns=['source','target','distanciaR','distanciaT']    # Formato para generar grafo de relaciones
Ngrams.to_csv('Aristas.csv',header=True,insdex=False)          # Guarda Datos procesados
```

In [30]:

```
nodop=Ngrams['target'].value_counts().reset_index()           #Lista de nodos y sus atributos Artículo abreviatura, palabra frecuencias
nodop.columns=['label','freq']
nodop['tipo']='B'
Art_aprov.columns=['label','freq']
Art_aprov['tipo']='A'
nodop=nodop.append(Art_aprov)
nodop['id']=nodop['label']
nodop=nodop[['id','label','tipo','freq']]
nodop.to_csv('nodos.csv',header=True,index=False)             #Guarda lista de nodos
```

In [31]:

```
#Fuente: http://stackoverflow.com/questions/19130512/stopword-removal-with-nltk
from nltk.corpus import stopwords
```

```
from nltk.tokenize import wordpunct_tokenize


stop_words = set(stopwords.words('spanish'))
stop_words.update(['.', ',', "", "'", '?', '!', ':', ';', '(', ')', '[', ']', '{', '}']) # remove it if you need punctuation
list_of_words=[]
#Listado de palabras
for doc in Muestra707577['texto']:
    list_of_words+= [i for i in wordpunct_tokenize(doc) if i.lower() not in stop_words]
```

In [32]:

```
texto = nltk.Text(list_of_words)
fdist = nltk.FreqDist(list_of_words) # Frecuencia en la muestra de palabras


#https://github.com/alukach/nltk-experiments/blob/master/natural_language_processing_toolkits_cheatsheet.md
tokensmasfrecuentes=sorted([w for w in set(texto) if len(w) >4 and fdist[w] > 5]) #Filtrado de palabras frecuentes
fdist2 = nltk.FreqDist(tokensmasfrecuentes)    #fdist2<-más frecuentes
```

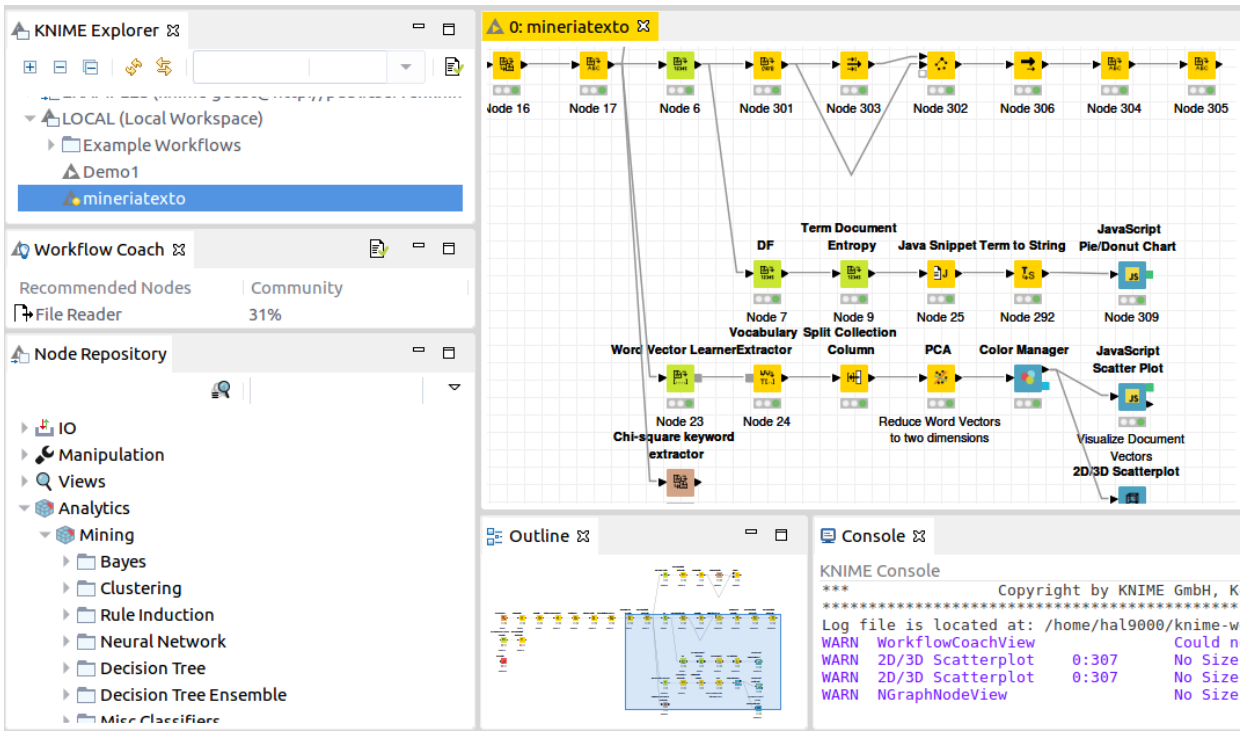
In [33]:

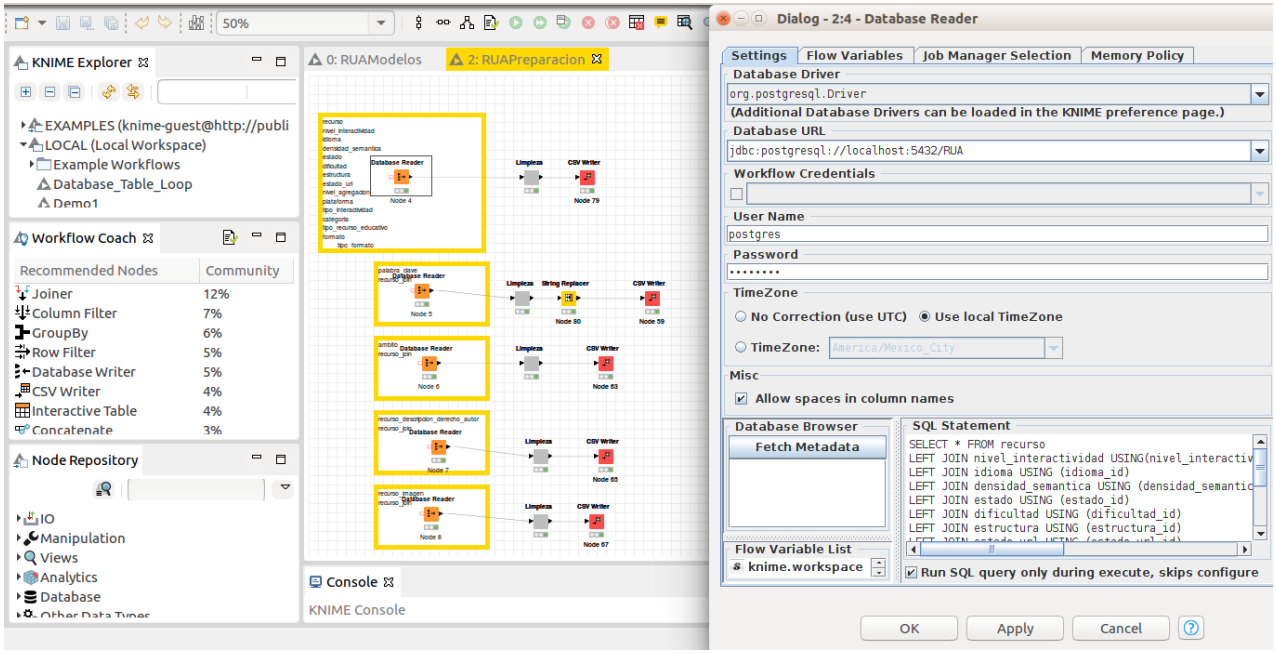
```
texto.dispersion_plot(tokensmasfrecuentes)
# Checar ventanas emergentes
# Gráfica la dispersión de palabras más frecuentes
```

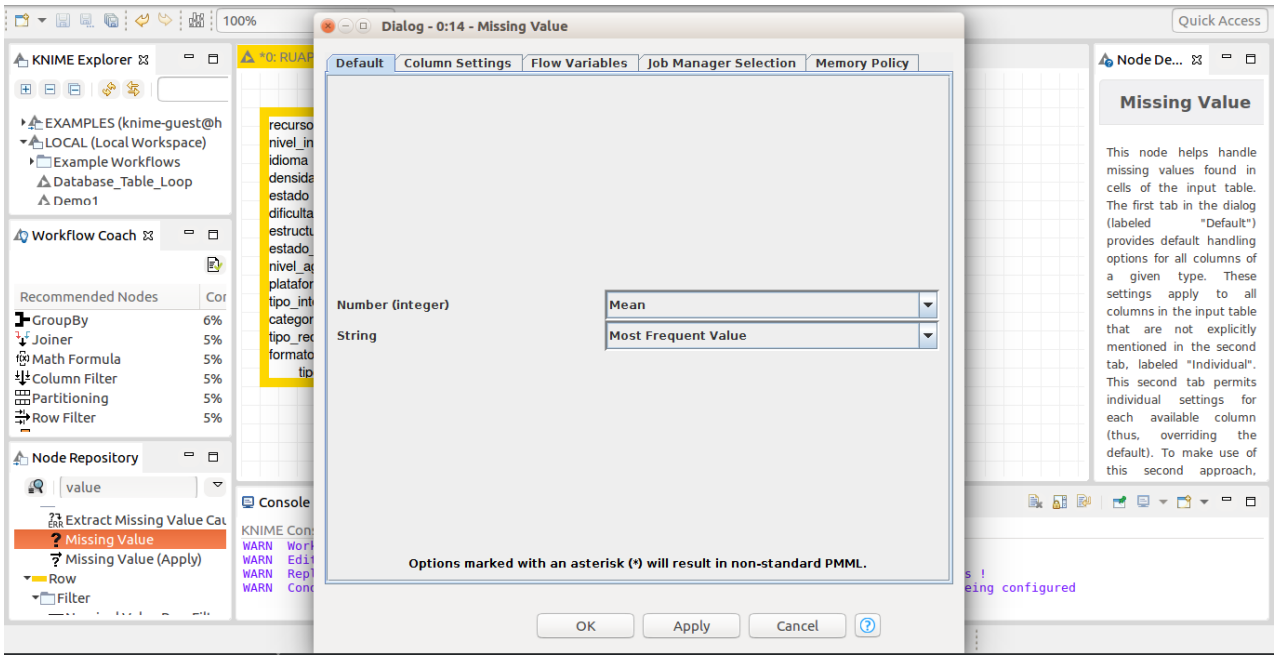
In [34]:

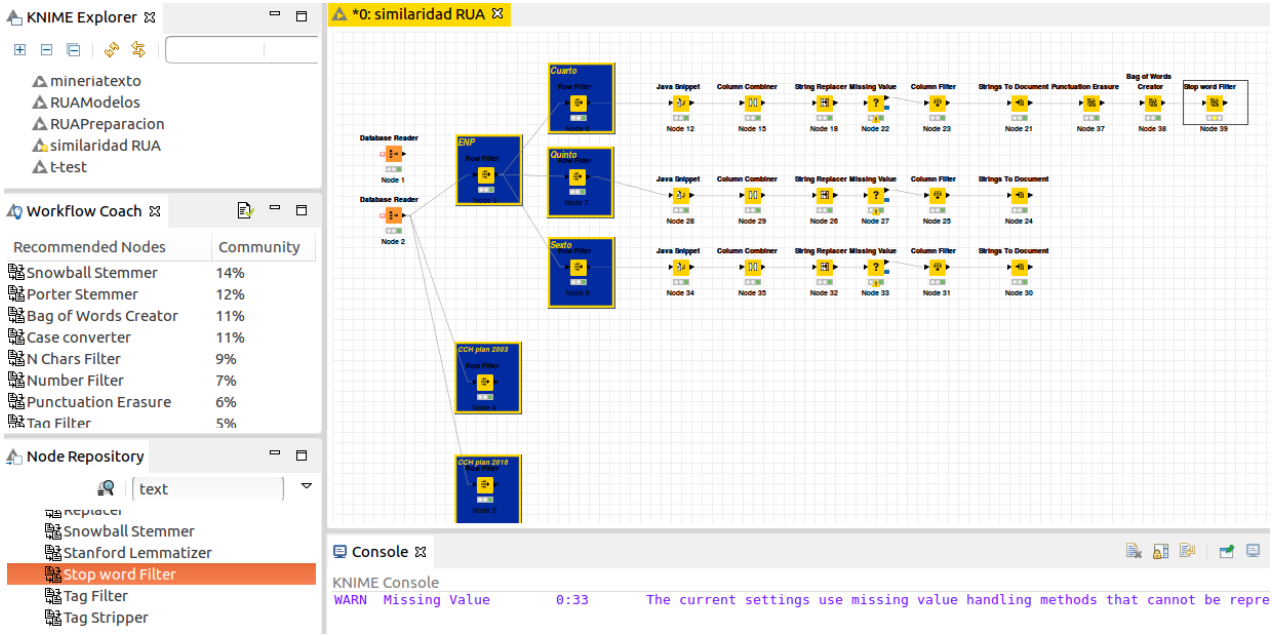
```
fdist2.plot(25, cumulative=True)          # Frecuencia acumulada
```

F Ambiente de Knime






Dialog - 0:14 - Missing Value
Default
Column Settings
Flow Variables
Job Manager Selection
Memory Policy
Number (integer)
Mean
String
Most Frequent Value
Options marked with an asterisk (*) will result in non-standard PMML.
OK
Apply
Cancel
Missing Value
This node helps handle missing values found in cells of the input table. The first tab in the dialog (labeled "Default") provides default handling options for all columns of a given type. These settings apply to all columns in the input table that are not explicitly mentioned in the second tab, labeled "Individual". This second tab permits individual settings for each available column (thus, overriding the default). To make use of this second approach,
KNIME Explorer
Workflow Coach
Node Repository
Console


KNIME Explorer
mineriatexto
RUAModelos
RUAPreparacion
similaridad RUA
t-test
Workflow Coach
Recommended Nodes
Community
Snowball Stemmer 14%
Porter Stemmer 12%
Bag of Words Creator 11%
Case converter 11%
N Chars Filter 9%
Number Filter 7%
Punctuation Erasure 6%
Tag Filter 5%
Node Repository
text
Snowball Stemmer
Stanford Lemmatizer
Stop word Filter
Tag Filter
Tag Stripper
*0: similaridad RUA
Console
KNIME Console
WARN Missing Value 0:33 The current settings use missing value handling methods that cannot be repre

G Programa dibujo histogramas sobre base de datos en Python

```
import pandas as pd
import sqlalchemy as motor
import matplotlib.pyplot as plt
import matplotlib
import numpy as np
##matplotlib.style.use('ggplot')
```

In [ ]:

```
Controlador = motor.create_engine("postgresql://postgres@localhost:5432/RUA");
```

In [ ]:

```
def histograma(df,nombre):
    %matplotlib inline
    plt.figure()
    dominio =df.index
    pos = np.arange(0,len(df.index))
    frecuencia = df['Cuenta'].values


# change the bar color to be less bright blue
    bars = plt.bar(pos, frecuencia, align='center', linewidth=0, color='lightslategrey')
# make one bar, the python bar, a contrasting color
    #bars[0].set_color('#1F77B4')

# soften all labels by turning grey
    plt.xticks(pos, dominio, alpha=0.8)
# remove the Y label since bars are directly labeled

    plt.title(nombre, alpha=0.8)

# remove all the ticks (both axes), and tick labels on the Y axis
    plt.tick_params(top='off', bottom='off', left='off', right='off', labelleft='off', labelbottom='on')

# remove the frame of the chart
   # for spine in plt.gca().spines.values():
     #   spine.set_visible(False)
    #
# direct label each bar with Y axis values
    for bar in bars:
        plt.gca().text(bar.get_x() + bar.get_width()/2, bar.get_height() - 5, str(int(bar.get_height())) ,
                ha='center', color='w', fontsize=11)
    x = plt.gca().xaxis
```

```
    plt.subplots_adjust(bottom=0.25)

# rotate the tick labels for the x axis
    for item in x.get_ticklabels():
        item.set_rotation(90)
```

In [ ]:

```
def grafica(nom,df):
    for cn in df.columns:

        df[cn]= df[cn].astype(str)
        df[cn].replace(r'\s*','',regex=True,inplace=True)
        df[cn].replace(r'None','NA',regex=True,inplace=True)
        df[cn]=df[cn].sort_values().copy()
        Columna=df[cn].copy()
        Columna=Columna.reset_index().groupby(cn).count()
        Columna.columns=['Cuenta']
        histograma(Columna,cn)
        plt.show()
        #plt.savefig("Tabla="+nom+"[columna="+cn+"]");
```

In [ ]:

```
def generaestadisticas(nom,df):
    print("---------------------------------------")
    print("[tabla:"+nom+"]")
    #print(df.describe().T)
    grafica(nom,df)
```

In [ ]:

```
x="recurso_join"
tabla = pd.read_sql("SELECT * FROM "+x+"",con=Controlador);
generaestadisticas(x,tabla)

    # ////////////////////////variante para todas las tablas
for x in Tablas.iloc[:,0]:
    tabla = pd.read_sql("SELECT * FROM "+x+"",con=Controlador);
    if(len(tabla.iloc[:,0])>1):
        generaestadisticas(x,tabla)
```

H Programa para gráficar distribución normal en Python

```
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
import math
%matplotlib inline
```

In [46]:

```
plt.figure(figsize=(12,10))

plt.fill_between(x=np.arange(-4,-2,0.01),
                 y1= stats.norm.pdf(np.arange(-4,-2,0.01)) ,
                 facecolor='green',
                 alpha=0.25,linewidth=3)

plt.fill_between(x=np.arange(-2,2,0.01),
                 y1= stats.norm.pdf(np.arange(-2,2,0.01)) ,
                 facecolor='white',
                 alpha=0.35,linewidth=3)

plt.fill_between(x=np.arange(2,4,0.01),
                 y1= stats.norm.pdf(np.arange(2,4,0.01)) ,
                 facecolor='green',
                 alpha=0.25,linewidth=3)

plt.fill_between(x=np.arange(2.5,4,0.01),
                 y1= stats.norm.pdf(np.arange(2.5,4,0.01)) ,
                 facecolor='blue',
                 alpha=0.35,linewidth=3)
plt.tick_params(top='off', bottom='off', left='off', right='off', labelleft='off', labelbottom='off')
plt.text(x=-1.3, y=0.15, s= "Hipótesis nula",fontsize=30)
plt.text(x=-0.5, y=0.13, s= "CI: 95%",fontsize=30)
plt.text(x=2, y=0.06, s= "Región de rechazo",fontsize=18)
plt.text(x=-4, y=0.06, s= "Región de rechazo",fontsize=18)
plt.text(x=2, y=0.02,s=r'$t_\alpha$',fontsize=30)
plt.text(x=2.5, y=0.025,s=r'$t_p$',fontsize=30)
for spine in plt.gca().spines.values():
    spine.set_visible(False)
help(plt.text)
```