**Terminología**

|  |  |
| --- | --- |
| Título del proyecto de tesis: | Caracterización de recursos computacionales para la fase de preprocesamiento de minería de textos |
| UPIITA-IPN: | Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas del Instituto Politécnico Nacional. |
| INECOL: | Instituto de Ecología. |
| Text Mining: | Es el término en idioma inglés, que generalmente se refiere a la extracción automática de información interesante y no trivial de un texto no estructurado. Generalmente, su propósito no radica en comprender todo o parte de lo que dice un hablante/escritor en particular, sino más bien en extraer patrones de una gran cantidad de documentos (Kao & Poteet, 2007). |
| PNL: | La minería de textos está relacionada con el Procesamiento del Lenguaje Natural (PNL), que incluye técnicas de inspiración lingüística, es decir, un texto se analiza típicamente desde un punto de vista léxico y sintáctico utilizando una gramática formal, la información resultante se interpreta semánticamente y se utiliza para extraer información sobre lo dicho (Kao & Poteet, 2007). |
| Preprocesamiento: | Se refiere a la realización de operaciones o transformaciones sobre el texto en algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis. Esta fase se lleva a cabo sobre un conjunto de documentos objetos de estudio determinando el tipo de representación o patrones contenidos en los textos (Taeho, 2019). |
| Transformación pdf a texto: | Los textos se transforman en representación estructurada o semiestructurada que facilite su posterior análisis. Se define el conjunto corpus de documentos. Estos documentos deben ser representativos y seleccionarse aleatoriamente o mediante algún método de muestreo probabilístico. Se debe evitar, en esta etapa, la duplicación de documentos dentro del corpus (Cortez, 2018). |
| Eliminación de ruido: | La eliminación de palabras vacías se refiere al proceso de eliminar palabras vacías de la lista de tokens o palabras derivadas (Kowalski & Maybury, 2000). Las palabras de parada son las palabras gramaticales que son irrelevantes para el contenido del texto, por lo que deben eliminarse para una mayor eficiencia. La palabra clave de la lista se carga desde un archivo y, si están registrados en la lista, se eliminan. |
| Tokenización: | Tokenización: La tokenización se define como el proceso de segmentar un texto o textos en tokens por el espacio en blanco o los signos de puntuación. Es capaz de aplicar la tokenización al códigos fuente en C, C ++ y Java (Aho et al., 2007) así como los textos que están escritos en un lenguaje natural. Sin embargo, el alcance está restringido al texto en este estudio, a pesar de la posibilidad. |
| Etiquetado de textos: | Etiquetado de textos: Se conoce como derivación es una técnica de preprocesamiento que se ocupa de reducir una palabra a su forma básica, llamada raíz. En este contexto, no es importante que la palabra elegida sea una sola palabra, sino que las variantes morfológicas del mismo término, que en la mayoría de los casos tienen interpretaciones semánticas similares, se asignen todas a la misma raíz (Kowalski & Maybury, 2000). |
| Python: | Python está desarrollado bajo una licencia de código abierto aprobada por OSI (modelo de interconexión de sistemas abiertos) lo que lo hace de libre uso y distribución, incluso para uso comercial. La licencia es administrada por Python Software Foundation (Python, 2021). |
| NLTK: | NLTK (Natural Language Toolkit, de sus siglas en inglés) es una plataforma líder para crear programas de Python para trabajar con datos de lenguaje humano. Proporciona interfaces fáciles de usar para más de 50 corpus y recursos léxicos como WordNet, junto con un conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, lematización, etiquetado, análisis y razonamiento semántico, contenedores para bibliotecas NLP de potencia industrial, y un foro de discusión activo (NLTK, 2021).  RapidMiner lleva la inteligencia artificial a la empresa a través de una plataforma de ciencia de datos abierta y extensible. Creado para equipos de análisis, Unifica todo el ciclo de vida de la ciencia de datos, desde la preparación de datos hasta el aprendizaje automático y  la implementación de modelos predictivos. Más de 700,000 profesionales de análisis utilizan productos RapidMiner para generar ingresos, reducir costos y evitar riesgos (RapidMiner, 2021). |
| Clustering: | El proceso de seccionar un grupo de objetos o datos en una colección de subclases relevantes y comprensibles se denomina clustering. El clustering se utiliza principalmente para hacer un conjunto de documentos y archivos similares (Tandel et al., 2019). |
| CRISP-DM: | CRISP-DM (Cross Industry Standard Process for Data Mining), el cual es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema (Chapman et al., 2000). |
| TDSP: | Es una metodología de ciencia de datos ágil e iterativa para proporcionar soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente. TDSP ayuda a mejorar la colaboración y el aprendizaje en equipo al sugerir cómo los roles de equipo funcionan mejor juntos. Incluye procedimientos recomendados y estructuras de Microsoft y otros líderes del sector para ayudar a implementar correctamente iniciativas de ciencia de datos. El objetivo es ayudar a las empresas a que se den cuenta de las ventajas de su programa de análisis (Microsoft, 2021). |
| Guía experimental de McGeoch: | Es una guía experimental la cual muestra dos etapas. La primera es la planificación del experimento, en la que partimos de la formulación de una pregunta, construimos o utilizamos herramientas (guiones de programación, paquetes de software) y finalizamos con el diseño mismo del experimento. Luego, en la fase de ejecución, el experimento se lleva a cabo y los datos recogidos se analizan y se publican a través de los reportes respectivos. |
| Unidad experimental: | Son los objetos, individuos, intervalos de espacio o tiempo sobre los que se experimenta. |
| Variable de interés o respuesta: | Es la variable que se desea estudiar y controlar su variabilidad. |
| Factor: | Son las variables independientes que pueden inﬂuir en la variabilidad de la variable de interés. |
| Factor tratamiento: | Es un factor del que interesa conocer su inﬂuencia en la respuesta. |
| Factor bloque: | Es un factor en el que no se está interesado en conocer su inﬂuencia en la respuesta pero se supone que ésta existe y se quiere controlar para disminuir la variabilidad residual. |
| Niveles: | Cada uno de los resultados de un factor. Según sean elegidos por el experimentador o elegidos al azar de una amplia población se denominan factores de efectos ﬁjos o factores de efectos aleatorios. |
| Tratamiento: | Es una combinación especíﬁca de los niveles de los factores en estudio. Son, por tanto, las condiciones experimentales que se desean comparar en el experimento. En un diseño con un único factor son los distintos niveles del factor y en un diseño con varios factores son las distintas combinaciones de niveles de los factores. |
| Observación experimental: | Es cada medición de la variable respuesta. |
| Tamaño del Experimento: | Es el número total de observaciones recogidas en el diseño. |
| Interacción de factores: | Interacción de factores: Existe interacción entre dos factores FI y FJ si el efecto de algún nivel de FI cambia al cambiar de nivel en FJ. Esta de ﬁnición puede hacerse de forma simétrica y se puede generalizar a interacciones de orden tres o superior. |
| Diseño equilibrado o balanceado: | Es el diseño en el que todos los tratamientos son asignados a un número igual de unidades experimentales. |
| Variable independiente: | Es una variable que representa una cantidad que se modifica en un experimento. |
| Variable dependiente: | Representa una cantidad cuyo valor depende de cómo se modifica la variable independiente (Khan Academy, 2021). |