1. TITULO: Proyecto Integrador – Parte 1 – Procesamiento de Texto.

AUTORES: Diana Catalina Velásquez Gaviria 201919006228

Jairo Loaiza Gutiérrez 201919005228

Juan Sevastián Moreno Zapata 201917508228

Julián Castelblanco Benítez 201910040228

Mariana Arboleda Florez 201910015114

AFILIACIONES: ¿??

1. RESUMEN DE LOGROS DEL AVANCE

Se realizaron las siguientes actividades:

* Se cargaron los archivos txt en el servidor
* Se leyeron la totalidad de los archivos (980 archivos)
* Se cargó el texto de todos los archivos en un único string
* Se seleccionaron únicamente los caracteres que estén compuestos por letras mayúsculas o minúsculas
* Se tokenizó el string, con el fin de crear un vector que contuviera todas las palabras del texto
* Se cargó el archivo de stopwords y se eliminan del texto las palabras que pertenecen a este, con el objetivo de eliminar los artículos y conectores, por ej. 'i', 'me', 'my'
* Se cargó el stemmer y luego se aplica al string, con el objetivo de unificar todas aquellas palabras que contiene una misma raíz, por ej. girls🡺girl, better 🡺best
* Se ejecutó la lematización, con el fin de cambiar las palabras que significan lo mismo en tiempo presente, por ej. is, are, were, was 🡺 be
* Se creó la lista de palabras únicas (voc)
* Se realizó el análisis de frecuencias de las palabras con el objetivo de identificar las palabras que más se repiten en la bolsa de palabras

1. Introducción y Marco teórico

* Tokenización
* Stopwords
* Stemmer (Lancaster y Porter\*¨seleccionado)
* Lematización (Hard – sof)

1. Problema

Se tiene un conjunto muy grande (big data ) de documentos tipo texto (txt, html o pdf) y los metadatos sobre dichos documentos, realizar una sistema/aplicación para la ingesta, almacenamiento , indexación , búsqueda, recuperación , navegación y visualización de documentos. Estos documentos pertenecen al mismo dominio de interés, por ejemplo: ingeniería, medicina, o astronomía, entre muchos otros campos de conocimiento o aplicación.

Estos sistemas presentan desafíos en cada una de sus etapas así como la visión conjunta del sistema, integración entre la computación (programación, desarrollo de software, algoritmos, infraestructura TI, escalabilidad) y las matemáticas/estadística que permite desarrollar los modelos de analítica requeridos para poder realizar este tipo de sistemas, analítica principalmente relacionada con el campo del Procesamiento Natural del Lenguaje, particularmente en dos campos: Text Retrieval & Text Mining. (Text Retrieval perteneciendo a otro campo más amplio conocido como Information Retrieval, y Text Analytics perteneciendo a otro campo más amplio conocido como Data Analytics).

Búsqueda y Recuperación de Documentos:

Como insumo se tiene un conjunto grande de documentos científicos de acceso público, a los cuales se tiene acceso tanto en texto completo, así como sus metadatos. Normalmente el texto completo se encuentra en PDF y los metadatos en un formato tipo Dublin Core, RDF o similar, típicamente 2 3 codificado en XML, XHTML, JSON o similar.

El equipo de trabajo deberá analizar las alternativas de procesamiento conjunto entre documentos y metadatos, para mejorar la ‘precisión’ o pertinencia en la respuesta (conjunto de documentos seleccionados a partir de una consulta -query-) para el usuario.

Deberán explorar alternativas de indexación del texto, representación de documentos/metadatos, modelos de búsqueda, mecanismos de ranking o priorización de los resultados, entre otros elementos.

1. Datasets

Se tendrá tres (3) conjuntos de datos relacionados:

1. papers-pdf.zip (980 papers)

2. papers-txt.zip (980 papers)

3. papers-dc.xml

URL: https://goo.gl/k3SQtF

● ‘papers-pdf.zip’ contiene un conjunto de artículos científicos en formato PDF

● ‘papers-txt.zip’ contiene el mismo conjunto de artículos científicos de ‘papers-pdf.zip’ pero

convertidos a txt para facilitar su procesamiento.

● ‘papers-dc.xml’ contiene los metadatos de los artículos científicos en formato XML en estándar Dublin

Core Metadata Initiative (DCMI)

Estos artículos están en idioma Inglés.

1. Arquitectura

Para el desarrollo de este proyecto se tiene la siguiente plataforma:

* Un servidor Ubuntu Linux con:
  + 72 nucleos,
  + 512 GB de RAM,
  + 252 GB de Disco SSD y
  + 2 Teslas 480
* Plataforma Jupyter para ejecutar código Python y MarkDown

1. Modelos

El primer seguimiento del proyecto, consta de la PREPARACIÓN DE DATOS, en la cual se han aplicado los siguientes modelos de procesamiento de datos:

* Stemming
* Lemmatization
* Representación
* BoW, etc para Text Analytics)
* Estarán concentrados en los modelos y mejores prácticas para la PREPARACIÓN DE DATOS EN TEXTO 8

1. Evaluación y resultados