Universidad de Carabobo

Facultad Experimental de Ciencias y Tecnología

Departamento de Computación

Recuperación de Información:

Informe Clusterización

de documentos por categorías

Alumna:

Mayerling Nazareth Aguilera Suarez

Cédula de Identidad: 24.290.965

Octubre 2019

Informe de búsqueda

Objetivo: Se desea realizar un programa para la búsqueda de documentos en pdf

Para la realización del programa se uso el Lenguaje de Programación Python con las siguientes librerias:

1. os: Funciones del sistema.
2. PyPDF2: Extrae el texto de pdf
4. nltk : Libreria de procesamiento de lenguaje natural.
5. sklearn: Libreria para clusterizacion.
6.numpy: Computo cientifico.
7.math: Libreria matematica.

En el desarrollo de la implementación de la búsqueda fue necesario extraer el texto de los pdf usando la función PdfFileReader de la librería PyPDF2 , con lo cual se leyeron cada una de las paginas transformándolas a texto plano para ser almacenadas en un archivo .txt

Ademas se pre-procesa el texto de los pdf usando la librería nltk. Se divide el texto en palabras denominadas tokens con la funcion word_tokenize, luego de este proceso se eliminan los signos de puntuación y artículos para reducirlos a su palabra raíz obteniendo como resultado una cadena del texto limpio.

A continuación se calcula la frecuencia de ocurrencia de términos en la colección de documentos inicializando un vector de tf-idf con la funcion TfidfVectorizer(), para luego construir el vocabulario de todo el documento tokenizando.

Se hace uso de un conjunto de etiquetas para asociar los resultados del TFIDF con el fin de reconocer con facilidad a que documento perteneces cada termino del conjunto.

Por ultimo se utiliza el clasificador de clusterizacion k-means que proviene de la librería SKLEARN, se inicializa un modelo que permite clusterizar los documentos en 3 clusters, una vez

realizado este proceso se entrena el modelo con el conjunto de datos que consiste en un vector que contiene los tf-idf con sus respectivas etiquetas.

Con los pasos mencionados anteriormente se posee un modelo que permite clasificar un conjunto de documentos en clusters para la realización de predicciones basados palabras claves otorgadas por un usuario.

Para las pruebas se poseen 3 categorías las cuales son: HCI, AE y TEED asociadas a una n-cantidad de documentos.

Pasos para ejecutar el código:

Si se tiene dos versiones de python usar el comando pip3 y python3:

1. pip install -r requirements.txt
2. python main.py