Ejercicio

Rapid Keyword Extraction

1 Propósito

La finalidad de este ejercicio es utilizar el algoritmo RAKE (RApid Keyword Extraction) para obtener información acerca de documentos que forman nuesro corpus.

Las actividades a realizar son:

1. Generación del corpus
2. Selección de documentos a extraer
3. Extracción de palabras clave usando RAKE
4. Extracción de palabras clave usando rake_nltk
5. Comparación de resultados

Cada uno de los pasos va a generar un entregable. Al final todos los entregables generados se deberán subir a Canvas junto con los scripts que hayas usado para cada uno de los pasos. La recomendación es que los scripts los generes usando notebooks.

2 Generación del corpus

Para este ejercicio vamos a utilizar los scripts que usamos para obtener documentos de “The Guardian”, modificándolos para hacer los siguiente:

- La fecha de inicio y fin de la búsqueda será el lunes 2 de octubre de 2023
- Solamente vamos a guardar los textos de los artículos que hayan sido publicados en las secciones de “fooball” y “us-news” (En caso de que no haya artículos en esas secciones, podemos ampliar los días de búsqueda, por ejemplo, la semana del 25 de septiembre al 2 de octubre)
- En el archivo de trabajo, se debe crear una subcarpeta “textos” y dentro de la misma, una subcarpeta para cada sección. De la siguiente forma:
    - CarpetaDeTrabajo
        - textos
            - us-politics
            - football

3 Selección de documentos a extraer

Vamos a seleccionar de forma aleatoria 3 (tres) documentos de cada una de las secciones que acabamos de bajar.

4 Extracción de palabras clave

4.1 Comparativa de las librerías

Vamos a extraer las palabras clave usando cada una de las librerías (RAKE y rake_nltk), para ello puedes usar como base los notebooks vistos en clase.

A manera de comparativa entre ambas herramientas, modifica el notebook para que:

- Se extraigan las palabras clave con RAKE y luego con rake_nltk.
- Se generen los DataFrame con las (5 <= n <= 20) frases claves extraídas por las herramients
- Presenta los dos DataFrame uno al lado de otro para poder comparar más fácilmente
- Repetir el proceso para todos los documentos

5 Entregables

Al haber concluido todos los pasos en este ejercicio, debes tener:

- 2 notebooks correspondientes a:
  - obtención del corpus
  - extracción de frases clave y comparativa de las herramientas
- Un directorio con 2 subdirectorios cada uno con los archivos de texto resultado de los pasos vistos anteriormente.

Genera dentro del mismo directorio de trabajo un nuevo directorio “notebooks” y guárda allí los notebooks que desarrollaste. Por favor, agrega a cada notebook un título que describa de qué se trata y un espacio donde te identifiques con nombre y número de expediente.

Agrega todo en un archivo “Ejercicio02_tuNombre.zip” y súbelo en el espacio designado en canvas para este ejercicio.