3.2 Actividad 2 | Más expresiones Regulares y matrices de conteo y tf-idf - Semanas 3 y 4

Empezar tarea

- Fecha de entrega Lunes a las 23:59
- Puntos 12
- Entregando una URL de página web

Objetivos

1.5 Ejemplificar los conceptos de tokenización y aplicarlos para la generación de las matrices de conteo (document-term-matrix) y tf-idf.

1.6 Explicar los conceptos que involucran el pre-procesamiento de un texto: limpieza de un texto, identificación de las palabras de enlace o stopwords, aplicar criterios de normalización y obtención de matrices de conteo mediante matrices dispersas.

Instrucciones

Desarrolla esta actividad siguiendo las siguientes indicaciones y que te permitirán familiarizarte con el proceso de tokenización y preprocesamiento de documentos de texto en un problema de análisis de sentimiento en NLP. Los datos muestran comentarios en inglés sobre si les gustó o no una película, servicios de comida y productos comprados en línea. Este es un tipo de problema muy común en el área de NLP, que de hecho se le conoce como "análisis de sentimiento" y que estaremos abordando continuamente a lo largo del curso.

Instrucciones:

1. Descarga el archivo de JupyterNotebook **MNA_NLP_semanas_03_04_Actividad.ipynb (https://experiencia21.tec.mx/courses/575069/files/226236689?wrap=1)** **(https://experiencia21.tec.mx/courses/575069/files/226236689/download?download_frd=1)** y ábrelo por ejemplo en Google Colab.
2. En esta actividad deberás utilizar los datos de los archivos llamados **amazon_cells_labelled.txt**, **imdb_labelled.txt** y **yelp_labelled.txt**, que puedes encontrar en el archivo llamado **sentiment labelled sentences.zip** en la liga de la página de la UCI:

https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences# (https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences#)

3. Sigue las indicaciones dentro del archivo de JupyterNotebook para realizar la actividad de estas dos próximas semanas.
4. Una vez terminada la actividad, deberás subir tu archivo de JupyterNotebook a GitHub con el nombre indicado en la sección "Especificaciones de entrega", de más abajo.
5. Dentro de Canvas, en esta página de la actividad, deberás incluir solamente la liga a tu archivo de la actividad en GitHub.

Puedes consultar el siguiente documento para saber cómo trabajar con archivos "ipynb" de Jupyter-Notebook en Google-Colaboratory: **Accesando los archivos de Jupyter Notebook** (https://experiencia21.tec.mx/courses/575069/files/226236796?wrap=1) (https://experiencia21.tec.mx/courses/575069/files/226236796/download?download_frd=1) .

Puedes consultar el siguiente documento para las instrucciones de cómo subir y compartir en GitHub tu archivo de la actividad de esta semana: **Archivos en GitHub** (https://experiencia21.tec.mx/courses/575069/files/226236791?wrap=1) (https://experiencia21.tec.mx/courses/575069/files/226236791/download?download_frd=1) .

Especificaciones de entrega

- **Modalidad:** Individual.
- **Medio de realización/entrega**: Subir el archivo en GitHub, y la liga al archivo a través del botón "Entregar tarea" de esta actividad.
- **Formato:** Archivo de Jupyter Notebook (ipynb).
- **Nombre del entregable:** Matricula_semanas3y4_actividad_02.ipynb

Criterios de evaluación

Esta actividad se evaluará con los siguientes criterios de evaluación:

Criterio	Valor
1 - Actualización de lista de stopwords.	5

2- Ajuste para los 1000 registros.	10
3 - Comentario datos perdidos.	5
4 - Limpieza de datos.	15
5 - Limpieza adicional.	10
6 - Segmentación y nube de palabras.	15
7 - Frecuencia mínima de tokens.	5
8 - Matrices enrtenamiento, validación y prueba.	5
9 - Entrenamiento de modelos con matrices count.	15
10 - Entrenamiento de modelos con matrices tfidf.	10
11 - Conclusiones.	5
Total	**100**