Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones 2020

Detección de Sarcasmo

Materia: Análisis y Curación

Introducción

Para este práctico realizaremos algunas tareas correspondientes a la limpieza de los datos, normalización y comparación de los datos con respecto a los obtenidos en el práctico anterior. Usando el corpus anterior no se filtraron por ejemplo los campos "NaN", se espera limpiar estos campos con algún criterio que ustedes consideren (eliminarlos, asignar un valor por defecto)

En el práctico anterior se tomaron los tres archivos correspondientes al corpus y se hizo un concat. Para este práctico, primero se realizará la curación de cada sub-corpus por separado y luego se procederá a trabajar con la concatenación de los mismos. Además se realizará el análisis TFIDF para analizar la variación de frecuencia de algunas palabras con respecto a los distintos documentos y con respecto al corpus concatenado.

A diferencia del práctico de Análisis y Visualización, usaremos la librería Spacy para el tratamiento de los datos.

Tareas a realizar

La entrega del práctico tendrá que tener el formato de una jupyter notebook (.ipynb) en donde se esperarán encontrar los siguientes análisis:

- Aplicar stemming y lematization. Comparar con los histogramas obtenidos en el práctico de Análisis y Visualización.
- Investigar acerca de los n-gramas, obtenerlos de ser posible.
- Identificar similitud de palabras, para esto usar distancia euclídea y del coseno.
- En el caso que haya palabras con mucha variabilidad, por ejemplo "2", reemplazar dichas palabras por un placeholder (DIGIT en ese caso).
- Intentar definir una cota inferior para la frecuencia de palabras y decidir qué hacer con aquellas que estén por debajo de esta cota.
- Hacer un análisis de reconocimiento de entidades y visualizar si es posible o si existe alguna relación entre estas entidades y las frases que son sarcásticas y las que no.

Visualizaciones esperadas

- Comparación de histogramas entre los resultados obtenidos en este práctico con el anterior
- Para los que no hayan realizado el heatmap, hacerlo.
- Visualización de n-gramas encontrados

Una vez realizadas estas tareas sacar conclusiones acerca de lo observado.

Entrega

La entrega será por mail donde se adjuntará la notebook con la que hayan resuelto los enunciados, deberá tener el nombre “grupoN_lab2” donde N pertenece al conjunto {1, 2} según el grupo al cual pertenezcan.
Como asunto del mail poner “entrega_mentoria”.

Las consultas que tengan las vamos resolviendo por los mismos canales habilitados para el práctico anterior.

Cronología tentativa:

Reunión orientativa: 15/07/2020
Reunión de seguimiento: 20/07/2020
Reunión pre-entrega: 24/07/2020
Entrega: 26/07/2020

A continuación, algunos links donde pueden sacar información sobre los temas necesarios:

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

https://en.wikipedia.org/wiki/Cosine_similarity

https://en.wikipedia.org/wiki/Euclidean_distance

https://spacy.io/