Maestría en Inteligencia Artificial Aplicada / MNA / Tecnológico de Monterrey
Actividad en Equipos - Semanas 6 y 7:
Modelado de Temas : LSI / LDA

ITESM Campus Guadalajara

Escuela de Ingeniería y Ciencias

Dr. Luis Eduardo Falcón Morales

Nombre(s): _____

Matrícula(s): _____

En la actividad de esta semana trabajarás en equipos en el tema de modelado de temas (topic modeling).

1. Descarga el archivo **noticiasTopicModeling.txt** que se encuentra en Canvas. Este archivo consiste en 5658 noticias de varios periódicos de España. El archivo de texto es una lista en el siguiente formato:
   [{"titular":"Encabezado", "texto":"Cuerpo"}, … , {"titular":"Encabezado","texto":"Cuerpo"}]
   Donde "titular" es el encabezado de la noticia y "texto" es el cuerpo del texto de dicha noticia. En particular en esta actividad trabajarás solamente con los cuerpos de las noticias, sin incluir los encabezados. Carga dicho archivo y genera un DataFrame de Pandas llamado "df" y que contiene una única columna llamada "noticia" con 5658 renglones formados por los cuerpos de las noticias.

2. Realiza un proceso de limpieza. Aplica el preprocesamiento que consideres adecuado para texto en español. Recuerda que el objetivo es identificar los tokens (palabras) que describan mejor la distribución de cada tema.
   NOTA: Recuerda que esta es una técnica no supervisada, por lo que no requerimos hacer una partición de los datos.

Parte 1: Indexación semántica latente (LSI):

3. Encontrar la matriz Tf-idf de la columna de noticias. Despliega los primeros 5 renglones con algunas de sus columnas con sus nombres, donde las columnas son los tokens. ¿Cuál es el significado de cada renglón? ¿Y el significado de cada columna?

4. Aplica el método de descomposición de valores singulares truncado a la matriz Tf-idf anterior con 10 componentes y obtener el gráfico de la importancia relativa de estas.

5. Obtener la matriz tokens-temas (term-topic) a partir de la matriz V^T de la descomposición SVD. Despliega sus primeros 5 renglones donde se incluya el nombre de las columnas.

6. Con base a la cantidad de conceptos latentes que determinaste en el ejercicio anterior, obtener cada uno de sus gráficos con sus 10 términos/tokens más importantes. ¿Cómo describirías cada uno de dichos conceptos latentes? ¿Se identifican claramente las temáticas de cada uno de ellos?

Parte 2: Asignación de Dirichlet Latente (LDA):

7. Utiliza la librería Gensim para implementar ahora la técnica de LDA. Revisa la documentación correspondiente y aplica de preferencia el modelo paralelizable:
   https://radimrehurek.com/gensim/models/ldamodel.html
   https://radimrehurek.com/gensim/models/ldamulticore.html

8. Con base a esta técnica ¿qué cantidad de tópicos consideras que es la más adecuada? Compara tus resultados con el método LSI. ¿Qué encuentras de coincidencias y diferencias? ¿Cuál consideras puede ser el mejor resultado, es decir, cuál consideras puede ser la mejor cantidad de tópicos a considerar?

9. Incluye tus conclusiones finales de la actividad.