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 que describan mejor la distribución de cada tema.

3. ¿Por qué en este caso no requerimos hacer una partición en entrenamiento, validación y prueba?

Parte 1: Indexación semántica latente (LSI):

4. 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?

5. 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.

6. 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.

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

8. Para cada uno de los 10 conceptos latentes obtenidos con la descomposición SVD del ejercicio 5, obtener el gráfico de barras que muestre los 10 términos/tokens más importantes de cada uno de ellos.

9. Con base a los resultados del ejercicio anterior, ¿cuántos tópicos o conceptos latentes identificas con claridad? ¿Cómo os describirías, es decir, de qué temáticas estarían hablando?

10. La librería de código abierto Gensim, incluye la implementación de la técnica LSI. Ver documentación: https://radimrehurek.com/gensim/models/lsimodel.html

    Investiga la documentación de este modelo LSI en Gensim y aplica dicha técnica a tu conjunto de datos. En particular, obtener los modelos de 2 a 10 temas/tópicos y para cada uno de ellos obtener su valor (score) de coherencia. Con base a dichos valores, ¿qué cantidad de tópicos sería el adecuado? ¿Coincide con lo que obtuviste previamente, usando la técnica SVD de manera directa?

Parte 2: Asignación de Dirichlet Latente (LDA):

11. 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

12. 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?

13. La librería pyLDAvis nos ayuda a visualizar de manera interactiva los temas y tokens de los documentos analizados. Revisa la documentación y utilízala para visualizar tus mejor resultado.

    https://pypi.org/project/pyLDAvis/

14. Incluye tus conclusiones finales de la actividad.

NOTA: Esta actividad está distribuida en 2 semanas. Aunque la primera parte de esta actividad se considera para evaluar la primera semana y la segunda parte para la segunda semana, la entrega de toda la actividad será hasta el final de la segunda semana en un solo documento. Al hacer la evaluación de la actividad se distribuirán las calificaciones en cada semana.