PRÁCTICA 3
IA en la empresa

Aplicaciones Avanzadas de la IA

Versión 1. 6 de marzo de 2025

Objetivo

El tema de esta práctica es el análisis de temas en documentos. Para ello usaremos una base de datos de noticias de prensa en inglés

Haremos dos tareas diferentes:

1) Clasificación de temas, modelo supervisado con los datos etiquetados en cinco temas diferentes.
2) Modelado de temas, no spuervisado, ignorando las etuqietas

El modelo supervisado será un clasificador multi-clase que prediga la categoría a la que pertenece un documento. Se entregarán tanto los resultados de validación cruzada como en un conjunto de datos de test independiente que se publicará más adelante.

Para la tarea de modelado de temas (*(topic modeling,* no supervisado) trabajaremos con todos los documentos, ignorando las etiquetas. Es decir, supondremos que hemos obtenido el conjunto de documentos y queremos identificar los temas principales tratados en los textos. El resultado será una agrupación (clúster) de documento similares y un análisis de los temas principales a partir de las palabras más frecuentes en cada clúster. El número de temas a utilizar es uno de los principales elementos a elegir en este modelo, y no necesariamente tienen que ser las cinco clases correspondientes a las etiquetas del conjunto de datos. Se recomienda valorar diferentes aproximaciones (k means, LDA,…), aunque finalmente no se prueben todas. Entre los resultados a entregar es recomendable incluir la nube de palabras más representativas de cada uno de los temas.

Conjunto de datos

Las noticias de prensa provienen del “BBC Full Text Document Classification” [enlace]. Es una colección de 2225 documentos del portal web BBC news de los años 2004 y 2005. Los documentos están etiquetados en cinco clases: business, entertainment, politics, sport y tech
La versión que usaremos contiene solo el 80% de los documentos, ya que el 20% restante se reserva para publicarlo como conjunto de test. Está disponible en Aula Global
El original contiene alrededor de 2200 documentos, con un volumen de 5Mb

Referencia del dataset:
D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.

Herramientas

Para alcanzar el objetivo de esta práctica, el alumno podrá utilizar la representación de los documentos y el las herramientas que considere más adecuadas.

Algunos ejemplos son

1. Representación como matriz de términos por documento
2. Representación con Word embeddings
3. Representación con sentence embeddings

Algunas herramientas recomendables son

- Python con librerías como nltk, spaCy, sentenceTransformers, gensim[i]
- Rapid Miner (ver la Guía disponible en Aula Global)
- Matlab

Resultados

Los resultados a entregar deben contener toda la información que se considere necesaria, por ejemplo:

- Resultados de la fase exploratoria de los datos realizada
- Descripción de la representación elegida para el texto, incluyendo el pre-procesamiento si es necesario.
- Proceso de refinamiento del modelo hasta llegar a la solución final (primeros pasos, limitaciones, cambios realizados)
- Tipo de modelo de clasificación, hiperparámetros, entrenamiento
- Método no supervisado para identificar los temas
- Resultados, eligiendo las medidas de rendimiento más adecuadas para cada problema
- Valoración de los resultados

Entrega de la práctica

- La práctica deberá realizarse en grupos
- La entrega constará de:
    - Memoria que detalle la propuesta desarrollada, las herramientas utilizadas, los resultados y su análisis, conclusiones, etc…

No hay un formato de entrega establecido. Sin embargo, es importante que toda la información se muestre de forma clara y su presentación también es considerada para la nota de la práctica.

- El código y otros elementos que sean necesarios para replicar los resultados

Evaluación de la práctica

Aspectos que evaluar en la corrección de práctica:

- Planteamiento y desarrollo del problema: 25%
- Resultados del problema: 25%
- Análisis de resultados y conclusiones: 25%
- Presentación: 25%

[i] Para utilizar un modelo de Word Embedding se utiliza Gensim es una biblioteca de Python de procesamiento de lenguaje natural (PLN), que permite construir Word Embeddings y realizar diferentes tareas de Topic Modeling. Además, ofrece una amplia variedad de algoritmos incluyendo Word2Vec, Doc2Vec, FastText, LDA y otros