Proyecto Final
Análisis Temático de Proyectos de Investigación

Procesamiento de Lenguaje Natural
Máster en Inteligencia Artificial Aplicada
Curso 2022/2023

December 1, 2022

1 Introducción

En este proyecto, los alumnos harán uso de los conocimientos y técnicas adquiridos durante el curso para resolver una tarea de aprendizaje sobre documentos textuales. Los alumnos trabajarán individualmente o por parejas sobre documentos que recogen los resúmenes de los proyectos de investigación financiados por la Comisión Europea dentro del Programa Horizonte 2020 (H2020).

Las tareas a realizar incluirán necesariamente:

- Procesado y homogeneización de textos
- Vectorización de documentos
- Modelado de tópicos con el algoritmo LDA
- Resolución de una tarea de clasificación
- Cálculo de distancias semánticas entre documentos, y cálculo y representación de grafos.

El proyecto tiene una valoración máxima de 4 puntos. Consta de dos partes:

- Proyecto básico: 3 puntos
- Extensión: 1 punto.

A continuación se indican los requisitos de cada una de las partes.

2 Descripción del conjunto de datos

El conjunto de datos seleccionado se proporciona en el fichero excel *"projects.xlsx"*. Dicho dataset consta de 35.378 entradas correspondientes a los proyectos financiados por la Comisión Europea dentro del Programa H2020. Para cada proyecto se proporciona la siguiente información:

- projectID: Identificador del proyecto
- acronym: Acrónimo del proyecto
- title: Título del proyecto
- summary: Resumen del proyecto
- startDate: Fecha de comienzo de ejecución
- endDate: Fecha de finalización
- totalCost: Coste total del proyecto
- ecMaxContribution: Financiación de la UE
- countryContr: Financiación recibida por cada país
- coordinatorCountry: País que coordina el proyecto
- euroSciVocCode: Código que codifica la temática del proyecto. Las categorías asociadas a cada código pueden consultarse en el fichero *"SciVocCodes.xlsx"*.
- publicationID: Identificadores de las publicaciones asociadas a cada proyecto
- patentID: Identificadores de las patentes asociadas a cada proyecto

La base de datos ha sido creada a partir de la información publicada por CORDIS en el portal de datos de la Unión Europea[1].

De forma alternativa, se permite usar otra base de datos que pudiera resultarle más interesante. En cualquier caso, si utilizara un dataset diferente al proporcionado, justifique brevemente su elección en la memoria, y asegúrese de que el dataset seleccionado permite realizar todas las tareas requeridas en este proyecto. En particular, el dataset debe incluir al menos un campo de texto no estructurado así como un conjunto de variables que permita la construcción de un modelo de clasificación.

[1] `https://data.europa.eu/data/datasets/cordish2020projects?locale=es`.

3 Proyecto básico

El proyecto básico consistirá en el análisis temático de la colección proporcionada, y en la construcción de un modelo de clasificación para determinar la categoría de la taxonomía EuroSciVoc asociada a cada proyecto. Además, deberá implementar un cuadro de mandos que permita la exploración conjunta de varias de las variables proporcionadas.

Los pasos que debe seguir en su trabajo son los siguientes:

- Paso 1: Implementación de un pipeline para el preprocesado de los textos. Para esta tarea puede usar las librerías habituales (NLTK, Gensim o SpaCy), o cualquier otra librería que considere oportuna.
- Paso 2: Representación vectorial de los documentos utilizando:
  - TFIDF
  - modelos de word embeddings
  - el algoritmo LDA
- Paso 3: Entrenamiento y **evaluación** de un modelo de clasificación para asignar a cada proyecto la **categoría de primer nivel** de la taxonomía de EuroSciVoc. Deberá explorar al menos las siguientes opciones:
  - Prestaciones al representar cada documento mediante su vectorización TFIDF, basada en word embeddings, o con las representaciones proporcionadas por LDA.
  - Prestaciones al utilizar únicamente variables basadas en los campos textuales, o al emplear adicionalmente los metadatos disponibles. Nótese que puede utilizar transformaciones de las variables proporcionadas, por ejemplo, el número de artículos asociados al proyecto, o el número de países participantes.
- Paso 4: Cálculo de distancias semánticas entre documentos, y cálculo y representación de un grafo.

En la memoria deberá incluir una descripción de los embeddings utilizados, así como del modelo de tópicos obtenido explicando cómo ha llevado a cabo la selección del número de tópicos. En cuanto al modelo de clasificación, debe discutir la metodología empleada para su entrenamiento, y analizar las prestaciones obtenidas según las variables de entrada. Las prestaciones deben estimarse empleando alguna metodología de validación que también deberá explicar en la memoria.

Finalmente, incluya en la memoria al menos una captura de uno de los grafos que haya calculado, y describa las opciones empleadas: e.g., criterio de coloreado o para el tamaño de los nodos, cálculo del *layout*, etc.

Tenga en cuenta que el objetivo de la memoria es describir el trabajo realizado y realizar un análisis crítico de los resultados obtenidos. Apóyese para ello

en gráficas u otras representaciones que considere oportunas. No es necesario describir los algoritmos utilizados, aunque sí deberá explicar cómo ha realizado el ajuste de sus parámetros.

4 Extensión

El trabajo de extensión es libre: deberá ampliar el proyecto básico en la dirección que considere más oportuna. Por ejemplo:

- Explorar el potencial de técnicas de NLP como el uso de bigramas, part-of-speech tagging, tesauros, etc.
- Exploración de alternativas a las vectorizaciones del proyecto básico, utilizando modelos de lenguaje contextuales (transformers).
- Resolución de un problema de clasificación empleando niveles más profundos de la taxonomía EuroSciVoc.
- Estudio temporal de tópicos o visualización dinámica utilizando la etiqueta del año.
- Generación de un cuadro de mandos para exploración conjunta de las variables disponibles.

Tome esta lista como una mera sugerencia, puede elegir cualquier otro tema siempre que encaje dentro del ámbito de la asignatura.

En el trabajo de extensión se valorará la creatividad y originalidad en la elección. Si tiene dudas sobre la idoneidad de la extensión elegida, consulte con el profesor. Evite embarcarse en trabajos de extensión demasiado ambiciosos que puedan comprometer la entrega en plazo del proyecto.

Si tiene cualquier duda sobre la idoneidad de cualquier trabajo de extensión, consulte con el profesor.

5 Entrega

Los alumnos deberán proporcionar los siguientes entregables para la evaluación del proyecto final:

1. Memoria descriptiva del trabajo realizado en formato .pdf y una extensión máxima de 14 páginas (excluyendo únicamente portada y referencias).
2. Script de Python con el código implementado debidamente comentado. En caso de haber implementado su código como un notebook, deberá descargar y entregar un script ejecutable (exportar como ".py")
3. (Opcional) Un vídeo corto (no más de 5 minutos) en el que describa su trabajo de extensión.

La memoria no debe incluir en ningún caso el código implementado, pero sí debe constar de cuatro apartados principales:

- Proyecto básico (máx. 10 páginas de memoria)
- Extensión (máx. 3 páginas de memoria)
- Manual de usuario del código, que explique cómo ha de ejecutarse el script entregado (máx. 1 página).
- Reconocimiento de autorías. Inexcusablemente, la memoria (o el notebook) debe respetar el principio de reconocimiento de autorías. Si ha utilizado fragmentos de código ajenos o cualquier material procedente de fuentes externas, debe especificarlo claramente en la memoria.

6 Evaluación

El proyecto se evaluará de acuerdo con los criterios siguientes:

- Proyecto básico (3 puntos)
  - Metodología (1.2)
  - Calidad de la memoria (1.2)
  - Calidad del código $(0,3)$
  - Reproducibilidad de los resultados $(0,3)$
- Extensión (1 punto)
  - Originalidad $(0,2)$
  - Calidad del trabajo y metodología $(0,8)$
- La evaluación por pares del trabajo de extensión mediante los vídeos enviados podrá suponer la obtención de $0,5$ puntos adicionales.

La entrega se realizará vía Aula Global. La fecha límite será el 30 de diciembre, a las 23,59 horas.