**Examen Interciclo Optativa Text Mining**

* Usted dispone de 115minutos para completar el examen
* Aquellos trabajos que no sean registrados en la plataforma virtual no tendrán una calificación
* Todos los archivos que sirvan como entrada a los diferentes procesos deben ser enviados. Si un proceso requiere un archivo que no está especificado en la práctica, entonces no se procederá a calificarlo.

**Objetivo**

El objetivo del examen es generar un corpus a partir de una página web sobre proyectos de investigación. La meta es construir un corpus de una lista de enlaces de direcciones URL en funcionamiento desde donde sea posible extraer contenido y luego analizarlo.

1. **Proceso de construcción del corpus**

El paso inicial es crear una lista de URLs que permitan acceder a información sobre proyectos. La fuente que se usará es el sitio web “ERC FUNDED PROJECTS”. Esta página contiene un índice con los países en donde se ha financiado proyectos por parte de la Unión Europea

<https://erc.europa.eu/projects-figures/erc-funded-projects>

Para poder identificar los tipos de extracción necesarios primero se requiere entender el corpus. Este proceso se puede inferir al ver la estructura HTML que es donde se encuentran los elementos de datos a extraer.

1. **Creación del Repositorio**

El primer procedimiento para construir el repositorio es generar una lista de URLs con identificadores únicos (representados por la variable **country** en la dirección URL) que define el nombre del país

La meta es extraer el nombre del país desde la sección “Country of Host Institution” y generar una dirección URL como la siguiente

<https://erc.europa.eu/projects-figures/erc-funded-projects/results?f%5B0%5D=country%3AAustria>

<https://erc.europa.eu/projects-figures/erc-funded-projects/results?f%5B0%5D=country%3ABelgium>

A pesar que la página ya contiene la URL en su conjunto será necesario crear un proceso que permita juntar esta información (URL + Country)

URL:<https://erc.europa.eu/projects-figures/erc-funded-projects/results?f%5B0%5D=country%3A>

Country: Belgium

Todo este listado de enlaces a los proyectos financiados en los distintos países deberá grabarse en una hoja de Excel.

1. **Extraer información relevante**

Desde cada URL almacenada en la hoja de Excel se deberá recopilar información del proyecto como:

* Nombre Proyecto (Project)
* Institución Anfitriona (Host Institution)
* Resumen del Proyecto (Summary)
* Financiamiento máximo (Max ERC Funding)

A efectos de reducir el costo computacional que significa acceder a toda la información en su conjunto, se pide priorizar únicamente aquellos proyectos que han recibido financiamiento en los países listados a continuación:

* Estonia (9 proyectos)
* Slovenia (8 proyectos)
* Romania (6 proyectos)

El acceso a estos enlaces (desde la hoja de Excel) puede ser ejecutado manualmente **solo en caso de que no sea posible** automatizar esta tarea.

Toda esta información debe almacenarse apropiadamente en el repositorio que lo desee, por ejemplo una hoja de Excel o una base de datos.

1. **Agrupamiento de proyectos**

Ejecute un proceso de agrupamiento de proyectos similares (de los países establecidos en el numeral previo) tomando en consideración el Resumen del Proyecto (Summary) o el Nombre del Proyecto (Project). Se debe identificar el número de agrupamientos más apropiado luego de efectuar varias pruebas sobre los datos.

Es necesario recordar que para ejecutar el proceso de agrupamiento será necesario primero ejecutar un proceso que permita almacenar las palabras más frecuentemente utilizadas en estos proyectos. Este conteo de palabras debe ser posterior a ejecutar un pre-procesamiento del resumen del proyecto que incluya al menos: tokenización, stemming y eliminación de stopwords. Cualquiera tarea adicional que influya en los resultados obtenidos será valorada en el examen.

En cualquier escenario se debe tener una identificación única del proyecto (p.ej. el país y título) previo al proceso de agrupamiento.

**ENTREGABLES**

1. El o los procesos que permitan cumplir con los requerimientos especificados
2. Hoja de Excel resultado del numeral dos (2)
3. Hoja de Excel o Base de datos donde se almacene información de los proyectos especificados en el numeral tres (3)
4. Explicación sobre el número de grupos (clusters) que ofrezca los mejores resultados, identificando además los valores de los centroides entre cada uno de los grupos
5. Identificación de los proyectos que pertenecen a cada grupo establecido