# 

Ingeniería Lingüística

2019-2020

Práctica 2

# 

# Alumnos:

# Lydia González Cid

# Carmen Bermejo Hernández

# 

# 

# 

# Índice:

# 1- Introducción

# 2- Preparación de los documentos

# 3- Entrenamiento

# 4- Pruebas de rendimiento

# 5- Conclusiones

# 1- Introducción

Se trata de hacer un clasificador de documentos que sea capar de distinguir entre diferentes categorías de noticias.

Para implementarlo hemos escogido el lenguaje Python y la librería Sklearn que incluye la mayoría de la funcionalidad que necesitábamos para esta tarea.

# 2- Preparación de los documentos

Hemos obtenido los documentos de la plataforma Social Clipping, que realiza una búsqueda de las últimas noticias en multiples medios online. Para el ejercicio hemos elegido las categorías salud, política y tecnología.

Hemos descargado el resultado de la búsqueda en estas tres categorías, obteniendo noticias de los últimos dos meses.

**2.1 – Scripts de limpieza y generación de documentos:**

Como ocurre en muchos casos cuando se trabaja con datos que no han sido preparados para trabajar con ellos, el documento obtenido presentaba varios problemas, que hemos resuelto implementando dos scripts.

El primer script, clean\_documents.py resuelve los siguientes problemas:

- Elimina líneas intercaladas en el documento que no pertenecen a la noticia, como la que indica el número de página de la búsqueda.

- Arregla caracteres especiales del español, como la ñ y los acentos, que la fuente original no leía bien.

- Añade un delimitador que separa claramente cada una de las noticias contenidas en el documento.

El segundo script, generate\_documents.py, parte los documentos en las diferentes noticias, generando un documento por noticia.

**2.2 – Sets de entrenamiento y test:**

Al descargar todos los documentos disponibles que nos permitía social clipping, esto es todas las noticias disponibles para cada categoría en los últimos dos meses, hemos obtenido diferente cantidad de noticias para cada categoría. 73 noticias de política, 64 de salud y 67 de tecnología.

Para el set de entrenamiento hemos escogido las 40 primeras noticias de cada categoría, dejando para el dataset de testing el resto.