Proyecto Final

Fecha: 3 de diciembre de 2018

Materia: Análisis y Procesamiento Inteligente de Texto

Maestro: Octavio Augusto Sánchez Velázquez

Alumno: José Antonio Velázquez Sánchez

Objetivo:

Hacer minería de tópicos en los subtítulos de películas hechos en distintos idiomas para:

a) Construir un clasificador de películas según su género

b) Encontrar similitudes entre vocabularios de los distintos idiomas

Estructura de la carpeta:

/clean_dataset – carpeta con el corpus previamente lematizado y limpiado

/download-scripts – carpeta con el código para descargar el corpus original

/otros-formatos – carpeta con los cuadernos de jupyter exportados a html y pdf

/pickles – carpeta con los objetos de python serializados que son utilizados en distintas ocasiones

0_bitacora-preliminar.pdf – Bitácora que fui llevando hasta que comencé a utilizar los Jupyter Noteboooks. Está documentado el cómo conseguí el corpus. Lo que se dice ahí se resume en el primer cuaderno de jupyter.

1_exploracion-del-corpus.ipynb – Se explora el corpus previamente lematizado y se crean y serializan las estructuras de datos que utilizaré en los siguientes cuadernos.

2_modelado-topicos-espaniol.ipynb - Se entrenan distintos modelos de LDA con distintos parámetros sobre los subtítulos en español. Se exploran manualmente los resultados.

3_clasificacion-con-topicos.ipynb – Se automatiza la elección de parámetros de LDA. Se crean modelos de tópicos y clasificadores de películas para inglés, español y francés.

4_exploracion-multilingual.ipynb – Se crea un corpus combinando los subtítulos de inglés, español y francés y se obtienen tópicos de él. Se exploran similitudes en el vocabulario entre estos idiomas.

myutils.py – modulo de funciones auxiliares utilizadas en los cuadernos jupyter

requirements.txt – listado de modulos instalados en el entorno virtual, necesarios para ejecutar los cuadernos de jupyter