Procesamiento de Lenguaje Natural (NLP)

Especialización y Maestría en Analítica Estratégica de Datos

2020-II

Plan de Proyecto Final

1. **Estudiante(s) — Máximo 2:**   
   •Leonardo Rubio Salcedo  
   •Angela Cristina Villate
2. **Título provisional del proyecto**

**¿Amor u odio?**

**Por qué nos gusta una película: caracterización de marcadores léxicos de éxito y fracaso a partir de técnicas de Twitter scraping & machine learning**

1. **¿De qué se trata el proyecto?**  
   Nos proponemos aplicar técnicas de procesamiento de lenguaje natural y de machine learning para caracterizar y analizar marcadores léxicos a partir de los que se pueda explicar por qué gustan o se rechazan películas, catalogadas como "cine comercial".

Y nos servirá para

* Comprender los comportamientos de consumo de cine de la población objeto de estudio.
* Determinar los factores que explican el éxito o el fracaso de una película.
* Identificar las expresiones lingüísticas asociadas a un sentimiento positivo, negativo o neutro.
* Aplicar técnicas de scraping y machine learning a fenómenos de consumo de masas y usos del lenguaje común.

1. **¿Cuál es la motivación?**

Utilizar y aplicar las herramientas de NLP para comprender el comportamiento de aceptación y rechazo de uno de los intereses comunes de los autores de este proyecto: el cine.

Queremos aplicar algunas de las metodologías aprendidas de la clase de NLP a situaciones cotidianas en las que se pueda tener acceso a registros del lenguaje en contextos actuales y espontáneos.

Las redes sociales son una plataforma en la que se expresan de forma sincrónica los modos en los que nos expresamos; es decir: el léxico que empleamos, los tiempos verbales en los que describimos eventos, las tradiciones discursivas (según el contexto de uso), etc.

Tradicionalmente, los estudios del lenguaje eran diacrónicos; esto es, se tenía que acudir a corpus históricos para reconstruir las formas de uso de las lenguas. Gracias a las herramientas y metodologías del NLP, podemos construir o acceder a corpus actuales y comprender los usos del lenguaje en tiempo real.

1. **¿Cómo se relaciona con NLP?**

Son varios los puntos de conexión:

* El propósito de este ejercicio es hacer un pequeño análisis del lenguaje natural a través de herramientas propias de NLP, como el Scraping.
* Se estructurarán (categorizarán y lematizarán) una cantidad considerable de datos
* Se analizarán los resultados según los contextos de uso del lenguaje (sentimientos positivos o negativos), lo que permite comprender algunas de formas estandarizadas en las que se expresan los mensajes.

1. **¿Dónde va conseguir los datos? ¿Son etiquetados?**

Contamos con dos fuentes de datos:

1. Base de datos con 4700 registros sobre películas: nombre, compañía, director, género, precio, puntuación, votos, escritor, año de lanzamiento, fuente IMBD.
2. Extracción de datos de twitter a través de métodos de scraping. Seleccionaremos hashtags relacionados con la muestra en los idiomas: inglés.
   1. Muestra
3. Extracción de datos de rotten tomatoes para obtener un corpus más robusto sobre las opiniones sobre las películas
4. **De las cosas que hemos visto (o vamos a ver) en clase, ¿qué piensa usar?**

**Sraping de datos**

* Extracción y preprocesamiento de datos
* Scraping de datos en twitter y páginas web
* Crear bucles para extraer datos de los tweets
* Cálculo de la distancia euclidiana entre las expresiones regulares TFIDF
* Proceso de Wor2vec a través de skip gram, que nos permite predecir contextos semánticos.
* Establecer las expresiones regulares: para limpiar los datos
* Repetir proceso cuantas veces sea necesario.

**Machine learning**

* Generar un clasificador de sentimientos a partir de una regresión logística que se aplicará a datos de entrenamiento

1. **¿Cómo piensa presentar los datos y sus resultados?**

Elaboraremos un informe en un notebook de Python en el que presentaremos:

**Los datos:** obtener información a partir de bases de datos y tweets que permitan identificar los indicadores de éxito o fracaso de la muestra de películas seleccionada para el análisis.

**Su procesamiento**: preprocesar los datos obtenidos teniendo como criterios categorías léxicas que permitan identificar los “sentimientos de los espectadores” en positivos y negativos. Las categorías léxicas se identificarán en inglés y español. Repetir el paso anterior cuantas veces sea necesario para obtener datos depurados para el análisis.

**Análisis:** del resultado obtenidos través de métodos de estadística descriptiva, correlación y comparación.

1. **Otros comentarios:**