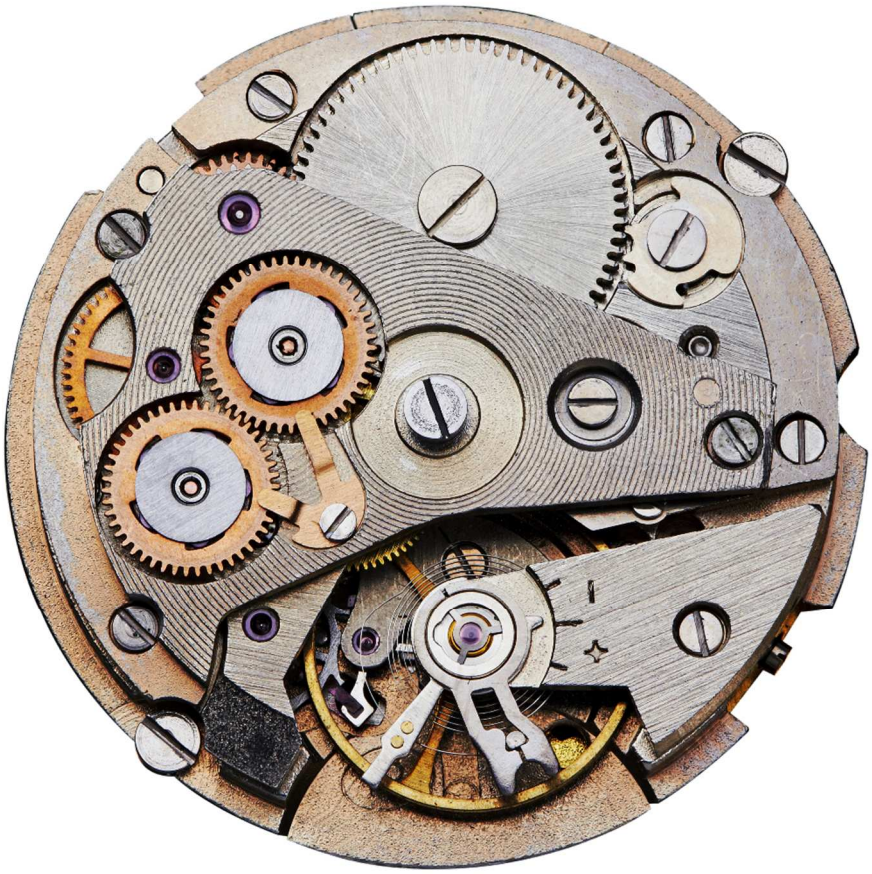
Deloitte.



Procesamiento del Lenguage Natural,

Análisis de sentimientos

20 de Septiembre, 2019

Autor: Miguel Ángel Fernández Guerrero, Analista de datos
mfernandezguerrero@deloitte.es

Introducción

Cada día se crean en internet más de 2,5 billones de bytes de datos, cifra que va en aumento. Saber tratar, analizar y entender esta información será una de las capacidades más demandadas en los próximos años. En este escenario las disciplinas del análisis de datos y el Machine Learning están siendo entendidas a día de hoy como el pilar fundamental sobre el que deben apoyarse decisiones en sectores de todo tipo, tan diversos como el Farmacéutico, el Bancario o el Marketing.

De toda la información que se genera actualmente, la mayor parte es en forma de texto, pensemos en aplicaciones de mensajería (Whatsapp), comentarios en redes sociales (Facebook, Twitter), búsquedas en internet (Google), opiniones en sitios web… Por esta razón, procesar la información en texto es actualmente objeto de investigación y gran interés. Éste será el contexto del problema propuesto.

Con el objetivo de introducir a los participantes en un problema real, les presentaremos Kaggle, una plataforma online de acceso público en la que compañías como Google, Netflix, Banco Santander y muchas otras compañias publican problemas reales donde datascientists de todo el mundo compiten por dar la mejor solución (comúnmente modelos predictivos). El problema propuesto ha sido publicado por Google.

***Key words**: NLP, Text Mining, Sentiment Analysis, Machine Learning, Binary Classification, Kaggle.*

Descripción del problema

Los alumnos dispondrán de dos sets de datos (en formato Excel o csv) que llamaremos **Train** y **Test**, cada uno compuesto por 25,000 textos en inglés de críticas a películas del portal IMDB. En el set Train, los alumnos sabrán si cada crítica es positiva (1) o negativa (0), pero desconocerán esta etiqueta para los textos en el archivo Test.

El problema será ajustar un modelo que sea capaz de clasificar una crítica en las dos categorías: positiva o negativa. Para ello los alumnos deberán emplear el set de datos Train y pasar por las distintas fases que requiere el Procesamiento del Lenguaje Natural. Estas fases se detallan en el plan de trabajo y contarán con un seguimiento en cada una de ellas. Además, para poder evaluar la calidad de los modelos ajustados, los participantes tendrán acceso a una dirección web (dentro de la plataforma Kaggle) en la que subir las predicciones de las críticas en Test y obtener así su precisión. De esta manera podrán comparar modelos y entender qué acciones hacen mejorar o empeorar el resultado.

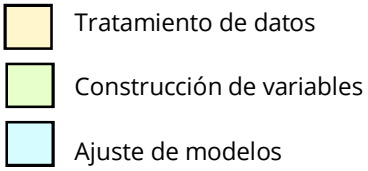
Objetivos

1. Conseguir una introducción al Machine Learning adecuada a unos alumnos de grado.
2. Conocer las distintas fases que requiere el Procesamiento del Lenguaje Natural.
3. Comprender el fundamento matemático de los distintos modelos que se plantearán a los alumnos (que irán de menor a mayor dificultad en función de cómo vayan avanzando con los tiempos).

Plan de trabajo

Octubre 2019

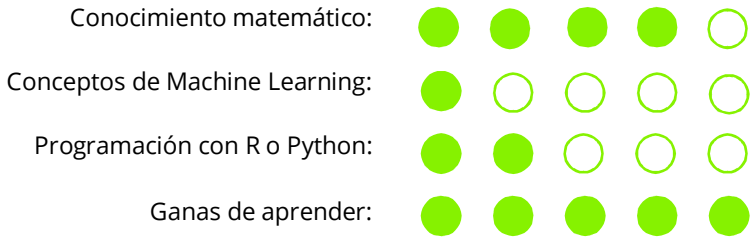
Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
	1	2	3	4 Kick Off	5	6
7	8	9	10	11 Reunión	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
8	9	30	1 Entrega			



Siguiendo la recomendación recibida, se fijará una reunión de seguimiento del día 11 de octubre con los alumnos. No obstante, los alumnos dispondrán de una persona que pueda dar apoyo y ayudarles previo aviso los días que sean necesarios.

Requisitos

En todo el proceso el alumno necesitará un ordenador en el que, apoyándose muy recomendablemente en software libre (R o Python), realizar cada una de las tareas:



Referencias Bibliográficas

a. A nivel teórico NLP https://web.stanford.edu/~jurafsky/slp3/
b. A nivel teórico NLP y apoyo en programación: Text Mining with R: A Tidi Approach (J Silge y D. Robinson)
c. A nivel teórico y práctico en Machine Learning : An introduction to Statistical Learning: With Applications in R (D.Witten, G. James, R.Tibshirani y T. Hastie)