Práctica 12 de Sistemas de Información para la Web

Ángel García Menéndez

Universidad de Oviedo

Introducción

A continuación se exponen los métodos y sistemas explorados a la hora de obtener un predictor de las fluctuaciones del índice *Down Jones* de la bolsa estadounidense mediante las noticias más relevantes del día.

Aproximación

Existen numerosos ejemplos de la forma de abordar el problema. El ejemplo canónico suele ser el de la clasificación de cuentos clásicos, y pese a las similitudes entre los problemas, existen notables diferencias a considerar.

Por un lado, en el caso de los cuentos lo que se explora es una propiedad inherente a los mismos, como es la temática. Se puede esperar una relación casi directa. Sin embargo, el caso que nos atañe es más complejo, pues la bolsa es un dominio ya de por sí errante y ya de por sí difícilmente predecible. De aquí se puede derivar que la correspondencia entre el contenido de las noticias y las variaciones del índice tendrán una relación menos clara, y por ende será necesario aplicar algo más de creatividad al problema.

Si se intenta elaborar el clasificador siguiendo el sistema de la navaja de Okham, agrupando los textos en *bags of words* y proceder a elaborar el clasificador con el resultado. Empero, el desempeño de esta técnica es bastante mejorable, siendo primero necesaria la eliminación de palabras vacía propias del dominio, y finalmente ninguno de los sistemas de clasificación ofrecidos por la herramienta consigue un resultado aceptable. El método alternativo empleado para el presente caso sigue la siguiente reflexión: la bolsa, como regla general, suele responder de buena forma a noticias positivas, mientras que ante las que se puedieran considerar negativas suele actuar de forma más recelosa.

Dentro del abanico de herramientas de *Orange* existe una que, dado un corpus, es capaz de dar una puntuación a los sentimientos que produce el texto, agrupados en:

- Positivo
- Negativo
- Neutro
- Conjunto

Así, es posible conseguir un valor numérico (el compuesto) que nos de una idea de los sentimientos producidos por las noticias, y de esta forma asignar un valor numérico que emplear a la hora de clasificar. Finalmente, se experimentará con diferentes técnicas y parámetros hasta encontrar una que de una respuesta aceptable.

Resultados

Los resultados experimentales han resultado bastante mediocres, pues apenas superan el 51%, o lo que es lo mismo, tiene una ligera mejoría con respecto a la total aleatoriedad. Además, tampoco se obtienen resultados tan malos como para poder recurrir a la negación del resultado del clasificador. Está en el punto en el que se puede decir que, simplemente, funciona mal. Las razones para esto pueden ser diversas, y es de interés el explorarlas. En primer lugar, habría que tener en cuenta la naturaleza del conjunto de datos. Si bien es verdad que los devenires munidales tienen un innegable impacto en el comportamiento del mundo bursátil, la naturaleza de este es compleja, y en determinados casos podría verse hasta arbitraria. La cantidad de factores que se pueden considerar, así como la forma en que responde el mercado a estos es un

campo de estudio en sí mismo, y por ende la relación entre las noticias más destacadas y las variaciones del *Down Jones* puede no ser tan directa como instintivamente pudiera creerse.

En segundo lugar, los sentimientos que transmiten las noticias no son tan claros y marcados como podría pensarse. Existe una tendencia en la prensa a resaltar lo negativo, y por tanto la tónica general es pesimista tanto en días de subida del índice como de bajada. Y también se puede apreciar en la gráfica que se genera en Orange que la distribución de los puntos no está muy bien diferenciada, es muy dispersa y no se aprecian concentraciones excesivamente evidentes. Por tanto se puede concluir que al no haber una distribución clara el algoritmo de clasificación no tiene matices claros con los que trabajar, de ahí el paupérrimo desempeño.

Finalmente, a modo de reflexión, una alternativa a la ejecución de este experimento pudiera ser un cambio en el conjunto de datos. La idea seguiría siendo la misma, sólo que en vez de emplear las noticias generales, se podría hacer uso de las publicadas en r/Economics, pues este subreddit está dedicado por entero a noticias de caracter económico, y generalmente pertenecen a medios norteamericanos, como el índice a analizar. Es entonces plausible que exista una relación más directa entre noticias y variociones, aunque casi con seguridad habría que descartar la idea del empleo de sentimientos, pues las noticias económicas suelen ser secas y técnicas.