PROCESAMIENTO DE LENGUAJE NATURAL
PROYECTO FINAL DEL CURSO
2013-2014

Introducción

Empresas de todo el mundo están interesadas en monitorizar las opiniones de los usuarios en redes sociales como Facebook, Twenty o Twitter. Esta información es clave para el desarrollo de nuevos productos, proyectos de marketing y evitar la difusión de opiniones negativas mediante la interacción con los usuarios. El caso de la red social Twitter, categorizada como "micro-bloging", es especialmente interesante dado que refleja opiniones en tiempo real (al igual que Facebook o Twenty) pero además es de libre acceso al igual que los blogs tradicionales. Por ejemplo, muchos consumidores consultan en Twitter antes de comprar un determinado producto.

El primer escollo al que se enfrenta la monitorización de opiniones en Twitter es detectar las referencias a la empresa o producto que se desea monitorizar. Por ejemplo, una ocurrencia del término "Subway", puede hacer referencia tanto a los establecimientos de comida rápida como a cualquier estación de metro.

El objetivo de este proyecto consiste en desarrollar un sistema de recuperación de información (sección 23.1 del libro base) que reordene un conjunto de 500 tweets, de forma que los tweets mejor posicionados referencien con mayor probabilidad a la marca en cuestión, ahorrando así tiempo al experto encargado de supervisar las opiniones en la WEB. Para ello se dispone de información a priori de la empresa. Concretamente, disponemos de 200 documentos recuperados con el motor de búsqueda de Yahoo usando como consulta el nombre completo de la marca para evitar ambigüedades. La metodología consistirá en generar un modelo de lenguaje (capítulo 4 del libro base) a partir de dichos documentos para luego re-ordenar los tweets según su perplejidad (sección 4.4 del libro base) respecto a dicho modelo de lenguaje. Los resultados se evaluarán en términos de curvas de precisión y cobertura y MAP (sección 23.1.4 del libro base) sobre un corpus de tweets manualmente clasificados como relevantes y no relevantes.

Este problema ha sido abordado en la competición internacional WEPS-3 (http://nlp.uned.es/weps/weps-3) como tarea 2: "Online Repuatation Management", en la que se generó un corpus anotado (relevante/no relevante) de unos 20.000 tweets asociados a 100 compañías distintas para evaluar sistemas desarrollados por distintos equipos de investigación. En este proyecto nos centraremos en 10 de esos nombres de compañías para los que más de un 40% y menos de un 70% de los tweets son relevantes, es decir, están relacionados con la compañía.

Material disponible.

Para la realización de este proyecto consideraremos 10 de las 100 entidades abordadas en la competición WEPS. La siguiente tabla describe el conjunto de esos diez nombres de compañías,la consulta no ambigua empleada para recuperar los documentos de referencia mediante el motor de búsqueda de Yahoo, la consulta empleada para recuperar los tweets que se desea re-ordenar, el número de tweets y el porcentaje de tweets anotados como relevantes en el corpus.

Término de consulta	**Nombre completo**	**Ratio de tweets relevantes**	**Página web**
harpers	Harper's Magazine	0,38	http://harpers.org/
mgm	MGM Grand Hotel and Casino	0,41	http://www.mgmgrand.com
boingo	Boingo Wireless	0,42	http://www.boingo.com
bart	Bay Area Rapid Transit	0,47	http://www.bart.gov/
luxor	Luxor Las Vegas	0,49	http://www.luxor.com/
rover	Land Rover	0,62	http://www.landrover.com
cadillac	Cadillac cars	0,63	http://www.cadillac.com/
bayer	Bayer Phramaceutical Chemical Company	0,66	http://www.bayer.com/en/homepage.aspx
fender	Fender Musical Instruments Corporation	0,67	http://www.fender.com/
blockbuster	Blockbuster Inc.	0,69	http://www.blockbuster.com

El material disponible para la realización de la práctica incluye, para cada nombre de compañía:

1. La lista de las 200 URLs devueltas por yahoo para cada nombre de compañía (directorio URLS).

2. Un directorio con los 200 documentos recuperados para cada nombre de compañía. Para la recuperación de estos documentos se ha empleado el paquete lynx de Linux. La colección no ha sido depurada, por lo que algunos documentos están vacíos o son PDFs (directorio DOCUMENTS).

3. Un documento de texto con la información de cada uno de los tweets que componen el corpus. Por cada tweet se incluye la siguiente información en un fichero de texto separados por tabuladores:
   - Nombre de la compañía
   - Número de tweet para la compañía.
   - Identificador general del tweet.
   - Contenido del tweet.
   - LABEL: Categoría (true/false) Esta categorización se ha revisado cuando existe falta de acuerdo entre anotadores.
   - Número de anotadores que anotaron el tweet como "related".
   - Número de anotadores que anotaron el tweet como "non related".
   - Número de anotadores que anotaron el tweet como "undecidable".
   - Meta-anotación.

Pasos a seguir.

1. En primer lugar, recomendamos familiarizarse con el problema. Para ello, conviene ojear las publicaciones accesibles desde la página WEB de la competición WEPS-3 (http://nlp.uned.es/weps/weps-3). En las publicaciones indicadas el problema se aborda en términos de clasificación, no de reordenamiento o ranking. Es decir, además de reordenar por relevancia los tweets, el sistema debe establecer un punto de corte entre relevantes y no relevantes. En nuestro caso, nos centraremos sólo en la fase de reordenamiento.

2. Generación de modelos de lenguaje a partir de la colección de documentos de referencia asociados a cada compañía. Entre las variantes en la generación de modelos, el estudio debe incluir distintas longitudes de n-gramas (sección 4.2 del libro base) y técnicas de suavizado o smoothing (sección 4.5 del libro base)

3. Cálculo de la perplejidad o perplexity (sección 4.4 del libro base) de cada uno de los tweets de la colección respecto al modelo de lenguaje correspondiente, y ordenación de los tweets de cada compañía.

4. Evaluación en términos de curvas de precisión y cobertura y MAP (sección 23.1.4 del libro base) de los resultados obtenidos, tomando como referencia los valores de relevancia anotados en el corpus.

5. Análisis de resultados.

6. **Voluntario:** Introducir en el sistema un filtro por entidades nombradas (sección 22.1 del libro base). Es decir, considerar únicamente expresiones que se correspondan con algún tipo de entidad en el proceso de aprendizaje. Para ello, puede emplearse la herramienta GATE (http://gate.ac.uk/).

Documentación a entregar

Se podrá usar cualquier lenguaje de programación. La documentación a entregar deberá incluir una descripción de:

1. El sistema desarrollado, tecnologías empleadas. modularización, etc.
2. Problemas encontrados durante el desarrollo y soluciones empleadas.
3. Analisis comparativo y justificación de los resultados obtenidos.