Procesamiento de Lenguaje Natural

Guión de prácticas final (Clasificador de opiniones)
Salud María Jiménez Zafra

Objetivos

Desarrollar un clasificador de opiniones empleando los conocimientos adquiridos a lo largo de las prácticas.

Descripción detallada

La práctica final consiste en el estudio, diseño y desarrollo de un sistema de clasificación de opiniones que reciba como entrada un texto y que devuelva como salida la polaridad que el sistema haya obtenido de acuerdo a una escala basada en dos valores: positivo y negativo. Para probar la validez del sistema se utilizará el corpus SFU, que está formado por comentarios escritos en inglés sobre 8 dominios diferentes: libros (books), coches (cars), ordenadores (computers), utensilios de cocina (cookware), hoteles (hotels), películas (movies), música (music) y teléfonos (phones). Este corpus se ha dividido en tres conjuntos de datos: train (para entrenar el sistema), dev (para probar el sistema durante la fase de desarrollo) y test (para evaluar el sistema).

Se propone el desarrollo de un sistema de clasificación basado en lexicón mediante el uso de la lista de palabras de opinión de Bing Liu[1] o, en su lugar, también se podrá desarrollar un sistema de clasificación supervisado o un sistema híbrido que combine ambos tipos de sistemas. Se valorarán positivamente los siguientes aspectos:

- Utilizar un pesado diferente para los adjetivos y verbos.
- Utilizar un pesado diferente para las oraciones introductorias y para las oraciones finales.
- Realizar un tratamiento específico de la negación.
- Realizar un tratamiento específico de los adverbios modificadores.

El sistema deberá producir como salida un archivo de texto con el siguiente formato:

nombre_archivo \t dominio \t polaridad

Ej. 10.txt BOOKS negative

Para poder probar el sistema, se proporciona el gold standard (archivo con las anotaciones correctas) del conjunto de desarrollo y un script que permite calcular la precisión, cobertura y exactitud del sistema.

Documentación a entregar

Se deberá entregar el código fuente generado, un archivo con las anotaciones que produce el sistema para el conjunto de test y un documento donde se detalle el funcionamiento interno del sistema. El documento

[1] Se trata de un lexicón formado por 2006 palabras positivas y 4783 palabras negativas indicadoras de opinión.

deberá contener, al menos, la siguiente información:

- Introducción, explicando de forma general cómo se ha realizado el clasificador.
- Desarrollo, con detalles más concretos de implementación y con las decisiones adoptadas.
- Ejecución de la aplicación.
- Parámetros de la aplicación.
- Resultados.
- Conclusión y valoración personal.

El resultado de esta práctica deberá entregarse en Docencia Virtual y tiene como **límite de entrega** las **23:59 horas del día 1 de mayo**. Se entregará un fichero .zip con el código fuente, con el archivo de anotaciones y con un documento en el que se detalle el funcionamiento del clasificador.