Laboratorio 2013

2da. Entrega

Resumen de Opiniones

Objetivos

El objetivo de este trabajo es familiarizarse con herramientas de Procesamiento de Lenguaje Natural (PLN), y utilizarlas para la resolución de problemas prácticos.
En particular, en este laboratorio se trabajará en la implementación de un sistema de resumen de opiniones sobre restaurantes. A lo largo de las entregas se pretenderá ir mejorando los resultados utilizando nuevas técnicas de PLN.

Descripción del Problema

En este laboratorio se intentará resumir los comentarios que los clientes de restaurantes han publicado en la web sobre los servicios recibidos en varios restaurantes de Montevideo.
El objetivo es procesar este corpus, extraer los comentarios e intentar resumirlos agrupándolos según su similitud.

Herramientas

En esta segunda parte del laboratorio se utilizarán las mismas herramientas de la parte anterior sumando las siguientes:

POS Tagger

Su función es realizar el etiquetado gramatical (Part of Speech tagging) de un texto. A cada una de las palabras de un texto asigna su correspondiente etiqueta gramatical, usualmente el texto se recibe previamente tokenizado. En este laboratorio se entrenará un POS Tagger a partir de un corpus etiquetado en español [1].

Wordnet

Wordnet [2] es una base de datos léxica que agrupa las palabras en conjuntos de sinónimos denominados *synsets*. También registra varios tipos de relaciones semánticas entre estos synsets. Originalmente fue creado para el idioma Inglés, pero se han realizado transformaciones [3] a otros idiomas como el Español, el cual utilizaremos en este laboratorio.

Clustering

Clustering es el proceso por el cual un conjunto de objetos es particionado en varios conjuntos o clusters. Los elementos de un mismo cluster poseen algún tipo de relación, que usualmente es medida a través de cierta medida de similitud.
Existen varios algoritmos y medidas de similitud para realizar tareas de clustering, en este laboratorio intentaremos medir la similitud de las opiniones de los clientes de restaurantes y agruparlas según su similitud utilizando paquetes de la biblioteca Scikit-learn [4].

Se pide

Se deberá construir un sistema que:

- Procese el corpus entregado.
- Extraiga comentarios del corpus.
- Analice los comentarios y calcule la similitud entre los mismos.
- Construya grupos de comentarios según su similitud.

Los pasos a implementar serán solicitados en cada una de las celdas descritas en el IPython notebook adjunto a esta letra.

Datos a Procesar
Los textos a procesar fueron obtenidos mediante un web spider que extrajo comentarios de clientes que expresaron su opinión acerca de varios restaurantes en el sitio Salir a Comer[1].

Formato y fecha de entrega
Cada grupo deberá entregar un archivo *.zip* de nombre *entregaGrupoX.zip* (donde *X* es el número de grupo) conteniendo:

- El informe en formato *pdf* que explique brevemente las etapas realizadas y analice los resultados obtenidos.
- El IPython notebook con el código de las soluciones.
- Otros archivos que consideren pertinentes a la entrega.

El trabajo puede entregarse hasta las 24 horas del día Viernes 11 de Octubre, utilizando un formulario que será habilitado en la página del curso.

Evaluación
Para la evaluación del trabajo se tomará en cuenta:

- Resultados obtenidos por la solución.
- La calidad del informe entregado, en particular la explicación y justificación de las decisiones tomadas, así como el análisis de los resultados obtenidos.

Referencias
[1] Wikicorpus - Sitio Oficial - http://www.lsi.upc.edu/~nlp/wikicorpus/
[2] Wordnet - Sitio Oficial - http://wordnet.princeton.edu/
[3] Multilingual Central Repository - Sitio Oficial - http://adimen.si.ehu.es/web/MCR
[4] Scikit-learn – Sitio Oficial - http://scikit-learn.org/

[1] http://www.saliracomer.com/