Laboratorio 2013

1da. 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

Este laboratorio se implementará completamente en el lenguaje de programación Python en su versión 2.7[1]. Python es un lenguaje multipropósito y multiparadigma, que entre otras cosas, es muy utilizado en el área de PLN. Como parte de sus librerías básicas "baterías incluidas" posee varios módulos y funciones que facilitan el procesamiento de texto (por más referencias sobre Python ver [1, 2, 3]). Algunas de las funcionalidades muy útiles para este trabajo:

Estructuras de datos:

- List
- Dictionary
- Set
- sorted (función para ordenar estructuras de datos)

Procesamiento de texto

- **open** (abrir archivos de texto e iterar sobre sus líneas)
- **len** (obtener el tamaño de un objeto, ej: largo de lista, de string, etc)
- **startswith/endswith** (chequear fin y comienzo de un string)
- **strip** (quitar espacios en blanco de inicio y fin de string)
- **split** (partir un string en pedazos según un patrón)
- **upper / lower** (convertir string a mayúsculas / minúsculas)
- **find / replace / count** (buscar / sustituir / contar dentro de un texto)
- **join** (transformar una lista de strings en un string)
- **Slicing** (tomar una subsecuencia de un string o lista)
  s[0] s[2:7] s[6:] s[:5] s[-1]
- Manejo de **Unicode** y diferentes **encodings**: *ascii*, *utf-8*, *latin-1*, etc.
  El módulo *codecs* permite abrir, escribir y guardar archivos de texto en el encoding seleccionado.
  Las funciones *encode* / *decode* permiten codificar / decodificar entre diferentes encodings y unicode.
- módulo **re** (módulo para expresiones regulares)

[1] Es obligatorio en este laboratorio utilizar la versión 2.7 y no instalar la versión 3.0 en donde se han hecho grandes cambios sin compatibilidad hacia atrás.

Para extender aun más la potencia de Python, utilizaremos las siguientes librerías:

Nltk
Natural Language Toolkit (NLTK) [4, 5, 6] es un conjunto de librerías de código abierto para PLN, implementadas en Python. Es una de las librerías más utilizadas para PLN y está compuesta por muchas funcionalidades, entre ellas:

- Tokenizador de palabras
- Tokenizador de oraciones
- Etiquetador gramatical
- Chunkers
- Reconocedor de entidades
- Expresiones regulares
- Gramáticas
- Parsers
- Stemmers/Lemmatizers
- Wordnet
- Algoritmos de Aprendizaje Automático
- Corpus, grandes colecciones de texto

Scikit-learn
Scikit-learn [7] es un conjunto de librerías de código abierto para *Aprendizaje Automático* (AA), implementadas en Python. Es una de las librerías de AA más utilizadas en el área. Además de poseer implementaciones de algoritmos de AA, también posee varias utilidades para PLN. AA es un área muy ligada a PLN por su gran utilidad en la resolución de problemas mediante algoritmos que aprenden a realizar una tarea a partir de grandes volúmenes de datos, en particular texto.

IPython Notebook
IPython Notebook [8] es un ambiente de Python que se accede a través de un navegador. Es un ambiente que permite trabajar con código Python de una forma muy agradable e interactiva. En él se pueden combinar ejecución de código, texto, gráficos en un solo documento.
En este laboratorio se utilizará IPython Notebook como ambiente de desarrollo. Las entregas se realizaran entregando los archivos generados por esta herramienta.

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[2].

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 20 de Setiembre, 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] Python - Documentación Oficial - http://docs.python.org/2/
[2] Dive Into Python - http://www.diveintopython.net/
[3] Python Essential Reference - Addison-Wesley Professional; 4 edition (July 19, 2009) – ISBN 0672329786
[4] Natural Language Toolkit – Sitio Oficial - http://nltk.org/
[5] Natural Language Processing with Python - O'Reilly Media; 1 edition (July 7, 2009) – ISBN 0596516495
[6] Python Text Processing with NLTK 2.0 Cookbook - Packt Publishing (November 11, 2010) – ISBN 1849513600
[7] Scikit-learn – Sitio Oficial - http://scikit-learn.org/
[8] IPython Notebook – Sitio Oficial - http://ipython.org/notebook.html

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