Laboratorio 2015. 2da Entrega
Categorización de textos

Objetivos
El objetivo de este laboratorio es familiarizarse con herramientas de Aprendizaje Automático (AA), y utilizarlas para la resolución de problemas prácticos.
En particular, en este laboratorio se trabajará con conjuntos de datos de tipo texto muy utilizados para experimentos y en la implementación de clasificadores.

Descripción del Problema
La categorización automática de textos es un problema de interés para las aplicaciones de búsqueda de información. Uno de los modos de categorizar es asignar el/los temas de los que un texto habla. En nuestro caso se consideran textos en idioma inglés, de conjuntos de datos presentes en repositorios de AA. En uno de los casos hay 20 temas posibles y cada documento pertenece a solo uno de ellos, en el otro hay más de 100 temas y el problema es multietiqueta, a cada documento le podemos asociar varios temas. El primer caso se tomará como instructivo, siguiendo el detalle de los pasos especificados en el *notebook* y en la *Guía del Usuario* de *sklearn.* Para el segundo, los propios estudiantes organizarán el *notebook* de acuerdo a las especificaciones que se establecen.

Herramientas
La solución se implementará completamente en el lenguaje de programación *Python* [2,3,4] en su versión 2.7[1]. Python es un lenguaje multipropósito y multiparadigma que, entre otras cosas, es muy utilizado en el área de AA.

Se utilizarán bibliotecas de Python especializadas en análisis de datos y AA.

- **Scikit-learn**
  Scikit-learn [6,7] es un conjunto de bibliotecas de código abierto para AA, implementadas en Python. Es uno de los ambientes de AA más utilizados en el área y cuenta con la implementación de varios de los algoritmos más conocidos.

- **IPython Notebook**
  IPython Notebook [8] es un ambiente de Python que se accede a través de un navegador. Permite trabajar con código Python de una forma muy amigable e interactiva. En él se pueden combinar ejecución de código, texto y 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.
  Luego de instalado Python, Pandas, Scikit-learn y IPython, para abrir un notebook se debe de ejecutar el comando:

  ```
  > ipython notebook
  ```

  en el directorio donde extrajeron los archivos del laboratorio. Esto abrirá una ventana de navegador donde podrá abrir el notebook que contiene las preguntas. Asociada a cada pregunta hay al menos una casilla (de tipo

[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.

código o texto) que los estudiantes deberán insertar y luego cargar con el código ejecutable o el texto requerido.

Los pasos a implementar serán solicitados en cada una de las celdas descritas en el IPython notebook adjunto a esta letra y, en la segunda parte, en las especificaciones generales que aparecen en el notebook.

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

- El IPython notebook con el código de las soluciones, las respuestas a las preguntas, las nuevas preguntas y soluciones definidas por los estudiantes.
- Otros archivos que consideren pertinentes a la entrega.

El trabajo puede entregarse hasta las 24 horas del día 3 de Noviembre, a través de la plataforma EVA.

Evaluación
Para la evaluación del trabajo se tomará en cuenta:

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

Referencias
[1] Catálogo de datos abiertos - Sitio Oficial - https://catalogodatos.gub.uy/
[2] Python - Documentación Oficial - http://docs.python.org/2/
[3] Dive Into Python - Sitio Oficial - http://www.diveintopython.net/
[4] Python Essential Reference - Addison-Wesley Professional (July 19, 2009) – ISBN 0672329786
[5] Python Data Analysis Library – Sitio Oficial - http://pandas.pydata.org/
[6] Scikit-learn – Sitio Oficial - http://scikit-learn.org/
[7] Learning scikit-learn: Machine Learning in Python – Packt Publishing (November 25, 2013) – ISBN 1783281936
[8] IPython Notebook – Sitio Oficial - http://ipython.org/notebook.html