Sistemas de Gestión de Bases de Datos
$1^{er.}$ cuatrimestre 2019

Ingeniería Informática
Universidad Nacional de Avellaneda

Guía Práctica 1
Procesamiento de texto

1. Introducción

Los objetivos de esta práctica son:

- Revisar algunos conceptos de expresiones regulares
- Realizar análisis de texto
- Visualizar resultados mediante histogramas y nubes de palabras

La bibliografía utilizada es:

- *Data Science From Scratch*, Joel Grus, O'Reilly

2. Ejercicios

2.1. Ejercicio 1

Realizar los siguientes puntos utilizando expresiones regulares:

2.1.1.

Escribir un programa que reconozca los símbolos de los números romanos: recibe un string **S** por teclado e imprime "TRUE" si todos los caracteres de **S** corresponden a símbolos de números romanos o "FALSE" en caso contrario. Por ejemplo:

- input: "XL" → output: "TRUE"
- input "#CienciaDeDatos" → output: "FALSE"

Obs.: no hay que determinar que el número sea válido. Por ejemplo el input "IIII" debería devolver "TRUE".

2.1.2.

Dado un string con el siguiente formato: "nombre1,apellido1,DNI1/.../nombreN,apellidoN,DNIN", escribir un programa que lo procese y escriba la siguiente información por pantalla:

apellido1 nombre1

...

apellidoN nombreN

2.1.3.

Escribir un programa que elimine los signos de puntuación de un string.

Nota: una fuente de información sobre expresiones regulares en python es `https://docs.python.org/2/howto/regex.html`

2.2. Ejercicio 2

2.2.1.

Escribir un programa para procesar el archivo “king_lear.txt”:

- Pasar las palabras a minúsculas
- Descartar los signos de puntuación
- Separar y contar las ocurrencias de las palabras (Evaluar si conviene utilizar un hash-map o la clase “collections.Counter” de python)
- Ordenar de modo descendente las palabras por cantidad de ocurrencias
- Responder
  - ¿Cuántas palabras tiene el texto?
  - ¿Cuáles son las 5 palabras más usadas?

2.2.2.

Escribir un programa basado en el punto anterior que considere filtrar el texto mediante un archivo de “palabras prohibidas”. Más precisamente: se requiere crear un archivo de texto que contenga una palabra por línea y aquellas palabras de “king_lear.txt” que estén contenidas en dicho archivo deben ser descartadas.

El objetivo de este proceso de filtrado es descartar aquellas palabras que aportan poca información sobre un texto (ej.: adverbios, artículos, proposiciones).

2.2.3.

Realizar las siguientes visualizaciones:

- Un histograma de las 10 palabras representativas con mayor cantidad de ocurrencias en “king_lear.txt” (Ver sección ”Bar Charts”, pág. 75 y 97))
- Una nube de palabras de las 50 palabras más representativas con mayor cantidad de ocurrencias en “king_lear.txt” (Ver capìtulo 20 del libro, pág. 334-336)

El objetivo de estas visualizaciones es intentar determinar algunas características esenciales (ej.: el tema, los personajes, las acciones) del texto en base a la cantidad de ocurrencias de las palabras.

Obs.: la noción de “palabra representativa” es subjetiva por lo tanto queda a criterio personal. Al menos habría que filtrar los adverbios, los artículos y las proposiciones.