Ingeniería del Lenguaje Natural

Sesión de Prácticas 2
Procesamiento superficial de corpus

Objetivos

- Desarrollar procesamiento básico de textos y corpus (cadenas y palabras).
- Valorar su importancia.
- Detectar problemas al trabajar en niveles tan superficiales
- Conocer herramientas básicas y sencillas para procesamiento superficial de textos.
- Analizar bibliografía básica y localizar soluciones.

Desarrollo

Dado un corpus (corpus_plano.zip), realizar los siguientes ejercicios:

1. Tokenizar cada fichero y contar las palabras de cada uno, considerando sólo las palabras diferentes (dos o más apariciones de la misma palabra se considera una solo).
2. Extraer la lista de frecuencias de palabras del total del corpus. Ésta debe ser una lista en la que en cada línea aparezca la frecuencia de la palabras (número de veces que aparece en el corpus) y al lado la palabra. Ordenado por frecuencias.

   Ejemplo: 2 casa

   5 coche

   67 se

3. Extraer los bigramas del corpus y su frecuencia, ordenado por frecuencias.
4. El índice de dificultad lectora de un texto se mide por la cantidad de letras de cada palabra que lo forman (*Mw*) y la cantidad de palabras que forman cada oración (*Ms*). Así, el índice de dificultad lectora se puede calcular como:

   $$4.71 * Mw + 0.5 * Ms - 21.43$$

   Calcular el índice de dificultad lectora de cada fichero del corpus. Indicar el más difícil de leer y el más fácil

   Ver http://en.wikipedia.org/wiki/Automated_Readability_Index

5. Reflexiona por escrito las siguientes cuestiones:
   a. Al contar palabras, ¿crees que es necesario algún tipo de conocimiento lingüístico para considerar todas las apariciones de la misma palabra? ¿De qué tipo?

b. ¿Hay palabras diferentes que tengan la misma forma? Si los hay, ¿cómo se pueden discriminar estos casos?

Bibliografía

Para desarrollar estos ejercicios debes leer los siguientes trabajos:

- Church, K. W. “Unix for Poets”
- Bird, Klein & Loper (2007) *Natural Language Processing in Python*, *http://nltk.sourceforge.net/index.php/Book*
  Cap. 2 Programming Fundamentals in Python
  Cap 3 Words: The Building Blocks of Language