Ingeniería del Lenguaje Natural

Sesión de Prácticas 3
Tokenización y normalización

Objetivos

- Tokenización de textos.
- Probar y analizar diferentes stemmers
- Consulta de materiales de teoría y reflexión sobre conceptos básicos.

Desarrollo

Esta práctica se debe realizar con python y NLTK. Para ello, abre Python e importa los módulos NLTK. Trabajaremos con expresiones regulares, por lo que importa también el módulo "re".

Para realizar estos ejercicios debes consultar el capítulo 3 del libro *NLP with Python* de Bird et al. Consulta la web del NLTK.

Dado el corpus en inglés "bible-kjv.projgutenberg.txt", que encontrarás en el Campus Virtual, realizar los siguientes ejercicios.

1. Abrir el texto, almacenarlo en una variable y eliminar los "retorno de carro".

2. Tokenizar el texto. Para ello utilizad el tokenizador del NLTK basado en expresiones regulares que encontraréis en el punto 3.3.1 del libro.

   Debéis definir varias expresiones regulares para que la tokenización no se base sólo en espacios en blanco. Definir las expresiones regulares necesarias para que incluya: palabras aisladas, signos de puntuación y nombres propios (varios nombres con mayúsculas seguidos). Sobre las expresiones regulares consulta el apartado 2.7 del libro.

3. Aplicar el stemmer de Porter y de Lancaster. Ambos están incluidos en NLTK. Consultar apartado 3.3.2. Extraer parejas de palabra "token, raíz".

4. Analiza las 100 primeras palabras y contesta a las siguientes cuestiones:

   a. Indica las principales diferencias que veas entre un stemmer y otro.

   b. Especifica dichas diferencias por categoría gramatical: nombre, verbo, adjetivo, adverbio, etc. ¿Localizas algún patrón de comportamiento?

   c. Según tu opinión, ¿cuál de los dos es mejor para aplicaciones de PLN?

   d. Indica alguna aplicación de PLN donde sea útil utilizar un stemmer.

   e. Indica qué desventajas puede tener hacer uso de stemmers.