Práctica 1

Procesamiento de Lenguaje Natural
Facultad de Ingeniería, UNAM

Objetivo: Preprocesar un corpus a partir de métodos basados en lenguajes formales.
Pasos a seguir:

1. Escoger un corpus de cualquier idioma y de un tamaño mayor a 10 000 tokens (se puede tomar este corpus de la paquetería *nltk.corpus*).
2. Limpiar el corpus: eliminar signos de puntuación, de interrogación, admiración y elementos no léxicos.
3. Eliminar las *stopwords* (se puede utilizar listas pre-hechas como las de nltk).
4. Aplicar un algoritmo de Stemming a los tokens limpios (p. ej. el algoritmo de Porter).
5. Obtener las frecuencias de los tipos en el corpus.
6. Obtener la lista de tipos por orden de frecuencia (de mayor frecuencia a menor frecuencia).

Puntos a evaluar:

1. Entrega a tiempo de la tarea.
2. Haber elegido un corpus adecuado y con las características indicadas.
3. Haber eliminado adecuadamente los signos no alfanuméricos y las stopwords.
4. Haber aplicado adecuadamente el algoritmo de stemming.
5. Haber obtenido las frecuencias de los tipos en el corpus en el orden indicado.