Práctica 2

Procesamiento de Lenguaje Natural
Facultad de Ingeniería, UNAM

Dado el corpus adjunto (corpusHMM.txt), realizar un etiquetador POST tomando en cuenta los siguientes criterios:

1. El corpus está dividido en dos columnas, una con palabras y otra con sus etiquetas. La separación se ha dado por un tabulador. Por ejemplo:

   ```
   frequently \t RB
   ```

2. Realizar un modelo del lenguaje $\mu = (\Sigma, A, \Pi)$ con las etiquetas; es decir, utilizar las etiquetas de la segunda columna del archivo para generar el modelo (NO con las palabras).

3. Para generar el modelo del lenguaje usar un smoothing Laplaciano.

4. Generar una matriz de probabilidades de emisión $B = (b_{ij}) = p(o_i|s_j)$, donde s_j es una etiqueta y o_i es una palabra u observación (éstas deben pasarse a minúsculas).

5. Para la matriz de probabilidades de emisión usar también smoothiong Laplaciano.

6. Guardar el Modelo Oculto de Markok $HMM = (\Sigma, A, \Pi, B)$ en un formato legible por python (por ejemplo con la librería pickle).

7. Programar el algoritmo de Viterbi con base en el modelo oculto de Markov obtenido. (¿Qué pasa cuando el algoritmo no vio una palabra en el entrenamiento?)

8. Probar el etiquetado con dos oraciones arbitrarias.