Instrucciones para la Tarea 2

Realizar un modelo de lenguaje $\mu = (A, B, \Pi)$ a partir del corpus «corpus_entrenamiento.txt». Recuerde que:

- A es la matriz de transición tal que $i \in \{1, ..., n\}, a_i j \notin \Pi(\sum_j = 1^n a_i j = 1)$.
- B es la matriz de probabilidades de observaciones tal que $\forall i \in \{1, ..., n\}(\sum_j = 1^m b_i j = 1)$
- P_1 es un vector de probabilidades iniciales tal que $\sum_{i=1}^{n} \pi_i = 1$.

El corpus se compone de **333 documentos** cada uno señalado por "*document*" : [...]. Dentro de cada documento se encuentra un token con su respectiva etiqueta, ambos dentro de llaves {...}. Las indicaciones por cada documento son las siguientes:

- El token se señala por "*token*" : "..."
- La etiqueta se señala con "*tag*" : "..."

Recuerde que B es una matriz de $token \times tag$ y que A es una matriz cuadrada de $tag \times tag$. También recuerde que las probabilidades iniciales de Π deben corresponder a las etiquetas de inicio de cada documento. Es decir, el vector Π es del tamaño del número de elementos tag.