Lingüística Computacional
Ejercicio 12

Descargar del Proyecto Gutenberg las siguientes obras, que podríamos considerar como un pequeño corpus literario contemporáneo (más de dos millones de palabras):

Baroja (La Busca, Las Inquietudes de Shanti Andía, El Árbol de la Ciencia); Blasco Ibáñez (La Catedral, Los Argonautas, Los Cuatro Jinetes del Apocalipsis); Fernán Caballero (La Gaviota); Concha Espina (La Esfinge Maragata); Gómez de la Serna (La Quinta de Palmyra); Palacio Valdés (Marta y María, La Hermana San Sulpicio); Pardo Bazán (Los Pazos de Ulloa, La Madre Naturaleza); Pereda (Peñas Arriba, Sotileza); Pérez Galdós (Fortunata y Jacinta, Marianela, Cádiz); Valera (Pepita Jiménez, Juanita la Larga).

Depurar. Analizar con Freeling.

1. (10 points) Utilizar el Corpus obtenido con las 20 obras como modelo para extraer los parámetros de un HMM: probabilidades de transición y probabilidades de emisión. Considerar sólo la categoría principal.

   Obtener la secuencia más probable de categorías para:

   las víctimas de delitos sexuales soportan a menudo una carga excesiva y viven como una segunda violencia un sistema que las hace recordar una y otra vez el horror vivido (De *El País* del 10 de abril de 2021)