Práctica de Extracción de Entidades

Objetivo

El objetivo de esta práctica es realizar un script que extraiga las entidades que se presenten en un texto. Las entidades a extraer serán de tipo organización, localización y persona.
Para realizar esta tarea se utilizará un sistema ya entrenado, fundamentalmente el entrenamiento con que se cuenta en NLTK.

1. Obtenga dos ficheros en formato txt, uno de ellos en ingles y otro en español, que contengan textos con entidades a extraer. Busque textos con un conjunto amplio de entidades, no se limite a dos o tres de cada tipo
2. Lea estos textos desde un programa Python, y realice las labores de pre-procesamiento habituales: división en frases, división en palabras y conversión a las formas normales. Utilice instrucciones como **nltk.sent_tokenize()**
3. Realice un análisis morfológico de los términos incluidos. Utilice **nltk.pos_tag()**
4. Por último, extraiga entidades con la función **nltk.ne_chunk()**
5. Una vez realizados todos los análisis, liste en pantalla las anotacioes obtenidas: frases, palabras, categoría gramatical del análisis morfológico, localizaciones, organizaciones y personas
6. Comente las diferencias entre los resultados obtenidos en español y en inglés.
7. Intente obtener mejores resultados en español. Para ello se sugiere utilizar la librería **spaCy**. Como punto de partida consulte las siguientes páginas WEB:

    - https://spacy.io/models/es
    - https://spacy.io