Trabajo Práctico - Procesamiento de Lenguaje Natural

Ciencia de Datos

Lectura recomendada:
https://web.stanford.edu/ jurafsky/slp3/15.pdf

Asociación de palabras[1]

1. Levantar el corpus AP, separando cada noticia como un elemento distinto en un diccionario (`<DOCNO> : <TEXT>`).
2. Calcular el tamaño del vocabulario.
3. Para las 500 palabras con más apariciones, calcular el par más asociado según la medida presentada.

Información Léxica[2]

Bajar de Project Gutenberg el libro de Darwin *ON THE ORIGIN OF SPECIES*.

1. Procesar el texto, tokenizando eliminando signos de puntuación.
2. Siguiendo el artículo de la sección, calcular la autocorrelación para estimar la distribución de la palabra a lo largo del texto.
3. Armar una función que reciba una lista de tokens, una lista de palabras y un tamaño de ventana y devuelva una lista de probabilidades de encontrar la palabra en cada ventana para cada palabra pasada por parámetro.
4. Calcular la entropía de la distribución de palabras seleccionadas para distintos tamaños de ventana
5. Generar una versión randomizada del texto, y medir la entropía de las palabras randomizadas.
6. Distinguir las palabras del texto en artículos, sustantivos y adjetivos usando un POS-tagger. Verificar si las medidas separan a estos grupos de palabras.

Word embeddings, distancia semántica y WordNet

1. Utilizando el test WordSim353[1], comparar el rendimiento entre LSA[3] y Word2Vec[2][4].

2. Comparar los distintos *word embeddings* con las medidas definidas en WordNet.

Referencias

[1] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[2] Marcelo A Montemurro and Damián H Zanette. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153, 2010.

[3] Thomas K Landauer. *Latent semantic analysis*. Wiley Online Library, 2006.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[1] `http://alfonseca.org/eng/research/wordsim353.html`

[2] Ver *pre-trained word vectors* en `https://code.google.com/archive/p/word2vec/`