

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. Calcular la entropía de una selección de 100 palabras que abarquen el rango de frecuencia de aparición en el libro, es decir elegir palabras que son muy frecuentes y otras de baja frecuencia.
4. Calcular la entropía de las palabras seleccionadas anteriormente, randomizando su posición en el texto. Comparar con los resultados del punto anterior.

Word embeddings, distancia semántica y WordNet

1. Utilizando el test WordSim353¹, comparar el rendimiento entre LSA[3] y Word2Vec²[4].
2. Comparar los distintos *word embeddings* con las medidas definidas en WordNet.

¹<http://alfonseca.org/eng/research/wordsim353.html>

²Ver *pre-trained word vectors* en <https://code.google.com/archive/p/word2vec/>

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.