28/04/2023

Gilbert Márquez Aldana

B94560

Reporte #5

Esta semana me enfoqué en tener una idea más clara del proceso que va a conllevar la implementación de nuestro código y de la estructura general de nuestro diseño. Para ello me dediqué a completar el análisis y seguimiento del tutorial de tensorflow de Word2Vec, además de analizar código de la implementación de CBOW con tensorflow. Asimismo, investigué sobre la implementación de Word2Vec en la biblioteca Gensim y me leí un artículo que trata sobre la representación de palabras como vectores y la comparación entre las distintas técnicas que existen.

Efficient Estimation of Word Representations in Vector Space

Este artículo nos es de ayuda para fundamentar nuestra decisión con respecto a la arquitectura/modelo que vamos a utilizar para vectorizar las palabras. De acá primero decir que se usa Word2Vec porque se fundamenta en que dos palabras que comparten contextos similares, también comparten un significado similar y, en consecuencia, una representación similar en el modelo. Justo lo que nos interesa: identificar localizadores faltantes basándonos en el contexto.

Entre las dos arquitecturas, aunque no hemos tomado la decisión final, la que más se ajusta a nuestro problema es **Continuous Bag Of Words (CBOW)**. Este modelo se utiliza para predecir una palabra basada en el contexto, donde el contexto son *n* palabras antes y *n* palabras después. Esta *n* se conoce como tamaño de ventana, y es importante experimentar con su valor, pues mientras más grande se mejora la calidad de los vectores resultantes, pero también se aumenta la complejidad computacional.

Para la próxima semana nos vamos a dedicar al diseño e implementación del modelo de NLP, de modo que tengamos suficiente material para poder experimentar durante el próximo avance.

Referencias

Word2vec embeddings
https://radimrehurek.com/gensim/models/word2vec.html

Tensorflow word2vec cbow basic
https://gist.github.com/yxtay/a94d971955d901c4690129580a4eafb9

Tensorflow – Word2Vec tutorial
https://www.tensorflow.org/tutorials/text/word2vec

Efficient Estimation of Word Representations in Vector Space
https://www.researchgate.net/publication/234131319_Efficient_Estimation_of_Word_Representations_in_Vector_Space