24/03/2023

Gilbert Márquez Aldana

B94560

Reporte #2

Esta semana, nos reunimos con los profesores para explicar lo que habíamos hecho durante la semana y la forma que le estábamos dando a la investigación. Nos comentaron algunas cosas, como que debíamos encontrar algún dataset y que nuestro problema aún era difícil de comprender. Nos recomendaron demostrar que lo que estamos haciendo sí es viable y necesario. Algo parecido a una justificación de la investigación.

Luego me reuní con Daniel para comentar la retroalimentación brindada por los profesores. En definitiva, solo nos enfocaremos en el análisis de archivos HTML, pues la idea de también realizar un análisis de imágenes se descartó. Nuestra idea, consiste en utilizar Procesamiento del Lenguaje Natural, por lo que esta semana me dediqué principalmente a investigar y ver videos sobre dicho tema.

NLP in 5 minutes

Natural Language Processing Basics: Word2Vec, CBOW & Skip-gram

Word2Vec – Skip-gram and CBOW

Word Embedding and Word2Vec, Clearly Explained!!!

En estos videos y páginas web, se explica sobre NLP y cómo funciona Word2Vec, técnica de NLP que utiliza una red neuronal para asociar palabras. Probablemente nos sea de utilidad, pues con CBOW y Skip-gram se puede analizar el contexto para obtener una palabra faltante o encontrar un contexto a partir de palabras. Nosotros hemos pensado que nos será necesario analizar el contexto en los cambios de los archivos HTML.

Using HTML in NLP: En esta publicación, se habla sobre el hecho de que mucha información podría perderse al extraer HTML para NLP utilizando Beautiful Soup's, pues se tira toda la información de los markups. En estos existe información que podría ser relevante, por lo que se aconseja mantener cierta parte de dicha información. Se mencionan algunas herramientas html2text, Pandoc, html5lib, sentencepiece.

Además, leí el siguiente artículo:

Root Cause Analysis for HTML Presentation Failures using Search-Based Techniques: Es una continuación del artículo "Finding HTML Presentation Failures Using Comparison Techniques", cuyo objetivo es encontrar la causa principal del

error partiendo de un set de potenciales elementos HTML causantes y utilizando técnicas de búsqueda y algoritmos genéticos.

Para la próxima semana voy a investigar con más detalle las herramientas recomendadas en la publicación de “Using HTML in NLP” (html2text, Pandoc, html5lib, sentencepiece). También, voy a investigar más sobre testing, identificación y localización de errores en HTML. Así como ver si existen aplicaciones de Word2Vec a código y detección de errores.

Referencias

Natural Language Processing Basics: Word2Vec, CBOW & Skip-gram

https://medium.com/the-techlife/natural-language-processing-basics-word2vec-cbow-skip-gram-e0e034862b8c

NPL in 5 minutes
https://www.youtube.com/watch?v=CMrHM8a3hqw&ab_channel=Simplilearn

Sonal Mahajan, Bailan Li, and William G. J. Halfond. 2014. Root cause analysis for HTML presentation failures using search-based techniques. In Proceedings of the 7th International Workshop on Search-Based Software Testing (SBST 2014). Association for Computing Machinery, New York, NY, USA, 15–18. https://doi.org/10.1145/2593833.2593836

Using HTML in NLP https://skeptric.com/html-nlp/

Word2Vec – Skip-gram and CBOW https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

Word Embedding and Word2Vec, Clearly Explained!!!
https://www.youtube.com/watch?v=viZrOnJclY0&ab_channel=StatQuestwithJoshStarmer