Notas Reunión 09/03/2022

Presentes en la reunión: Prof. Benjamín Bustos, Matías Vergara

Trabajo realizado entre 03/03/2022 y 09/03/2022

- **Investigación de técnicas.**

→ Se revisa capítulo 7 del Modern Information Retrieval de Ricardo Baeza, 2da edición, correspondiente a procesamiento de texto.

→ En el libro no se declara un término como "normalización" u "homogeneización" sino más bien se habla de "procesamiento" del texto.

→ En el libro no se menciona lemmatization, solo stemming y otras técnicas como tesseract, que no aplican a nuestro caso.

→ Se revisa el libro Introduction to Information Retrieval de D. Manning, Raghavan y Schütze (apunte del curso de Stanford). Allí se explican en detalle stemming y lemmatization, así como las diferencias y ventajas de cada uno.

- **Se decide utilizar lemmatization y no stemming.**

→ La razón para esta decisión es, principalmente, que de usar stemming luego habría que agregar un post-procesamiento de etiquetas que las devuelva del *stem* a la etiqueta como tal (a fin de que sea útil para sugerencias).

→ Además, stemming no tiene la ventaja de considerar los contextos de las palabras a fin de identificar sinónimos, a diferencia de lemmatization (que lo logra basándose en WordNet).

- Se llevan a cabo experimentos preliminares

→ Se realiza un procesamiento de las etiquetas quitando en primer lugar los espacios, comillas y paréntesis.

→ Se trabajan luego las *stop words,* que son eliminadas mediante funciones del Natural Language Toolkit (NLTK) de Python.

→ Una vez eliminadas las *stop words*, se aplica lemmatization. Este proceso destaca en solucionar el problema de reunir plurales y singulares de un mismo sustantivo.

→ Luego se agrega un paso extra en el que cada etiqueta compuesta por más de una palabra es separada para dar lugar a una etiqueta por palabra.

→ Como resultado, se tienen ahora **339 etiquetas distintas.** Considerando que en un inicio esta cantidad era de 586, se consiguió entonces una **reducción del 43.25%.**

→ Se detectan sin embargo algunos problemas menores producto del procesamiento. Un ejemplo es que etiquetas como "andrew's cross" ahora son "andrew's" y "cross" por separado. Sin embargo, se cree que este aspecto no debiera representar un problema mayor, pues los algoritmos deberían ser capaces de reconstruir la relación.

Trabajo a realizar entre 09/03/2022 y 16/03/2022

→ Mapear la transformación de etiquetas de vuelta hacia los patrones, y con ello probar cómo cambian los resultados obtenidos preliminarmente.

→ En esta etapa se usarán los mismos descriptores utilizados hasta ahora, a fin de modificar una sola variable a la vez.

→ En experimentos posteriores se probará con los mejores descriptores obtenidos por el prof. Iván (para ello es necesario en primer lugar agendar una reunión con él a fin de entender las diferencias en las arquitecturas de origen).