Procesamiento de Lenguaje Natural (NLP)
Especialización y Maestría en Analítica Estratégica de Datos
2020-II

Plan de Proyecto Final

1. **Estudiante(s)**

   Arturo Hernández Carvajal

   Rubio de Jesus Vásquez Bustamante

2. **Título provisional del proyecto**

   ANALISIS DE QUEJAS RADICADAS EN LA SUPERINTENDENCIA DE SALUD Y ENTENDIMIENTO DE SUS PRINCIPALES CAUSAS RAICES EN LOS AÑOS 2019 Y 2020

3. **¿De qué se trata el proyecto?**

   En el contexto se busca analizar las quejas relacionadas al sistema de salud en Colombia, de acuerdo con las radicaciones de inconformidades, peticiones y quejas que ha recibido la Superintendencia de Salud durante los años 2019 y 2020. Para esto principalmente contamos con las observaciones (descripción detallada de la queja) de cada caso y tipificaciones generales del caso.

   Con lo anterior queremos identificar un modelo de Machine Learning que permita realizar clasificación de la queja y su causa raíz.

   El conjunto de datos principal lo podemos complementar con las siguientes fuentes de información:

   a. CIE 10: **C**lasificación **I**nternacional de **E**nfermedades
   b. Datos de la red de prestadores de salud

4. **¿Cuál es la motivación?**

   La idea surge del conocimiento previo que tiene Rubio del tema; ya que en conjunto con su esposa (quien trabaja y pertenece al medio de la salud) han observado posibles mejoras asociadas a los diferentes procesos. Con lo que con un análisis de quejas es posible encontrar y dar un mayor

fundamento al conocimiento previo. Así mismo se puede entender la influencia de los intermediarios de salud en la calidad del servicio.

Por otra parte, Arturo trabaja en áreas de servicio al cliente donde el análisis de quejas y entendimiento de causas raíz se vuelve de gran importancia para llevarlo a implementación a nivel laboral. Con lo cual este ejercicio permite enfocar y tener presente realidades, logros y dificultades que pueda haber en un proyecto de este tipo.

Adicionalmente estamos utilizando este mismo conjunto de datos y temática para los proyectos finales de análisis de caso, machine learning y visualización; con lo que esperamos ir complementando aplicación de técnicas y conocimientos de las diferentes materias con NLP.

5. ¿Cómo se relaciona con NLP?

En el caso de radicaciones de quejas, las observaciones (texto escrito por el asesor o por el mismo cliente) *"esconden"* más información que las tipificaciones resumen. Por lo que hemos considerado el análisis de textos como una herramienta para extraer información de gran importancia para entender la relación entre las quejas, sus causas raíz y los mismos prestadores del servicio de salud. Para esto debemos:

- Realizar limpieza de textos
- Usos de metodologías como las ya vistas TF - IDF, word2vec para identificas patrones.
- Uso de web scraping para adquisición de textos adicionales que complementen el conjunto de datos de radicación de quejas.

6. ¿Dónde va a conseguir los datos? ¿Son etiquetados?

Los datos son etiquetados, las bases de datos provienen de la Superintendencia de Salud. En general el conjunto de datos de quejas, se puede encontrar en Datos Abiertos Colombia www.datos.gov. co

7. De las cosas que hemos visto (o vamos a ver) en clase, ¿qué piensa usar?

- Limpieza de textos.
- Conteos de palabras, lematización, top de palabras
- TF – IDF
- Distancia del coseno
- Word2vec

- Adquisición de textos por web scraping
- Modelo de machine learning, enfocado en clasificación para predecir la causa raíz de acuerdo a ciertas variables que sean seleccionadas en el modelo.

8. ¿Cómo piensa presentar los datos y sus resultados?

- Estadísticos descriptivos en visualizaciones de Power BI o en tableros de alguna herramienta de software libre (RStudio)
- De acuerdo a los avances que se vayan realizando en la materia de visualización de datos presentación de gráficos adicionales (se ha propuesto presentar mapas de calor, para entender los tipos de quejas y sus volumetrías de acuerdo a las ciudades de radicación)
- Documento de proyecto basado en la metodología de CRISP – DM donde describamos cada fase del proyecto dando énfasis en los resultados obtenidos, procesos de limpieza del conjunto de datos y descripción del modelo de machine learning utilizado.

9. Otros comentarios: