Nicolás Torrealba – Eduado Jimenez
Priscila Rodríguez- Kiumarz Goharriz

Informe de avance 1: Proyecto final

1. Re-Scope

El objetivo del proyecto es sistematizar y visualizar información de las causas no contenciosas del Tribunal de Defensa de la Libre Competencia (TDLC) y, posteriormente, predecir cuál es la probabilidad de que el organismo emita una sentencia favorable para el denunciante. Del primer análisis realizado con text mining, se llega a la conclusión de que, si bien el objetivo es abordable, para el análisis de visualización se deberá analizar caso a caso, pues tener un corpus que sistematice toda la información no es lo más eficiente, tal como se puede ver en la ilustración 1. Posteriormente, para realizar la predicción, se deberán usar los documentos cuya correlación sea cercana a uno, pero es altamente probable, que los resultados no sean los más robustos. En este sentido, si bien el objetivo del proyecto sigue siendo el mismo, las expectativas en cuanto al poder de predicción que tendrá el modelo se han ajustado a la baja.

2. Datos

Los documentos se descargaron desde la página del Tribunal de Defensa de la Libre Competencia. Específicamente, se tomaron todos aquellos archivos .pdf clasificados dentro de la jurisprudencia como: instrucciones de carácter general; informes y; resoluciones. Las recomendaciones normativas, al estar tangencialmente relacionadas con las causas no contenciosas quedan fuera de la descarga.

3. Análisis Exploratorio

Como se observa en la nube de palabras -ilustración 1- de los documentos de causas no contenciosas TDLC; existen 6 grupos de densidades definidos, estos diferenciables según los colores que vemos. Nuestro desafío se basa principalmente en obtener información relevante a partir de la cual podamos obtener conclusiones para documentos nuevos que se quieran analizar. Claramente se observa que con frecuencia se hace mención a palabras con raíces: “merc”, “competent”, “tribunal”; lo cual hace mucho sentido con el contexto de nuestra propuesta. Por otra parte, son igual de usados los términos: “servici”, “oper”, “libr”, “inform”, “chil”. El resto de términos que se observan en la nube, si bien serán relevantes para futuros análisis, se analizarán en profundidad más adelante.

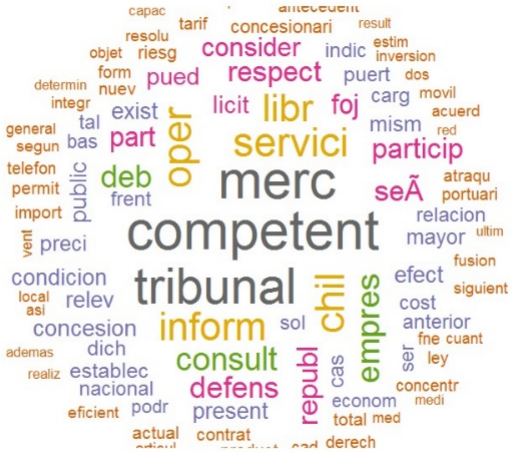

Ilustración 1

4. Dificultades

Si bien los informes se encuentran agregados de acuerdo a su categoría, los link con el reporte .pdf no funcionan en todos los casos, por lo que se debió realizar una exploración artesanal, que permitiera reemplazar la dirección del documento al que no se tenía acceso directo. Por otro lado, realizar el análisis de text mining ocupando las stopwords adeacuadas, resultó ser más complejo de lo que se tenía presupuestado inicialmente.