Búsqueda y recomendación para textos legales

Mentor: Jorge Pérez Villella

Integrantes: Zelaya, Adrián
Cardellino, Fernando

Índice de la presentación

- Objetivos de la segunda fase de la mentoría
- Revisión de resultados obtenidos con algoritmos de aprendizaje supervisado
- Revisión de resultados obtenidos con algoritmos de aprendizaje no supervisado
- Resultado obtenidos con la librería Gensim para recomendación de textos legales
- Conclusiones y Next Steps

Índice de la presentación

- Objetivos de la segunda fase de la mentoría
- Revisión de resultados obtenidos con algoritmos de aprendizaje supervisado
- Revisión de resultados obtenidos con algoritmos de aprendizaje no supervisado
- Resultado obtenidos con la librería Gensim para recomendación de textos legales
- Conclusiones y Next Steps

Objetivos de la segunda fase de la mentoría

- Aplicar técnicas de Aprendizaje Supervisado (Logistic Regression, Naive Bayes, y SVM) y No Supervisado (LDA y NMF), en base a los resultados obtenidos en la primera fase del proyecto
- Implementar un sistema de recomendación de textos legales similares a un texto que se provee como input.

Índice de la presentación

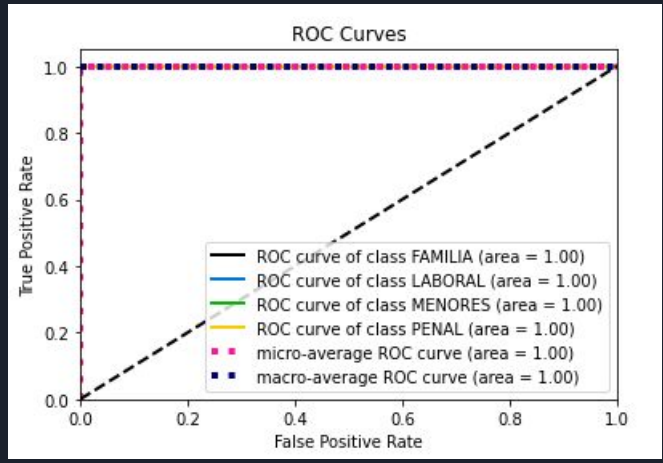
- Objetivos de la segunda fase de la mentoría
- Revisión de resultados obtenidos con algoritmos de aprendizaje supervisado
- Revisión de resultados obtenidos con algoritmos de aprendizaje no supervisado
- Resultado obtenidos con la librería Gensim para recomendación de textos legales
- Conclusiones y Next Steps

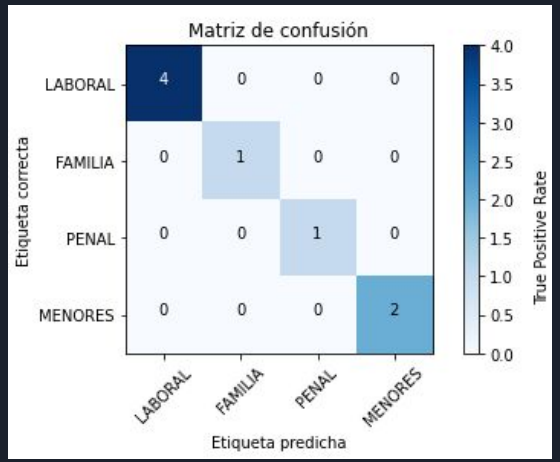
Resultados obtenidos con Aprendizaje Supervisado

Para clasificar los documentos por fuero, se corrieron y compararon modelos basados en algoritmos de Naive Bayes, Logistic Regression y SVM.

Si bien todos tuvieron una buena performance, en base al porcentaje de accuracy consideramos como mejores modelos los siguientes:

- Logistic Regression: **98.8% de accuracy**
- SVM: **98.2% accuracy**





Si bien Logistic Regression parece tener mayor accuracy según los resultados del grid search hyperparameter tuning, al correrlos con la data de evaluación, ambos muestran una accuracy del 100%.

Este elevado accuracy se debe en cierta medida a que se utilizaron pocos datos y con poca variabilidad entre los mismos.

Índice de la presentación

- Objetivos de la segunda fase de la mentoría
- Revisión de resultados obtenidos con algoritmos de aprendizaje supervisado
- Revisión de resultados obtenidos con algoritmos de aprendizaje no supervisado
- Resultado obtenidos con la librería Gensim para recomendación de textos legales
- Conclusiones y Next Steps

Resultados obtenidos con Aprendizaje No Supervisado

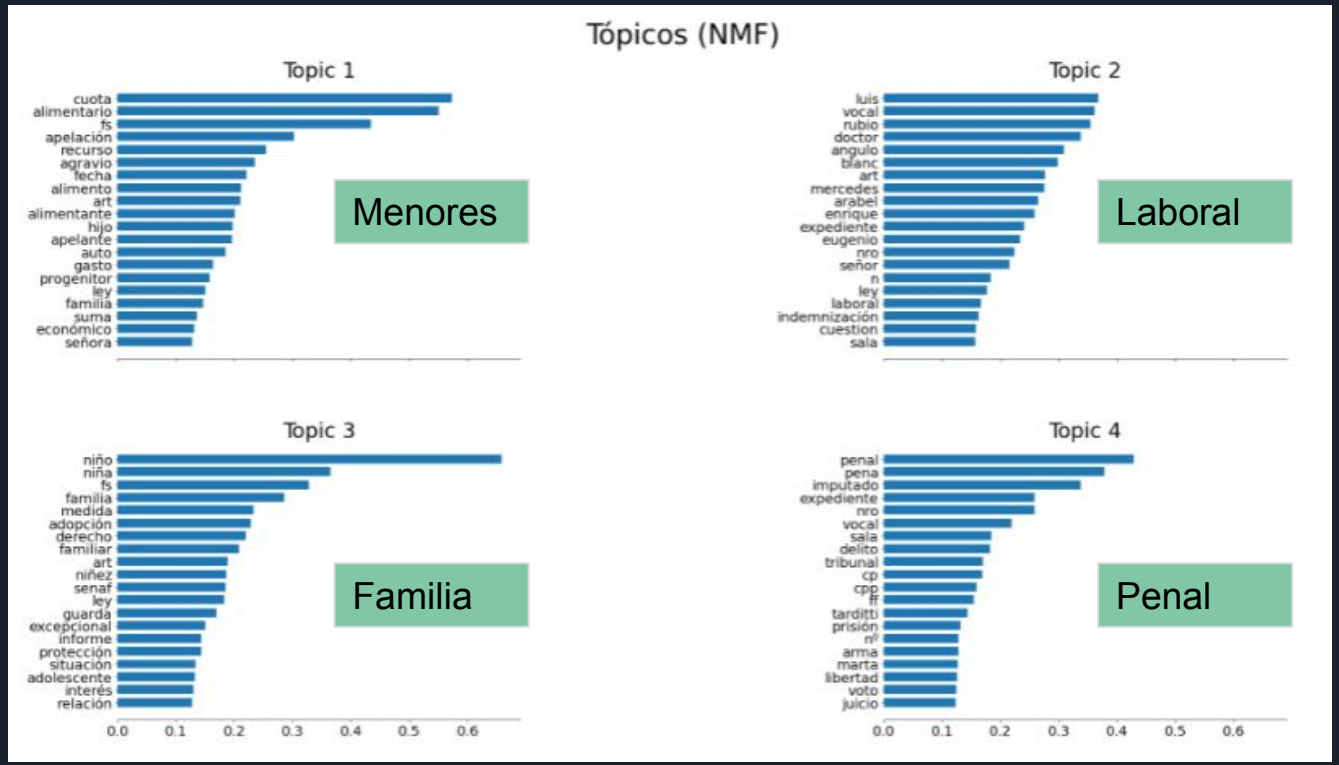
Se utilizaron los siguientes métodos para Topic Modeling:

- Latent Dirichlet Allocation (LDA)
- Non-negative Matrix Factorization (NMF)

De ambos métodos, el que presentó mejores resultados generando agrupaciones de textos fue NMF.

Esto tiene sentido ya que los vectores utilizados toman en cuenta la frecuencia inversa de ocurrencia en los documentos (TF-IDF).

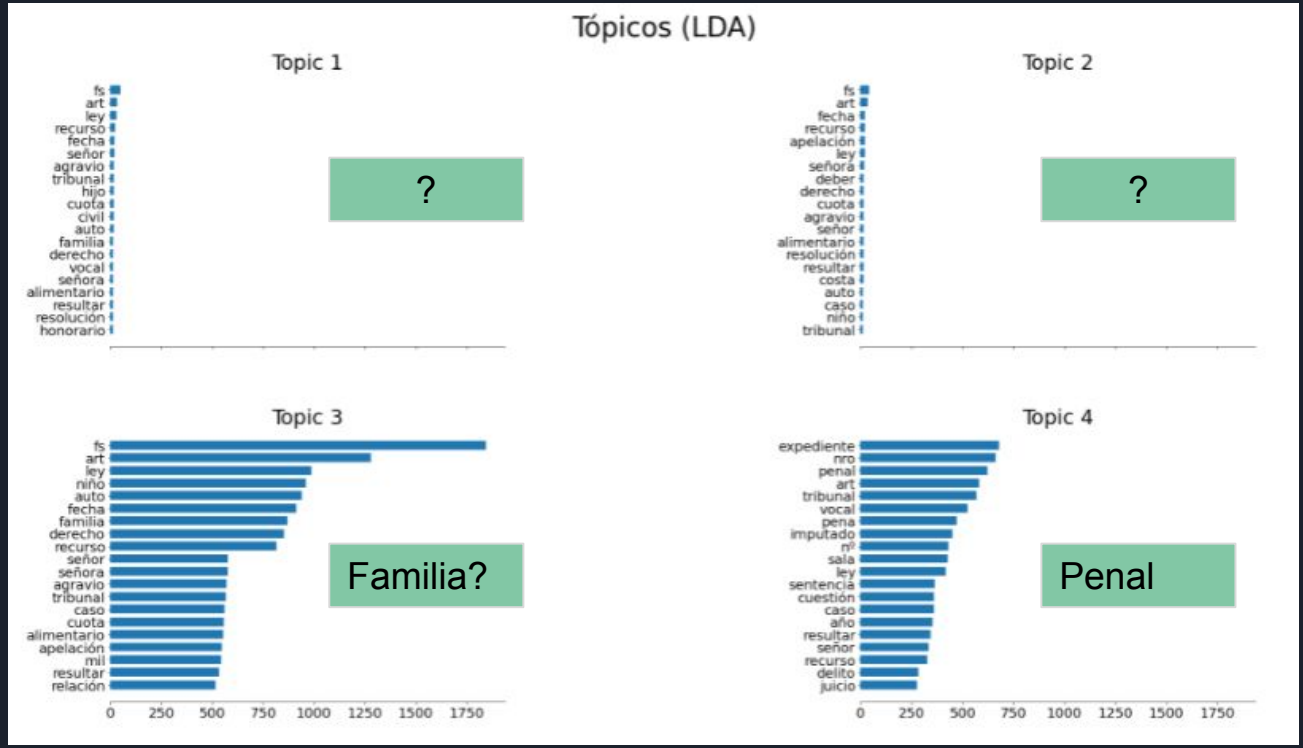
Resultados obtenidos con Aprendizaje No Supervisado



- El modelo NMF agrupó los textos según las categorías originales: Penal, Laboral, Familia, Menores.
- El modelo fue capaz de agrupar en tópicos los términos correlativos a todos los fueros.

Resultados obtenidos con Aprendizaje No Supervisado

- En el caso del modelo LDA, se sugirió tópicos con términos correlativos a solamente dos fueros: Penal y Familia.
- Los otros dos tópicos no son concluyentes. Esto se puede deber a que tanto las palabras específicas de cada fuero como las que son transversales tienen igual ponderación.



Índice de la presentación

- Objetivos de la segunda fase de la mentoría
- Revisión de resultados obtenidos con algoritmos de aprendizaje supervisado
- Revisión de resultados obtenidos con algoritmos de aprendizaje no supervisado
- Resultado obtenidos con la librería Gensim para recomendación de textos legales
- Conclusiones y Next Steps

Resultados obtenidos con Gensim

Implementando la librería Gensim se logró obtener, por cada fuero, los 5 documentos más similares a los textos que se utilizaron para evaluar el modelo.

`34 FERRETTI-HORIZONTE CIA ARGENTINA DE SEGUROS GENERALES.pdf.txt (LABORAL)`

	archivo	fuero	similarity
32	34 FERRETTI-HORIZONTE CIA ARGENTINA DE SEGUROS...	LABORAL	0.847813
25	64 MINO-MINISTERIO DE EDUCACION.pdf.txt	LABORAL	0.842251
10	14 MOLINA LILIANA - COOPERATIVA ELECTRICA Y DE...	LABORAL	0.837312
14	39 BENITO-SISTEMAS.pdf.txt	LABORAL	0.829025
18	219 LUJAN-KICZALUK.pdf.txt	LABORAL	0.825500

`7°-2018- C.M. y otros - Denuncia por Violencia Familiar.doc.txt(MENORES)`

	archivo	fuero	similarity
218	2°-2018- L., A. - Denuncia por Violencia Famil...	MENORES	0.648380
84	L, F A c. R, L A- Medidas urgentes .doc.txt	FAMILIA	0.555896
50	E., H. - DVF - GENERO - APELA ARCHIVO LA VICTI...	FAMILIA	0.554271
219	7°-2018- C.M. y otros - Denuncia por Violencia...	MENORES	0.553473
65	C.M.N. c. AFI.doc.txt	FAMILIA	0.535361

La única salvedad se da para los textos del fuero Menores, donde se recomendaron sentencias del fuero Familia. La explicación de esto puede radicar en que comparten términos similares.

Índice de la presentación

- Objetivos de la segunda fase de la mentoría
- Revisión de resultados obtenidos con algoritmos de aprendizaje supervisado
- Revisión de resultados obtenidos con algoritmos de aprendizaje no supervisado
- Resultado obtenidos con la librería Gensim para recomendación de textos legales
- Conclusiones y Next Steps

Conclusiones y next steps

Conclusiones:

- Se pudo llegar a la implementación de un prototipo de sistema de recomendación de textos legales por fuero.
- Se observó la necesidad de crear diccionarios de palabras específicas del ámbito jurídico para ampliar el universo de stopwords que generan ruido al momento de alimentar los modelos.
- Se determinó la mayor eficacia en modelos de aprendizaje no supervisado que ponderan las palabras por la inversa de la frecuencia (Tf-Idf). Con esto se logra agrupar textos con características similares en forma más precisa.

Next Steps:

- Probar con otros modelos como Transformers, Long Term Short Memory, etc.
- Lograr escalar el sistema para hacerlo más robusto y aplicable a otro tipo de textos legales.
- Ampliar el dataset

¡Muchas gracias!