Mentoría Textos Legales

Búsqueda y Recomendación de Textos Legales - Análisis y Visualización

Mentor: Claudio Sarate

Integrantes:

- Ezequiel Juarez
- Jorge Pérez
- Clara Quintana
- David Veisaga

Objetivos

- Búsqueda y recomendación de textos legales
- Agrupar textos similares (temáticas similares)
- Lo que buscamos es encontrar una manera de poder decir si un texto habla de lo mismo que otro, una medida de ***distancia*** entre 2 textos.
- Textos legales (InfoLEG)

Características del Corpus

- Cuando procesamos textos no contamos con una tabla de atributos o características. (como se puede ver en otras problemáticas)
- Cada palabra que contienen esos textos pasan a ser las características discretas y categóricas.
- Por otro lado, en el caso particular de esta mentoría, tampoco contamos con una columna de clases (como se puede ver en otras problemáticas)

Extracción de Características

- Las palabras deben codificarse numéricamente para que puedan ser utilizadas por los algoritmos de aprendizaje.
- Una de las técnicas más simples para representar texto numéricamente es la denominada “Bag of Words”.
- Para iniciar el pre-procesamiento contamos con una colección de textos legales seleccionados para el análisis, que almacenamos en un archivo denominado Corpus.

Limpieza de datos

1. Eliminar ruido
   - Formatos de archivos TXT, PDF, DOC, etc
   - Text Encoding: ascii, utf-8
2. Tokenización
   - Dividir el documento en palabras
   - Lengua Española
3. Normalización
   - Convertir el texto a minúsculas
   - Eliminar signos de puntuación
   - Corregir errores de ortografía
   - Eliminar stop words
   - Stemming
   - Lematizar

Limpieza de datos: Conclusiones

- Proceso iterativo
- Diferentes librerías para leer PDFs
- Idioma Español
- Abreviaturas y Errores de ortografía
- Lematizar y/o Stemming
- Stop word personalizado

Palabras Relevantes

- Para saber si 2 textos son similares
  - → de qué trata un texto.
    - → Palabras relevantes

Palabras Relevantes:
Resultados - Frecuencia de palabras

Palabras Relevantes:
Resultados - IDF para stop word

IDF:

Frecuencia Inversa

del Documento

	idf_weights
nacional	1.176183
artículo	1.195928
oficial	1.243746
registro	1.263845
archívese	1.284887
...	...
establecimiento instalaciones	8.131299
establecimiento instituirá	8.131299
establecimiento instrucciones	8.131299
establecimiento inferior	8.131299
útilmente según	8.131299

Palabras Relevantes: Resultados - Frecuencia sacando stop word

Eliminado las stop word obtenidas con con IDF

Las palabras más relevantes coinciden con las más frecuentes

Palabras Relevantes:
Resultados - TF-IDF

Palabras más relevantes usando TF-IDF

	tfidf
venezuela	0.337258
minera	0.311413
bolivariana venezuela	0.309724
bolivariana	0.309724
caracas	0.146379
metalíferos	0.106680
minera bolivariana	0.106680
cooperacion minera	0.106680
caracas bolivariana	0.101361
cooperacion	0.073029

Palabras Relevantes: Resultados

Buscamos las palabras más relevantes en cada documento

	CANT_PALABRAS	TOP_10	TOP_1
0	5815	venezuela minera bolivariana venezuela bolivar...	venezuela
1	577	suprema generan generan expropiación expropiac...	suprema generan
2	1794	remanentes tesoreria remanentes tesoreria vein...	remanentes
3	1360	reintegros restitución tarjeta magnética reint...	reintegros
4	6177	componente hogar uocra componente hogar gecal ...	componente

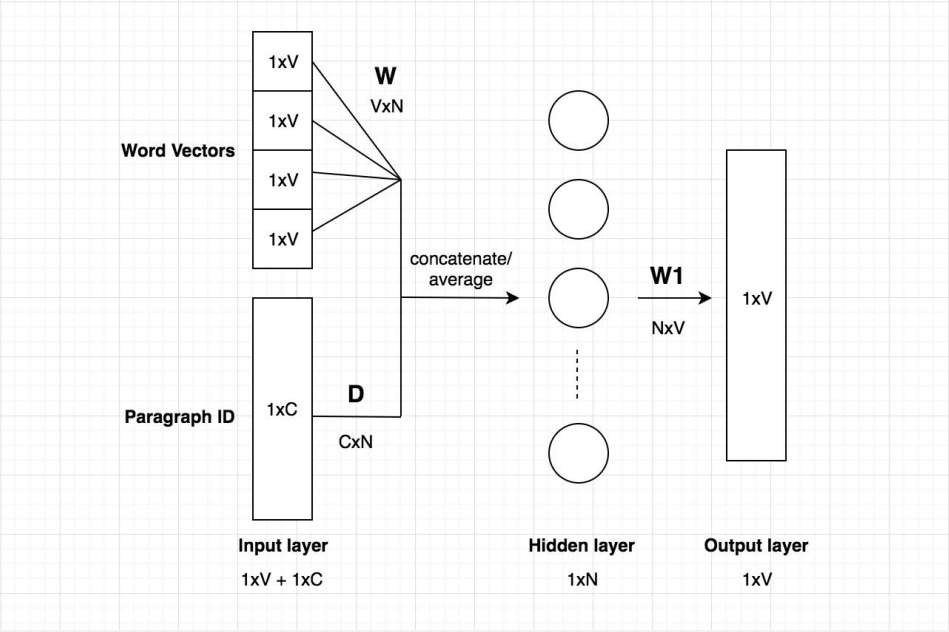
Intentamos generar clases y hacer aprendizaje supervisado, obtuvimos una accuracy del 61%

No supervisado

- LDA: modelo de tópicos generativo, asume que cada palabra en un documento es generada a partir de un tópico que es tomado de una distribución de tópicos para cada documento.
- NMF: Consiste en la descomposición de la matriz de frecuencia de palabras (V), en 2 matrices más pequeñas que representan los tópicos (H) y la relevancia de cada tópico en cada texto (W).

No supervisado

- Doc2vec: El modelo crea una representación vectorial de un documento utilizando la representación vectorial de cada palabra que compone el documento.



Conclusiones

- Proceso iterativo
- IDF es un buen método para encontrar stop words (éstas dependen de la temática)
- TF-IDF es una buena medida para encontrar palabras relevantes/representativas de cada doc
- No es un "problema" para ser tratado con aprendizaje supervisado.
- Aprendizaje no supervisado, todavía no hemos dado con la configuración apropiada para encontrar similitud entre documentos.