El corpus

El corpus está dividido en dos:

1-Las versiones taquigráficas de las sesiones plenarias por el debate del aborto que tuvieron lugar en la Cámara de Diputados entre el de abril y el de mayo.

2- Las versiones taquigráficas del debate por la ley de legalización del aborto en la Cámara de diputados.

Preprocesamiento

En las dos partes del corpus, se usaron Expresiones Regulares para captar los patrones que dividían, en los documentos taquigráficos, las líneas que introducían el discurso de cada uno de los expositores/diputados.

Esa extracción se bajó a una tabla que indica el nombre del orador, su profesión (no en todos los casos) y su posición a favor o en contra de la ley del aborto. RegEx también fue la estrategia que se usó para conectar apellidos en las transcripciones y en las tablas donde se indicaba profesión y postura frente a la ley. En este caso, la limpieza fue, en muchos casos, hecha ‘a mano’.

Finalmente, se extrajeron símbolos que podrían perjudicar el procesamiento posterior. Se trabajó únicamente con los discursos que estaban marcados con las ‘etiquetas’ ‘a favor’ y ‘en contra’, es decir, se dejaron de lado las intervenciones de los presidentes/moderadores del debate.

Lematización

Con la lematización se buscó reducir la dimensionalidad del vocabulario sin afectar a las formas completas, para que, al final del proceso, el output fuera legible. Al menos, esa fue la idea inicial.

Primero, se utilizó un lematizador muy simple: consiste en una lista de aproximadamente quinientos mil pares ‘forma de superficie’ ‘lema’, a dos columnas, separados por una tabulación.

Esta forma de diccionario tiene sus problemas. Primero, el diccionario disponible no poseía tildes, por lo que se procedió a eliminar las tildes del corpus, eliminando la diferencia evidente entre palabras con y sin tilde. Segundo, con este tipo de diccionario, se debe confiar en que el sea lo suficientemente completo y consistente como para ser realmente útil. Tercero, una misma forma de superficie puede ser entrada para más de una forma lema del diccionario. Pongamos el caso de la palabra ‘esta’, al eliminar la tildación, se convertía en la forma de superficie de dos lemas diferentes: la forma verbal ‘está’ para el lema ‘estar’ y el pronombre ‘esta’, para el lema ’ésta’; pero, al momento de lematizar de forma automática, la decisión de a qué forma llevarlo debía hacerse de ante mano. Se tomó, para todos los casos como este, la primera forma lema. Para el caso de la palabra ‘estado’, con importancia propia dentro del corpus, que era tomada como forma verbal y llevada al lema ‘estar’, se evitó la lematización (y consigo, la de las formas verbales reales. Conclusión: se redujo la dimensionalidad a costa de obtener un vocabulario que no fuera fiel al corpus original.

Sin embargo, no se abandonó la intención perseguida con el lematizador, y se pocedió al uso de otro que, con el uso de un PoS Tagger de Stanford provisto por la lirería NLTK, entrena de manera de decidir qué etiqueta sintáctica darle a cada palabra del corpus y, en base a esa etiqueta, aplica el diccionario de lemas, consiguiendo un vocabulario reducido y fiel (o al menos, mucho más fiel) al original.

Extracción de tópicos con LDA

La hipótesis perseguida en la extracción de tópicos era la idea de que cada postura frente al aborto (‘a favor’ y ‘en contra’) utilizarían distintos tópicos para apoyar su postura y contradecir la postura opuesta. Estos tópicos, se espera, incluirán distintas maneras de tratar el tema desde lo social, económico, político, científico, médico, etc.

Para explorar esta hipótesis, se aplicó LDA en el total de los textos y en los textos separados por la etiqueta ‘posición’. De paso, se aplicó en los corpus según lematización para ver qué diferencias aparecían en el uso de cada uno.

Una estrategia en la exploración de tópicos fue la de explorar la relación de los tópicos con ciertas palabras claves del debate. Las palabras fueron extraídas de un glosario del aborto. Los términos que se eligieron para explorar los resultados fueron:

Aborto

Vida

Muerte

Derecho