Resumen TFM: Análisis de datos de proximidad para exploración y clasificación de textos

El objetivo de este trabajo es mostrar la aplicación del análisis de datos de proximidad al análisis estadístico de textos.

El trabajo se encuentra dividido en 4 capítulos. En el primer capítulo se definen los conceptos mas importantes en el campo de la estadística textual, se realiza una breve revisión de su desarrollo histórico, se muestran sus principales aplicaciones y se describe una etapa propia de la estadística de variables textuales: el necesario tratamiento previo de estos datos para obtener un conjunto estructurado sobre el que poder aplicar las correspondientes técnicas estadísticas, introduciendo los diversos tratamientos aplicables para codificar el corpus lingüístico en tablas numéricas.

En el capítulo 2 se describen algunas técnicas estadísticas exploratorias para datos multivariantes:

- El análisis de componentes principales, técnica de aplicación muy generalizada por su potencia para abordar la reducción de la dimensionalidad y permitir la detección de variables latentes.
- El escalado multidimensional, técnica de aplicación específica sobre datos de proximidad que permite abordar el análisis desde el punto de vista de la similaridad.
- El análisis de correspondencias, técnica históricamente ligada al análisis estadístico de textos.

En la sección dedicada al escalado multidimensional se incluye la descripción del concepto de medida de proximidad para datos multivariantes, se definen los conceptos de distancia y similaridad, se introducen algunas de las métricas mas importantes y se describe la medida de similaridad mas utilizada para comparar textos: el coseno del ángulo entre dos vectores.

El tercer capítulo se dedica al problema de la clasificación mediante la metodología del análisis discriminante. Comienza con una breve descripción del problema de clasificación o discriminación y contiene una parte de desarrollo teórico centrada en el análisis discriminante clásico desarrollado por Fisher.

Por último, en el cuarto capítulo se muestra la aplicación de los conceptos y técnicas estadísticas anteriores sobre un corpus lingüístico. Para ello se realiza, en primer lugar, un análisis exploratorio de la variable textual objeto de estudio, analizando el vocabulario empleado en el corpus desde el punto de vista de la frecuencia relativa por documentos de las distintas palabras, de las palabras que mejor caracterizan los textos pertenecientes a cada uno de los autores y de la comparación en las frecuencias de uso de las palabras entre los mismos. Además, se aplica la técnica de *unfolding* para obtener una representación conjunta de los autores y las palabras mas representativas que permita interpretar su relación a través de las distancias en el espacio de representación conjunto. En una segunda etapa se aborda una metodología para resolver el problema de clasificación asociado a la identificación de los autores de cada uno de los textos mediante el análisis de las proximidades entre ellos. Esta tarea se realiza utilizando la distancia del coseno como medida de la similaridad entre dos textos y aplicando técnicas de escalado multidimensional con carácter previo a obtener un modelo de clasificación mediante análisis discriminante lineal.

Los resultados obtenidos muestran como la matriz de documentos-palabras asociada al corpus lingüístico no es apropiada para la aplicación del análisis discriminante, obteniendo un modelo que no mejora demasiado una clasificación por azar, mientras que la aplicación sobre esta matriz de la distancia del coseno entre documentos para posteriormente construir la matriz de disimilaridad asociada y aplicar MDS proporciona una configuración que, reduciendo considerablemente la dimensionalidad de la matriz, resulta, al aplicar análisis discriminante, en un buen modelo con un acierto en la identificación de autores cercano al 90 %.