Resumo do Traballo de Fin de Grao

A la hora de clasificar documentos de forma automática, es necesario adaptar su representación a un formato que los algoritmos de clasificación puedan aceptar como entrada. Actualmente, la forma más habitual de representar un documento se denomina *bag-of-words,* en el que cada documento pasa a ser un vector en el que cada elemento indica la frecuencia de aparición de un término en el mismo.

Sin embargo, esta representación conlleva serios problemas a la hora de tratar grandes cantidades de documentos debido a la gran dimensionalidad de los vectores resultantes.

Recientemente, los algoritmos de *Topicmodeling,* como los basados en LDA (Latent Dirichlet Allocation), se están empleando para descubrir de forma no supervisada los temas sobre los que tratan los textos.

El trabajo principal consistirá en aplicar los modelos LDA para reducir el número de características necesarias para representar un documento y así mejorar la eficiencia de los algoritmos de clasificación. Se realizará una implementación en el software de minería de datos Weka para su posterior evaluación.

En concreto, las tareas a realizar serán las siguientes:

1. Desarrollo de un modelo basado en LDA para representar documentos.
2. Implementación del modelo en la herramienta Weka.
3. Evaluación del modelo y comparación con otros algoritmos de reducción y transformación de características empleando corpus de datos biomédicos.