Resumo do Traballo de Fin de Grao

(Describa brevemente o traballo a desenvolver, xustificando o interese do mesmo, e indicando obxectivos, descrición técnica, proceso de desenvolvemento, e medios empregados. Engada tantas liñas como sexa necesario)

El manejo de grandes cantidades de datos para llevar a cabo tareas de aprendizaje automático (Machine Learning) requiere de un pre-procesamiento de los mismos para poder ser empleados. El preprocesamiento incluye habitualmente tareas de reducción de la dimensionalidad, es decir, un proceso de transformación del espacio de variables originales a un espacio de dimensión menor.

Cuando se dispone de unos datos con gran cantidad de atributos e instancias, resulta necesario reducir dichos datos a un número menor de variables o de casos perdiendo la menor cantidad de información posible. De lo contrario, el tiempo de computación puede ser demasiado elevado además de poder estar induciendo ruido al sistema. Además, un alto número de características conjunto de genera sobre-entrenamiento en la etapa de clasificación.

Por tales razones, es aconsejable reducir la dimensión de los datos, mientras la estructura original de los mismos se mantenga casi intacta. Así, debe mantenerse la dimensión del espacio de características tan pequeña como sea posible, en consideración a la precisión en la clasificación automática.

Teniendo en cuenta lo anteriormente mencionado, el objetivo principal de este proyecto es reducir la dimensionalidad de un conjunto de datos que representan conceptos sobre los que versan ciertos textos con el objetivo de determinar si son molestos (spam) o no para un usuario determinado. Para ello, se implementará un algoritmo "ad hoc" en Java que, empleando diccionarios ontológicos, será capaz de realizar tareas de generalización semántica sobre los atributos (hiperonimia) para agrupar representar un conjunto de atributos similares en uno sólo (por ejemplo, agrupar los atributos "perro", "gato" y "hámster" en uno nuevo denominado "mascota"). Usando esta idea, el algoritmo será capaz de mantener prácticamente la misma información que los datos originales.

Una vez obtenido el nuevo conjunto de características con menor dimensión, se comprobará con un modelo de aprendizaje en Weka si mejora o mantiene su rendimiento con respecto al uso de todas las variables originales.

El desarrollo de este proyecto se hará empleando la metodología Scrum.

Para el desarrollo se dará prioridad al uso de herramientas de código abierto (Maven, Git, etc,…). Concretamente, además de Git como repositorio de código y Maven para la gestión del proyecto, se empleará el IDE NetBeans. Para el desarrollo se empleará un equipo MacBook Pro (2017) de 13,3 pulgadas con procesador Core i5 2.3Ghz, Intel Iris Plus Graphics 640, 8 GB de RAM y 128 GB SSD.