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)

En las actividades de I+D, y de forma muy especial en aquellas relacionadas con la explotación de grandes volúmenes de datos (Big Data), resulta necesario contar con herramientas para gestionar, archivar y procesar datos en formato original para la realización de experimentos. Por otro lado, la construcción de conjuntos de prueba (test datasets) a partir de los datos almacenados en su formato original (es decir, la realización de las operaciones de preprocesamiento) es una tarea que debería ser facilitada. Por tanto, además de las gestiones de los datos archivados, para garantizar la reproducibilidad de los experimentos es necesario publicar y mantener los conjuntos de prueba (procesados y en formato original) que se han empleado en la experimentación original. La creación de este tipo de herramientas está cobrando mayor importancia en la actualidad debido al auge de la explotación de grandes volúmenes de datos.

Teniendo en cuenta lo anteriormente mencionado, se plantea llevar a cabo el desarrollo de una aplicación web, denominada BDRep que permite el almacenamiento, intercambio y creación de nuevos conjuntos de datos en diferentes formatos. Concretamente, se focalizará la herramienta en permitir la gestión de formatos textuales (sms, eml, warc, tweet, youtube, ...) aunque se valorará la posibilidad de desarrollar de forma genérica para la gestión de otros tipos de datos.

La realización de las tareas de preprocesamiento se realizará con la herramienta *BDP4J (Big Data Preprocessing for Java)* que se está desarrollando en la actualidad en el marco de desarrollo de otro trabajo fin de grado. En este sentido, el desarrollo de esta plataforma debe estar coordinado con el desarrollo de la herramienta BDP4J. BDP4J es una aplicación de escritorio multiplataforma y un framework, que proporciona soporte específico para la realización de actividades de preprocesamiento con distintos conjuntos de Big Data. BDRep permitirá descargar BDP4J para definir las actividades de preprocesamiento que se requieren para un determinado experimento. Dicha definición se podrá incorporar a un fichero XML que se empleará en la plataforma BDRep para realizar las actividades de preprocesamiento sobre los datos seleccionados para generar un conjunto de datos de prueba.

Por otro lado, la plataforma debe permitir el registro y validación de usuarios para que puedan identificarse y acceder a las funcionalidades que aporta BDRep. Solo una vez registrados podrán subir conjuntos de datos a la plataforma y combinarlos para generar conjuntos de prueba. La herramienta debe comprobar que cada conjunto de datos subido tiene el formato adecuado para cada tipo. Una vez BDP4J cree un conjunto de prueba con una configuración proporcionada por el usuario, se almacenará el resultado y si el usuario lo desea, se podrá hacer público. Cabe mencionar la necesidad de la existencia de un usuario administrador que tenga control sobre los usuarios y su actividad.

El desarrollo de este proyecto se hará empleando la metodología del Proceso Unificado.

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 portátil y los datos se almacenarán en un servidor del grupo SING.