## 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)

La explotación de grandes volúmenes de datos (Big Data) y su uso para la extracción de información y conocimiento o para la implementación de sistemas inteligentes, ha pasado a formar parte de las herramientas básicas de trabajo de empresas consultoras e instituciones de investigación. La explotación con éxito de grandes volúmenes de datos supone la necesidad de superar distintas dificultades relacionadas con: (*i*) el volumen, (*ii*) la velocidad en la recepción de los datos y (*iii*) la variedad de la información manejada. Para soportar la variedad de información manejada (*iii*) se deberán implementar mecanismos de preprocesamiento que ajusten la información a un formato único. A mayores, para resolver las problemáticas del volumen (*i*) y la velocidad de la información (*ii*), se debe contar con equipos de gran capacidad computacional que puedan ejecutar las tareas de preprocesamiento de forma que se maximice la cantidad posible de información por unidad de tiempo. Por tanto, el preprocesamiento de la información resulta un elemento clave para sacar partido de la explotación de grandes volúmenes de datos, y debe ser realizado de forma que se facilite la definición de las operaciones de preprocesamiento que se aplicarán sobre los datos, se minimicen los errores cometidos en la definición del proceso y se maximice la cantidad de información procesada por unidad de tiempo.

Dadas las características mencionadas en el párrafo anterior, este trabajo pretende abordar la construcción de un framework genérico y basado en la definición de una tubería (pipeline) de trabajos de preprocesamiento. El framework será implementado en Java y permitirá la definición de distintas tareas a aplicar sobre los datos así como de pipeline de trabajo entendido como la secuencia de tareas necesarias para completar el preprocesamiento de datos recibidos de forma heterogénea. El trabajo desarrollado se empleará como base para el preprocesamiento de la información manejada en el proyecto RETOS TIN2017-84658-C2-1-R "Integración de conocimiento semántico para el filtrado de spam basado en contenido" en el cual se pretende la extracción de información de diversas fuentes heterogéneas incluyendo redes sociales, mensajes de correo electrónico, sitios web diversos, etc.

El desarrollo de la herramienta partirá del estudio e identificación de los sistemas de preprocesamiento de información incorporados en herramientas de Big Data y Machine Learning incluyendo herramientas como Mallet (http://mallet.cs.umass.edu), TinkerPop (https://github.com/tinkerpop/pipes) ó AIBench (http://www.aibench.org). Se intentará proponer una solución que permita realizar el preprocesamiento de la información de una forma sencilla y resolviendo las problemáticas de las herramientas estudiadas.

El proceso de desarrollo estará guiado por la metodología SCRUM.

Para desarrollar el proyecto, se empleará el entorno de desarrollo IntelliJ, la herramienta de diagramación Visual Paradigm y el hardware empleado para el desarrollo constará de un ordenador portátil DELL XPS15 9560 (i7-7700HQ, 16GB RAM, 512GB SSD, GTX 1050 4GB).