Resumo do Traballo de Fin de Grao

El aumento en la digitalización en el mundo empresarial unido a hábitos y comportamientos de la sociedad en dispositivos conectados a internet, sirven para entender que actualmente tanto empresas como organismos públicos tienen a su disposición ingentes cantidades de datos. Estamos en los comienzos de la era del Big Data. Una de las caracterizaciones del Big Data son las llamadas 5V's, Volumen, Velocidad, Variedad, Veracidad y Valor. En este trabajo nos proponemos avanzar en la V de Veracidad, en el sentido de unificar formatos de la información extraída de diversos recursos de internet (correo-e, SMS, twitter, sitios web, ...) con el objetivo de asegurar la veracidad de los mismos. Entiéndase, por ejemplo, en correo-e la clasificación en correo spam y no spam.

El lenguaje de programación R permite disponer de extensas herramientas de preprocesamiento de datos como de clasificación. R además es un lenguaje funcional, es decir, el uso de paréntesis es muy habitual en el código, haciendo que su lectura y comprensión sea difícil. Matemáticamente semejante a la composición de funciones.

Para el desarrollo del código propondremos soluciones basadas en pipes, tanto para el procesado de los datos como para su análisis.

Objetivos concretos:

- Definir pipes con el objetivo de manipulación de cadenas de texto para transformar la información extraída de diferentes fuentes en palabras clave (tokens).
- Definir clases R6 para el tratamiento unificado de los diferentes tipos de datos.
- Extender los clasificadores estándares en Data Mining y Machine Learning de las clases S3 a R6. Usando inicialmente los procedimientos más populares (Naive Bayes, Logistic Regression y Random Forest entre otros) establecer estrategias para obtener diferentes sistemas de clasificación.
- Analizar los resultados de la clasificación, principalmente a través de la matriz de confusión.
- Determinar la mejor estrategia en el proceso de clasificación para el ejemplo propuesto.

Para desarrollar el proyecto, se empleará el entorno de programación RStudio y el hardware empleado en el desarrollo constará de un ordenador portátil.

Se utilizará la metodología SCRUM en el desarrollo del proyecto.