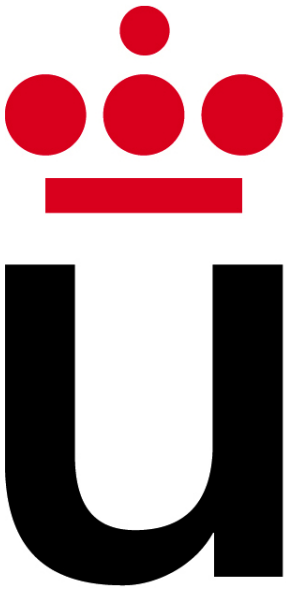
Foro de preguntas

Análisis de Big Data

Limpieza de datos

José Ignacio Escribano



Móstoles, 19 de marzo de 2016

Índice

1. Artículo

http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hu
html

2. Preguntas

2.1. ¿Cuáles son algunos de los problemas que impone la limpieza de datos no estructurados, tales como datos textuales?

Uno de los problemas es el incremento de las fuentes de información provenientes de sensores, documentos o bases de datos, ya que normalmente cada una de ellas tiene un formato distinto, teniendo que combinar esas fuentes de información en un solo formato. Este proceso puede ser bastante costoso debido a las inconsistencias que puede haber entre las fuentes de información y la gran cantidad de fuentes necesarias en las aplicaciones de hoy día.

Otro problema es la ambigüedad del lenguaje humano. Por ejemplo, una persona humana puede distinguir entre distintos sinónimos, pero un algoritmo debe ser programado para realizar esa tarea. Otro ejemplo claro son los dobles sentido que pueda tener un comentario: un humano es capaz de reconocerlo, pero un algoritmo seguramente no sea capaz, haciendo que los datos obtenidos por estos algoritmos no sean válidos para realizar una aplicación.

2.2. ¿Cuál es la principal actividad de negocio de la empresa Trifacta?¿Cómo crees que puede impactar este tipo de herramientas en la labor del científico de datos?

Trifacta ha desarrollado una herramienta para los profesionales de los datos. Su herramienta emplea distintos algoritmos de machine learning para encontrar, presentar y sugerir tipos de datos que puedan ser útiles en las tareas que se estén realizando.

Este tipo de herramientas pueden ser útiles para reducir en gran medida el tiempo que se emplea (entre el 50 % y el 80 % del tiempo del científico de datos) en obtener, preprocesar y agregar las distintas fuentes de información, empleando dicho tiempo en mejorar los algoritmos de la aplicación que se esté desarrollando en ese momento, y mejorar los servicios hacia los usuarios, pudiendo aumentando los beneficios de la empresa.