Propuestas de Trabajo Final del Diploma de Especialización en Big Data

Autoritas Consulting, SA

Análisis Morfo-Sintáctico en entorno Big Data

Muchas de las tareas de minería de textos requieren de pasos previos como el análisis morfo-sintáctico. Existen herramientas abiertas que efectúan esta tarea de manera eficiente en cuanto a precisión pero muy ineficiente en cuanto a tiempo de cómputo, lo que las hace inviables para su aplicación en entornos Big Data.

El objetivo del proyecto es definir y desplegar una arquitectura escalable para el análisis morfo-sintáctico en tiempo real de grandes volúmenes de texto. Un dato indicativo del volumen a procesar es del orden de 100 documentos cortos (tweets) por segundo por un volumen aproximado de 400 proyectos diferentes (100x400).

La tecnología de procesamiento morfo-sintáctico utilizada actualmente es Freeling (http://nlp.lsi.upc.edu/freeling/), en procesamiento batch sobre servidores Linux en Amazon AWS.

Se valorarán las siguientes tres fases (¿se puede convertir en un proyecto en equipo con tres tesinas diferentes?):

- El estudio de viabilidad de la tecnología utilizada y la comparativa con tecnologías equivalentes
- La propuesta de arquitectura escalable
- La implementación y puesta en funcionamiento de dicha arquitectura

Propuestas de Trabajo Final del Diploma de Especialización en Big Data

Autoritas Consulting, SA

Inteligencia Colectiva en Social Media

Autoritas dispone de un datawarehouse con documentos de más de 400 proyectos diferentes recuperados desde hace más de 5 años (cientos de millones de documentos). Los documentos son de distintos tipos: noticias, posts, tweets, vídeos… Cada documento ha sido generado por lo que en la terminología del sector se denomina influenciador. Un influenciador es el periódico El País. El País en la web escribe noticias de prensa bajo la dirección elpais.com y cuando escribe en Twitter lo hace bajo @el_pais. Otro influenciador sería Francisco Rangel que en la web escribe bajo el blog kicorangel.com y en Twitter lo hace bajo @kicorangel. Además, un influenciador menciona a otros influenciadores en sus contenidos, por ejemplo, @kicorangel puede comentar una noticia de elpais.com.

El objetivo del proyecto es por lo tanto generar una base de datos de relaciones de quién es quién y quién menciona a quién. Para ello se recorrerán los documentos almacenados en el datawarehouse y se construirá un grafo con los documentos como nodos de tipo documento y el autor del documento y los influenciadores mencionados como nodos de tipo influenciador, con la relación “autor” del influenciador al documento creado y la relación “menciona” del documento al influenciador mencionado.

Se valorará:

- Diseño de la arquitectura ETL. Se debe procesar un volumen considerable de información almacenada en tablas planas para extraer la información necesaria para construir el gráfico. Una ETL secuencial no es viable por el elevado tiempo que requeriría. Una ETL en paralelo podría colapsar las bases de datos de origen.
- Desambiguación de entidades. En ocasiones el influenciador está claro porque es un link a una noticia o una cuenta de Twitter marcada con @. En otras ocasiones no está tan claro, por ejemplo, cuando se indica el nombre de una fuente pero no su cuenta (Fuente: Blog de Francisco Rangel)
- Implementación del sistema

La tecnología a utilizar es Neo4j.

Propuestas de Trabajo Final del Diploma de Especialización en Big Data

Autoritas Consulting, SA

Intérprete Cypher y visualización avanzada

La incorporación de la tecnología Neo4j descrita en el Trabajo 3 nos permite la construcción de una macro-entidad que albergue inteligencia colectiva. Su explotación se puede realizar desde entornos como el que proporciona el propio Neo4j web, pero en este trabajo proponemos la integración de un intérprete dentro de la propia herramienta de Autoritas.

Autoritas ha desarrollado Cosmos en entorno Liferay. Liferay permite el desarrollo de porciones de código denominadas portlets. Una pantalla de la aplicación está compuesta por un conjunto de portlets que interactúan entre sí. Se propone diseñar un conjunto de portlets que permitan:

- La incorporación de sentencias Cypher para consulta de datos. Estas sentencias serán un subconjunto de Cypher limitado a búsqueda de información (no se permite crear ni eliminar). Este portlet consultará la API del servidor Neo4j y obtendrá los resultados
- La visualización de resultados de la consulta en dos opciones:
    - Fichas de resultados con tecnología PrimeFaces (e.g. http://www.primefaces.org/showcase/ui/data/datascroller/loader.xhtml)
    - Gráficas con tecnología D3js (e.g. http://d3js.org/) o Visjs (e.g. http://visjs.org/)

Se valorará:

- La capacidad de ejecutar sentencias Cypher y obtener resultados
- La visualización en fichas de resultados
- La visualización creativa con tecnología d3js

Las tecnologías a utilizar son:

- Liferay
- PrimerFaces
- d3js
- Neo4j

Propuestas de Trabajo Final del Diploma de Especialización en Big Data

Autoritas Consulting, SA - Universitat Politècnica de València

Propuesta particular del alumno

Además de los proyectos propuestos, se posibilita al alumno a proponer su propio tema y llevar una coordinación conjunta Autoritas - UPV. Algunos temas de interés pueden ser:

- Author Profiling: Identificación de edad, género o variedad regional del idioma
- Plagiarism Detection
- Documentación en entornos Big Data
- ...