TFM (Trabajo Fin de Máster) del TECI

Curso académico: 2014/2015

Título: Data Science en el ámbito de los Datos Culturales: Desambiguación y reconciliación de autoridades contra bases de datos de referencia

Tipo (marca una casilla): Académico ☐ **Profesional** ■

Institución: Biblioteca de la Fundación Juan March
Persona:
Ponente (si procede):

Observación:
Al rellenar los siguientes puntos hay que considerar que la carga de trabajo no debe superar las 300 horas para el estudiante

Problema a tratar:

La Fundación Juan March cuenta con un corpus de datos heterogéneo producido y gestionado por sus departamentos de Exposiciones, Música, Conferencias y Biblioteca. Este corpus incluye videos, audios, fotografías, publicaciones y una gran variedad de datos textuales. En los últimos años desde el área de Data Science de la Biblioteca se está tratando de integrar este corpus y enlazarlo con bases de datos de referencia en entornos de la web semántica.

Uno de los retos a los que nos enfrentamos es la construcción de un tesauro e índice de personajes único para toda la Fundación. Cada fuente de datos almacena información sobre personas e instituciones que se necesita normalizar para conseguir la integración y transversalidad en los contenidos de la organización. No solo eso, además existen bases de datos de referencia sobre autoridades como VIAF (http://viaf.org/) que sirven de instrumentos imprescindibles de cara a normalizar e integrar nuestra información de personas y autoridades con las del resto de instituciones culturales y científicas.

A través de este proyecto se pretende desarrollar estrategias de desambiguación y reconciliación de autoridades usando métodos de aprendizaje de máquinas replicando algunas de las funcionalidades que pueden aportar herramientas como Open Refine (http://openrefine.org/).

Objetivos:

El objetivo fundamental es que el alumno desarrolle su trabajo de fin de master guiado por el equipo de Biblioteca de la Fundación permitiéndole adquirir unos conocimientos y experiencia en el trabajo y explotación de datos que le pueda ser útil en otras empresas a la vez que nos ayude con el reto de desambiguar y reconciliar datos de autoridades de la Fundación.

- Trabajar con datos de autoridades de la Fundación y de las bases de datos de autoridades de referencia que permita adquirir experiencia en el acceso y tratamiento de datos del sector cultural.
- Utilizar métodos de aprendizaje de máquinas vistos durante el Máster (redes neuronales o k nearest neighbors) para reproducir las funcionalidades de *clustering* de la herramienta Open Refine para la normalización de nombres de personas y entidades.
- Estudiar la forma de replicar los algoritmos de Open Refine para la reconciliación con bases de datos de referencia y así ganar experiencia en el desarrollo de algoritmos contra servicios web y en entornos de web semántica.
- Utilizar herramientas como R o Phyton para todo lo anterior con el fin de adquirir una experiencia práctica en estas herramientas fundamentales en equipos de Data Science.