Semester project for SIN/SSC/DS Master students

Automatically building a hierarchical taxonomy of tags

Martin Rajman
Martin.Rajman@epfl.ch

Project description:

The goal of this project is to automatically build a hierarchical (tree-like) tag taxonomy from a corpus of 160'000 companies described by a set of 300'000 tags.

Some examples of the available tags are:
- "develop advanced biodegradable solutions"
- "20 Acquisition Channels"
- "touch interfaces"
- "ubiquitous computing"
- "Consumer Electronics"
- "Computer Hardware"

and an example of a company description using the available tags is:
- the company "Sinequa" is described by the following tags:
  - Big Data Analytics
  - Big Data search and Analytics
  - Business Intelligence
  - Cognitive Computing
  - Content Analytics
  - Enterprise Search
  - ...

A possible approach is to find an adequate (typically information theoretic) measure to efficiently prune the word lattice that can be derived from the available collection of tags. An alternative approach is to adapt the method described in the "Topic tree: Building a hierarchy" section of the article accessible at:
https://blog.feedly.com/data-science-behind-recommendations-in-feedly/

Proposed for:
1 Master student in Computer science, Communication systems or Data Science