15/05/2017

# Topic Modeling & Sentiment Analysis

## Description:

Uncovering topics from texts has become an important task for many applications. Topic modelling provides methods to organize large collections of textual information. It can be seen as the unsupervised version of text classification. In order to accomplish this task, conventional models implicitly capture document-level word co-occurrence patterns to reveal topics.

Sentiment analysis is the computational study of people's appraisals and emotions toward entities, events and their attributes. The goal is to determine whether an opinionated document (e.g. reviews) or sentence expresses a positive or negative opinion. Opinions are important because whenever someone wants to make a decision, he would like to hear other's opinions [1].

While determining the opinion of short texts is easily handled by sentiment analysis, topic modeling methods suffer from severe data sparsity. Some approaches have been developed for this kind of texts [2, 3]. These latter have been applied to very short texts such as SMS, chats, Twitter, Facebook posts, etc.

The idea of this project would be first to implement and compare several methods of topic modeling for short texts such as reviews. As a second step, we would run these on the *Yelp* dataset which contains restaurant reviews[1]. Finally, we would like to determine the sentiments of these and aggregate per former found topics.

## Proposed plan:

1. Implement the method in [2], compare it with BTM [3] and LDA [4][2] on a known dataset as well as *Yelp* dataset.
2. Infer the sentiments of the sentences using an existing classifier.
3. Aggregate the them by topics.

**Prerequisites:** Knowledge about Machine Learning.

**Supervisor**: Diego Antognini (diego.antognini@epfl.ch)

## References:

[1] https://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf
[2] http://www.aclweb.org/anthology/W15-1526
[3] http://www.bigdatalab.ac.cn/~lanyanyan/papers/2013/WWW2013-yan.pdf
[4] http://www.cs.columbia.edu/~blei/papers/BleiNgJordan2003.pdf

[1] https://www.yelp.com/dataset_challenge/dataset
[2] Implementations of LDA and BTM already exist