LINKÖPING UNIVERSITY
Department of Computer and Information Science
Division of Statistics
Mattias Villani

2013-05-05
Text Mining
732A47

Text Mining Project - Some Suggested Project Directions

You are encouraged to find your own topic for the text mining project.
The following is an unstructured list of some possible directions for projects. I will add more projects here when they come to my mind.

- First of all, have a look at the corpus included in **NLTK** corpora. See if there is some data there that you might use for some interesting project. It is of course not OK to repeat an analysis from the book, you have to come up with your own application.
- **Topic classification**. Online newspaper place their articles under different fixed topics. For example The New York Times (nytimes.com) classifies their articles under topics such as *Technology*, *Science*, *Health*, *Sports*, *Arts*, *Style*, *Travel* etc. Fetch a sample of webpages from the newspaper, and build a classifier that is able to predict the articles' topics. This is just one example of topic classification, I am sure that you can think of many similar projects.
  Contact person: **Mattias Villani**.
- Data from **Twitter** or other similar microblogs/social networks (Facebook, Google+, LinkedIn etc etc) is a good source of text data, which is relatively easily accessible. You can also use other type of information from Twitter, such as information on who follows who, and who retweets who, and so on. There are already books written on how to mine text from social networks such as Twitter (see for example http://shop.oreilly.com/product/0636920010203.do).
  Contact person: **Mattias Villani**.
  Here are some suggestions for Twitter-based projects:
  - *Sentiment analysis*. Collect tweets with a given hashtag (for example #windows8). Train a classifier that predicts peoples' sentiments regarding the object of the hashtag. Note that these types of projects requires you to label the sentiment "by hand" for each of the tweets, which takes some time, but typically goes faster than you would think, once you get going.
  - *Predicting political ideology* on the basis of the text in a person's tweets. The training data could be tweets from active politicians, which is attrative since you would then have labelled data = you know the political affiliation of the people in the sample.
    Update: Lyam Dolk came up with this idea independently of me (bravo!), so consider this specific project already taken by Lyam. But there are clearly variations of this idea that someone else can try.
  - What characterizes a tweet that gets re-tweeted by many people? What kind of tweets generate many (new) followers?

- **Term extraction** from a monolingual or bilingual corpus.
  Possible data sources: http://www.statmt.org/europral,
  http://www.nactem.ac.uk/genia/genia-corpus/ or a corpus from Microsoft's online help system.
  Contact person: **Lars Ahrenberg**.