Tentative Title: Topic Modeling

Team Members: Kevin An, Martin Barus, Daniël Okret

Brief Overview of Topic and Motivation:
Topic modeling is a very technique that is very relevant in our age of information. Increasing amounts of information is digitized and stored through blogs, newspapers, scientific articles, novels, sound and social networks. Topic modeling offers an alternative to the way navigating information is currently being done. Instead of searching with keywords, topic modeling tries to label each document of a large and unstructured collection with particular themes.
The topic uses techniques covered in the class *ECE 532 - Theory and Applications of Pattern Recognition* and allows students to apply the theories learned in class.

Topic modeling assumes each text has been generated with some model. Therefore the challange is to write an algorithm that can uncover these *hidden structures* such as the topics, per-document distributions or per-document per-word topic assignments using the collection of texts as input. A successful algorithm should be able to accurately infer the topic of each particular document, this topic should best describe the collection of words used in each document. It will also give the most probable terms associated with each topic. This project will teach students about the strengths and weaknesses of this approach and give students hands on experience in developing a topic model.

Challenges might arise with the usage of statistics and probability in the application of topic modeling. It is also uncertain if a basic topic modeling algorithm can deliver an accurate result.

Core Concepts:
The basic concept is that we will create a topic-term matrix. Each row will correspond to certain topic and each column will correspond to certain term (word). We will extract all words from all the articles and assign topics to them and then create this matrix, which will tell if the word belongs to some topic or if it doesn’t. (1/0)

One way to tell the topic of a new article is just to count occurrences of words and count the sum for each topic and then select the topic with highest sum of corresponding words. We could also try to express each topic as a combination of latent features (hidden features) and each word as a combination of these features. To do so, we could use the SVD. Then we could predict the topic of each new article by choosing the topic with most resembling sum of words features.

The reason why we would use the SVD is, because there can only be a limited number of relevant latent features (those with highest sigma values). In other words: we could perform a PCA, which might result in better performance.

Related Papers, Datasets, or Resources:

Probabilistic Topic Models
https://www.cs.princeton.edu/~blei/papers/Blei2012.pdf

Transfer Learning for Image Classification with Sparse Prototype Representations
https://drive.google.com/file/d/0B6LVktSU-SsQOTRTYWE0QzJKTG8/view?usp=sharing

Machine Learning in Automated Text Categorization
https://drive.google.com/file/d/0B6LVktSU-SsQZWdhNHNlb3pUQk0/view?usp=sharing

Topic Spotting on News Articles with Topic Repository by Controlled Indexing
https://drive.google.com/file/d/0B6LVktSU-SsQNFZSdGpSX1o2Mzg/view?usp=sharing

METHOD AND SYSTEM TO PREDICT THE LIKELIHOOD OF TOPICS
https://drive.google.com/file/d/0B6LVktSU-SsQNVdMMjFDMHI2b00/view?usp=sharing

Evaluation Methods for Topic Models
https://drive.google.com/file/d/0B6LVktSU-SsQRzByZFlxMll1dnM/view?usp=sharing

A Neural Network Approach to Topic Spotting
https://drive.google.com/file/d/0B6LVktSU-SsQVmhVZVdsbHg2TTg/view?usp=sharing

Any large collection of documents can be used as a dataset. Examples include: scientific journals, news articles, and periodicals.