Assignment 3: Text mining and exploratory analysis

Samuel Rönnqvist (sronnqvi@abo.fi)

February 19, 2015

This exercise touches upon unsupervised learning methods for analysis of data without any known output variable. Rather, we are interested in uncovering interesting structures within the data, especially looking beyond single or pairs of variables. Exploratory study is suitable when we are confronted with unknown data, when we might not know very well what they contain or what questions to ask about them. The exercise is also an introduction to working with textual data, and performing some basic text mining tasks.

Deadline: The report is due by **March 2nd.**

This document describes the analysis you should perform and contains the questions you are to answer. Take care to understand what you are doing, and let the *report* that you write reflect it. Explain your reasoning concisely, and submit as a *PDF of 2-4 pages*. Your final edited *code file*, which you use to perform your analysis, shall also be uploaded separately but will not be graded. You may work individually or in groups of up to three; submit once per group through Moodle. The exercise awards up to 20 points out of a total of 50 points for all assignments, 50% of the score in each assignment is required to pass the course.

1. Working with text data

The text mining process relies heavily on data preparation that aims to structure data, and support subsequent modeling. Initial steps in basically any text mining application include splitting text into sentences and their words into *tokens* (words etc.). In general, text preprocessing steps perform *feature extraction*[1] on the data, in order to filter relevant information and present it in more structured or abstract forms to the learning algorithms that follow.

The text data used for this assignment is a selection of patent application abstracts from the US Patent and Trademark Office, all stemming from the subclass 705/35 defined as finance (e.g., banking,

[1] In machine learning a *feature* is the same as input variable. Feature extraction/selection is a process of limiting and normalizing a large number of variables to a set of variables that are better suited for modeling.

investment or credit). You will be introduced to some elementary methods for analyzing the topics of these documents, as well as their similarity structures, in order to better understand what is being discussed and how different documents/topics relate.

Text statistics

Once the text has been tokenized, by the code you have been provided, we can start calculating basic statistics on the resulting terms, to learn something about the text and the structure of the language. The term frequency distribution of a text, which describes the number of terms that occur with different frequencies, follows a Zipfian distribution[2].

[2] This distribution is governed by a power law, and can be used to describe many other types of natural data, such as number of connections per node in a social or other type of natural network (scale-free networks).

Often in text mining, text is analyzed using a *vector space representation*, which represent a document as a vector where each term has its own position that can hold some value (e.g., term count). This representation can be seen as each term representing a separate dimension in a high-dimensional space, and the term counts may represent the a points position in each dimension (or length of the base vector). Treated in this way, text data is very high-dimensional and it is sparse since each data point (document) usually occupies only a fraction of all dimensions in the dataset (the vocabulary).

Inspect the distribution for the text file specified in the code to answer the following questions.

Q1.a Assess either visually or numerically how many terms occur only once, and how often the most frequent term occurs.

Q1.b What kinds of words are most frequent in the document, and do they describe the content of the documents?

Q1.c Explain in your own words what it means that text data is sparse, and why it might lead to difficulties when measuring similarity of two documents based on word counts.

Treating words in a text as independent, i.e., without regard to word order, is referred to as the *bag-of-words* approach. Calculate bigram[3] statistics on the text as well, to incorporate some word order information, and compare it to the unigram (single term) statistics.

[3] N-grams are sequences of N consequitve terms observed in a text, for $N = 2$ they are called bigrams.

Q2 What might be the advantages and/or disadvantages of using bigrams, instead of unigrams, as features in measuring topical similarity of documents?

Identifying keywords

Using the *TF-IDF* (term frequency - inverse document frequency) method and the document-specific term frequencies, it is possible to

rank terms according to how representative they are of a document. The method favors terms that occur frequently in a document, but infrequently in documents in general, yielding a higher score for terms that are more specific to the document at hand, compared to all other documents in the analyzed *corpus* (body of text). The ranked terms can be used as keywords to represent the topic of a document.

The provided code reads and calculates TF-IDF scores for the 100 first documents of the corpus. Inspect the top 20 keywords for the 5 first documents in order to answer the following questions.

Q3.a Evaluate the quality of the keywords by comparing them to the text of the documents. Do you think the method works well? Explain.

Q3.b Concider whether the text preprocessing could be altered in some way, in order to improve the results. What would you propose and what effects do you expect?

2. *Document clustering*

Clustering is an unsupervised learning approach that aims to identify meaningful structure and groupings in multivariate data, based on similarities between data points.[4] Since we are working with text data, we will demonstrate clustering of documents as a means of exploring a corpus.

[4] Typically, measures of dissimilarity are actually used in clustering, such as Euclidean distance.

Document similarity

The basis for any clustering is the ability to measure similarity between data points, in this case documents. The TF-IDF scores function as weights in comparing documents to documents, using the vector space representation[5] and the *cosine similarity*[6] function that produces a similarity score between 0 and 1. The similarity score is influenced by the number of terms shared between documents, subject to the term weights reflecting their importance. You will use TF-IDF and the cosine function to calculate similarities between the documents in our corpus.

[5] More background reading on the vector space representation in relation to TF-IDF is available *here*.

[6] The cosine similarity is the cosine of the angle between two document vectors, defined by their terms (dimensions) and TF-IDF weighting (basis length), in the high-dimensional vector space. It is a better similarity measure than Euclidean distance for sparse data.

Hierarchical clustering

To explore the calculated set of document similarities, we will perform *hierarchical clustering* that recursively merges pair-wise closest data points or clusters. The result is a hierarchical structure representing relationships in the data, which can be visualized as a *dendrogram* (a tree-like structure). Distances between clusters (branch points) are represented by vertical distance.

Hierarchical clustering can use a number of different linkage methods that define how they select the clusters to merge. The code uses *Ward's method* by default, try a few other methods as well to compare the results (e.g. single and complete linkage).[7] We will cluster only the 30 first documents, in order to make the results easier to evaluate in more depth.

[7] The R manual page for the *hclust* function lists the alternative linkage methods. Type *?hclust* in R.

Q4 Make some general comments about the differences in global structure between the dendrograms using different linkage methods. Judging from the visualization, which clustering do you think best reflects the underlying structure of the data (in terms of topic similarities), or are the differences in fact negligible? [8]

[8] The question calls for a qualitative analysis, but clusterings can also be evaluated quantiatively by a range of measures, which share the general aim of measuring the degree of separation between clusters (inter-cluster distance) and concentration within clusters (intra-cluster distance).

Partitioning clustering

Alternatively, clustering can partition the data set into a predefined number of flat, non-overlapping clusters. Such clusters might be easier to work with, while they loose more detail on the underlying distance structures they seek to represent. In our case, we will derive a partitioning from our hierarchical clustering by specifying a desired number of partitions.[9]

[9] Partitioning clustering can also be performed directly, without hierarchical clustering, for instance, using the k-means method.

Q5 Familiarize yourself with the contents of the documents in order to interpret your clusters. What would be a good number of clusters for the text documents that would reflect some meaningful, natural grouping of their topics, and can you find some suitable description for each cluster? Explain your reasoning.

3. *Topic modeling*

Analyzing text with methods such as TF-IDF suffers from the problem of data sparsity. Therefore, various methods are commonly used to introduce more abstract representations of text than the raw word level, or the generalizations obtained by simpler lexical operations such as stemming. For instance, representations of the semantics of documents both serve to reduce dimensionality and directly support interpretation of themes in text. Using TF-IDF, we operate at the word and document levels, whereas topic modeling introduces a more abstract layer of topics, which then links to documents and words. We will focus on topic modeling with the popular Latent Dirichlet Allocation (LDA)[10] algorithm, presented in:

[10] Do not confuse with Linear Discriminant Analysis, also abbreviated LDA.

[1] David M. Blei. Probabilistic Topic Models. In *Communications of the ACM*. Vol. 55 No. 4. 2012.

Latent Dirichlet Allocation attempts to extract meaningful topics from a set of documents based on word co-occurrence in the documents, in a fully unsupervised manner. The number of topics is

given as a parameter and LDA estimates topic probabilities for each document (topic assignments, a document can discuss a mixture of multiple topics) as well as word probabilities for each topic (describing the meaning of a topic).

Q6 Run the code to perform topic modeling on all files in the corpus, and study the topics that emerge, while testing a few parameter settings (primarily number of topics). Are you able to find meaningful topics? Does the topic model provide any new insight into the corpus, compared to your previous analysis?

4. Exploratory analysis and data science

Exploratory analysis is often an important step in seeking to understand data relating to some specific problem, and it may involve both modeling (such as clustering) and visualization. The data mining process is highly iterative, where problems may be very open-ended, and increased understanding of the task or data may call for changes to previous steps in the process. The below article discusses what may be referred to as the *data science* approach to analyzing data, and should give you a broader view of the field and where it is heading. Read the article in order to answer the next questions in your own words.

[2] Vasant Dhar. Data Science and Prediction. In *Communications of the ACM*. Vol. 56 No. 12. 2013.

Q7.a What is the role of feature engineering/extraction when working with unstructured data? Give some examples and highlight possible challenges.

Q7.b The article focuses primarily on predictive modeling, as a general end goal in data science. First, elaborate on the role and benefit of exploratory analysis, in relation to that context in general. Second, concider how the role and benefit might be different in the particular case of text mining.