Synopsis - Text Mining Kierkegaard

Introduction

As the interest of text mining, and topic modelling in particular, grows within the humanities, it is my aim with this study to gain knowledge of how this method can be applied within my fields of study and interest, namely literature studies and cognitive semiotics. As the data available for mining increases rapidly and several methods offers quick and convenient gazes into enormous text corpora, I find it both interesting and important to reflect on advantages and disadvantages of this, too me, new approach to text analysis that topic modelling offers. Especially when dealing with aesthetic language. There is no doubt how powerful topic modelling can be in terms of searching for and retrieving information, but there is something slightly counterintuitive to me as a former literary scholar and a present semiotician, about trying to understand big amounts of literary texts and derive meaning from them through a computer programme. This is partly due to the classic hermeneutic methodological approach we as students are taught when analysing and interpreting literature, an partly the basic assumption from semiotics that 'meaning' happens when someone interprets something in a context[1].

My purpose is to, without specific questions directed at my corpus, to dive into an authorship and thereby get a deeper understanding of how topic modelling is to be used.

Corpus

The corpus used in my group was the collected works of the Danish theologian, philosopher and writer Søren Kierkegaard (1813-1855). He is known for writing in an often complex language, full of humour and irony. He is furthermore using many voices in his literature, by writing under different antonyms. Through these antonyms, he is arguing from different points of view, and this combined with the many different types of texts - from love letters to journals, notebooks and long philosophical dissertations - makes his writings comprehensive and pluralistic. Although Kierkegaard is not a fiction writer as such, topic modeling his writings will presumably imply some of the general problems one will encounter when dealing with aesthetic material, due to the before mentioned notions about his language.

[1] C.S. Peirces definition of a *sign* in Hoffmeyer, 1993

Methods and pipeline

We used the *gensim* module to perform probabilistic topic modelling on our text corpus. The script can be found in my GitHub repository[2]. The texts were already made available for us, scraped from the web archive sks.dk., and it consisted of 214 texts of different length.

The texts were already separated from the metadata and the corpus was collected in a folder as txt.files, which made the retrieval and initial importing of our corpus pretty straight forward. Then followed the lower-case folding and tokenization, which was set to include æ, ø and å to accommodate our Danish corpus. With a function, we then generated a stop word list from the corpus, to be able to remove the most frequent words before starting the topic modeling.

Changing the number of stop words, from e.g. 50 over 100 to 250 revealed the importance of a stop word list, and also the crucial role of the *researcher* using a tool as topic modeling. With a non-strict stop word list (50 words), the topics that later on revealed themselves in the model was barely distinguishable, only contained high frequency words like shown in the example below:

```
Topic 2
[u'fra', u'min', u'noget', u'hans', u'bliver', u'blive', u'dig',
u'alt', u'derfor', u'hvor', u'kun', u'menneske', u'vilde', u'dem',
u'mere']
-----
Topic 3
[u'kun', u'noget', u'havde', u'vilde', u'vi', u'vel', u'hun',
u'dig', u'hans', u'hvis', u'hvor', u'alt', u'fra', u'bliver',
u'derfor']
```

A stricter stop word list (100 words), did still not result in informative topics, but filtered out the words “menneske”, “verden” and “gud” (“human”, “world” and “god”) - words one might consider quite central to the writings of a Christian, existential philosopher. The internally generated stop word list can thus be interesting, because it can, like in this example, reveal the most frequent used nouns. These nouns can be quite meaningful to understand the corpus as a whole. Other frequent words could be interesting if one is investigating the rhetorical habits of the author, or the time the author was writing. Despite the centrality, I still chose to omit these words, because they pretty much sum up the (my) general knowledge of Kierkegaard, and I would like to go beyond that.

These considerations should make it clear how the identification of a threshold for the stop word filter is a somewhat subjective process. It also implies that the threshold may differ according to the questions the researcher is asking - as it is up to the researcher to decide when the topics displayed are indeed meaningful.

[2] https://github.com/idajuutilainen/idajuutilainen

After generating the stop word list, we trained our model. In this step, we were able to manipulate the numbers of topics that the model gave us. The last step was to print the topics and here decide how many words each topic was to consist of. These three steps; generating stop word list with different levels of strictness, training the model, choosing different amount of topics and number of words in each topics, were repeated several times, as we were searching for the values that combined would give us the most meaningful results.
At last, we picked out five texts form the full corpus and used a loop to identify the percentage of each topics in the different texts.

Results

We ended up choosing 250 stop words and 10 topics. As we never made it to the visualization part, the full model can be found in my GitHub repository. The topics shown below are topic 0, 5, 7. They are among the most present topics in the five selected texts.

Topic 0	troen, aand, betydning, ak, strax, evige, vise, kierkegaard, dit, enkelt, sandt, viser, bestandigt, elske, alvor
Topic 5	christendom, viser, tør, ei, nye, bestandigt, forstaaet, retning, tager, mennesker, glæde, ligesaa, dertil, desto, sagen
Topic 7	aand, enkelt, glæde, hvorfor, strax, høiere, to, grund, betydning, mener, grad, evige, alvor, elske, imidlertid

The model sample clearly shows some language and spelling habits from Kierkegaard and the time he wrote in, e.g. the spelling of *strax* or *høiere* or the exclamation *ak*. The words *dertil* and *desto,* I believe, is due to his own rhetorical habits. The model also gives a clear 'coarse' overview of the existential and religious themes that seems pervasive through the complete authorship, as all the topics are full of abstract concepts like *truth*, *love*, *meaning*, *eternal* and *reason*. The presence of his own name in topic 0 could be a reflection of how more pre-processing is needed.

Discussion

Even though the model quite clear shows something about language use and main themes, I still find it quite hard to interpret and give all the different topics distinct labels. A visualization of the results[3], would definitely make the model much easier to explore. The problems with interpreting is definitely

[3] E.g. like this one from Ted Underwood: http://pmla.site44.com/

partly due to my own inexperience with topic modelling, text mining and programming in general. It also might be due to the language of Kierkegaard, and maybe this first encounter with topic modeling should have been done with more success with a classical fiction writer.

But either way, because of the bag-of-word approach to texts, some crucial parts definitely get lost in the process. With no prior knowledge of Kierkegaard, the researcher would have no chance of guessing how he is using irony as an important tool for communication in his writings. This might be the primary problem when dealing with aesthetics; that the meaning of poetic language, e.g. metaphors, are extremely context driven. Thus, some parts of thematic structures in literature might remain hidden due to the tokenization. This question however, could be interesting to pursue, using e.g. different satirical novels[4].

A great possibility of topic modelling could also be to challenge established assumptions about certain literary periods. Because the 'algorithmic approach' to literature is not under influence from prior knowledge, as the hermeneutic approach inherently is, new patterns of very well investigated periods, as the Danish golden age, might be discovered. It is not only the amount of data that can be hard to process for a human, also conventionalized knowledge might stand in the way of new perspectives. A computer programme can get around that, because it's search for patterns is unbiased, unlike the human way of interpreting and understanding.

Literature and sources

- www.sks.dk
- Blei, D. M. Probabilistic topic models. 2012, Communications of the ACM, 55(4), 77{84.
- Hoffmeyer, Jesper. *En snegl på vejen.* 1993, Rosinante
- Jockers, Matthew. Secret recipe for topic modeling themes, 2012.
- Roland, Teddy. Topic modeling: What humanists actually do with it. Digital Humanities, Berkeley, 2014
- Underwood, Ted. Topic modeling made just simple enough. 2012

[4] This could be done on Holbergs' "Niels Klims underjordiske rejse" - a politicial satire aimed at the king and nobility but disquised as a fantasy-novel - could a topic model capture the critique of the ones in power?