November 2014

Master's thesis project proposal

Title: Sub-corpus topic modeling and Swedish litterature

Keywords: topic modeling; macroanalysis; distant reading; culturomics

Goal: the goal of the Master thesis will be to: i) use/process a large Swedish text collection ii) experiment and apply topic modeling and consequently sub-corpus topic modeling (according the description by Tangherlini & Leonard, 2013) iii) adapt or create a visual, web based environment to explore the results (this will be done in various ways, preferably as a) network graphs (Smith et al., 2014); se for instance figure 1 and integrated them in b) a web based exploratory environment, such as a dashboard; se figure 2)

Background: Topic modeling (TM) is a way of identifying patterns, or clusters of words or "topics" in a (large) text collections. TM is based on unsupervised learning algorithm(s) and the most common technique is Latent Dirichlet Allocation (LDA) by Blei et al. (2003). According to Wikipedia <http://en.wikipedia.org/wiki/Topic_model>: given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would *probably* be about 9 times more dog than cat words. A TM captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.

Problem description: using the material in the Swedish Prose Fiction 1800-1900 database (SPF) and the Swedish Literature Bank (Litteraturbanken) you should apply sub-corpus topic modeling using a well understood corpus of (literary texts), such as A. Strindberg, or other (the Bible), in order to identify passages related to these well understood pieces of work. Details about the methodology is explained in detail in the paper by Tangherlini & Leonard (2013).

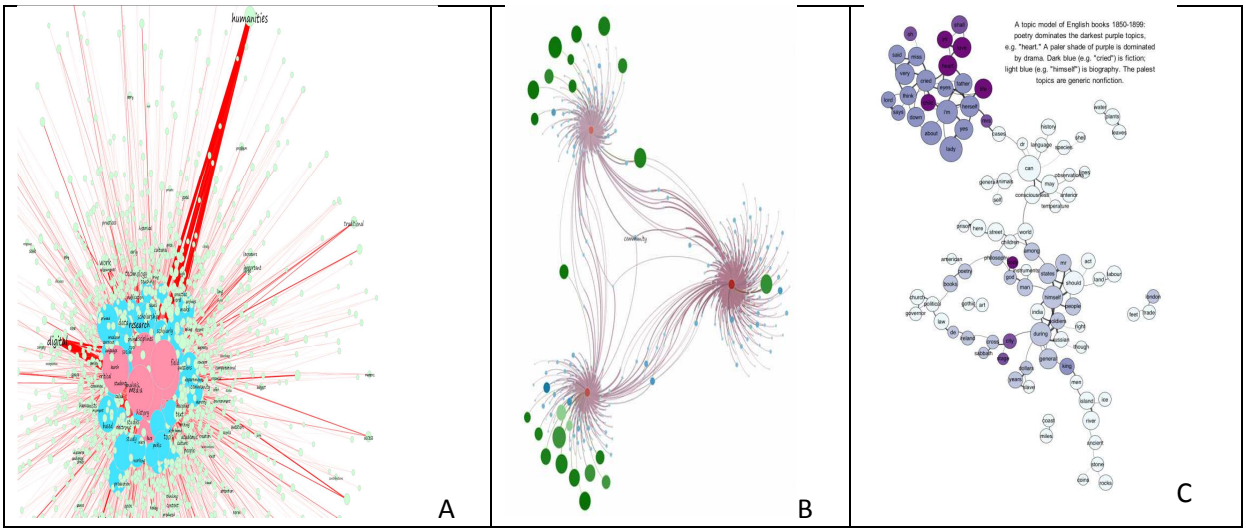

Figure 2. Graph representations of topics from Digital Humanities Specialist (Stanford) [A and B] and from Ted Underwood's blog [C]

Figure 2. Dashboard example for visualizing topics from Tangherlini & Leonard, 2013

Recommended skills: (a) programing skills in a script programming language, Python or Perl and/or javascript; (b) knowledge of web programing**;** (c) some familiarity with XML and particularly the Text Encoding Initiative (TEI) structure and (d) familiarity to execute shell programs and read the results (required for MALLET). **Note:** The use of the freely available implementation of topic modeling in MALLET is *recommended* for the assignment <http://mallet.cs.umass.edu/topics.php%20%20Topic%20Modeling>

References

- David M Blei, Andrew Y Ng and Michael I Jordan. 2003. Latent Dirichlet Allocation. Machine Learning Journal, 3:993–1022.
- Megan Brett. 2012. Topic Modeling: A Basic Introduction. J of Digital Humanities. Vol. 2:1
- Travis Brown. 2012. Telling New Stories about our Texts: Next Steps for Topic Modeling in the Humanities. Maryland Institute for Technology in the Humanities. University of Maryland, College Park. <http://mith.umd.edu/topicmodeling/wp-content/uploads/dh2012-slides.pdf>
- Alison Smith, Jason Chuang, Yuening Hu, Jordan Boyd-Graber, Leah Findlater. 2014. Concurrent Visualization of Relationships between Words and Topics in Topic Models. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 79–82. Baltimore, Maryland, USA
- Timothy R. Tangherlini and Peter Leonard. 2013. Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. Poetics. Volume 41:6. Pages 725–749.
- Ted Underwood: <http://tedunderwood.com/2012/11/11/visualizing-topic-models/>
- Litteraturbanken: <http://litteraturbanken.se/#!/start>
- Digital Humanities Specialist (Stanford) <https://dhs.stanford.edu/comprehending-the-digital-humanities/>
- SPF: <http://spf1800-1900.se/#!/start>

Supervisors: Dimitrios Kokkinakis (Department of Swedish) and Mats Malm (Department of Literature, History of Ideas, and Religion)