Garbage in, garbage out: Impact of sequence matching based text cleaning and phrase identification on unsupervised text mining

Arho Suominen[1], Hannes Toivanen[2]

[1]VTT Technical Research Centre of Finland,
Itäinen Pitkäkatu 4, Turku, P.O. Box 106, 20521 Turku, Finland,
arho.suominen@vtt.fi, Tel. +358 50 5050 354

[2]VTT Technical Research Centre of Finland,
P.O.Box 1000, 02044 Espoo, Finland,
Hannes.Toivanen@vtt.fi, Tel. +358 40 186 3882

In 2006, Daim et al. published the highly cited paper on forecasting emerging technologies with bibliometrics and patent analysis. In the paper, scientific publications and patent were used as a numerical input to for example system dynamic models or scenarios –elaborating on the current state and trend of technological development as a year-to-year indicator value. By forcing the indicator to the well-known growth models the analyst also had an indication of future development. This approach of quantifying instances of publication of patenting is to a significant extent valid in producing a “how much” indicator, but yields far less an indication on the “what” of technological development.

Looking to analyse the technologies more in-depth, studies have looked towards the automated analysis of semantic text to derive high-quality information on technologies. In practical tech mining, the researcher often has little background in the subject matter, making the use of unsupervised learning methods intriguing. Unsupervised learning methods do not require any training data and can be applied to a text mass directly. There has been a significant interest in for example using topic modeling, specifically Latent Dirichlet Allocation (LDA), in searching for hidden patterns text (e.g. Blei 2003). There is, however, some discussion on what is a practical approach to text pre-processing prior to running an analysis (e.g. Yau et al. 2014). It seems that, as with most methods, there is a clear need for pre-processing input data to avoid the well-known “garbage in, garbage out” effect.

In this article, we introduce a process of semantic analysis contributing to the “how much type” elementary indicators (Suominen, 2013). Our goal is to show the applicability of unsupervised learning methods in creating competitive technological intelligence. We show how we can synthesize textual information through an unsupervised process facilitating non-expert interpretation.

In practice, we show the impact of sequence matching based text cleaning, event extraction and bigram identification as a pre-processing for LDA. We created a Python –programming language based tool for matching tokens based on their similarity at different levels. This was done using a sequence matching algorithm implemented in the difflib library in Python, based on Ratcliff and Obershelps “gestalt pattern matching.” We also controlled the text for acronyms used by individual authors. The software tool also searched for bigrams, sequence of two adjacent elements, within the tokens, merging unique, tokens at different levels of co-occurrence.

We evaluated the use of a sequence matching based cleaning and bigrams in running a unsupervised learning method, LDA, on fuel cell related scientific publication abstracts (N=34900). We evaluated changes in token frequency against Zipf's law, perplexity of the topic model at a fixed number of topics and by a qualitative evaluation of the results.

Our results suggest that cleaning had a positive impact on the qualitative results. Seen in Figure 1 the frequency spectrum of words remained stable throughout the process following the Zips's law. However, while the token distribution remained stable, the cleaning had a clear impact on the terms in produced by the unsupervised learning process as key N-grams, such as Solid Oxide Fuel Cell, emerged to the forefront of several topics. Topics clearly pointed to key technology areas within fuel cells, making a practical synthesis for the non-expert.

Our results create an added layer, which can be overlaid on top of existing analysis such as presented by Daim et al. (2006). Work-in-progress focuses on incorporating automated labelling to the unsupervised topics. In addition, we are incorporating an event extraction algorithm to point to key events in each topic that could further extend the applicability of the text mining approach

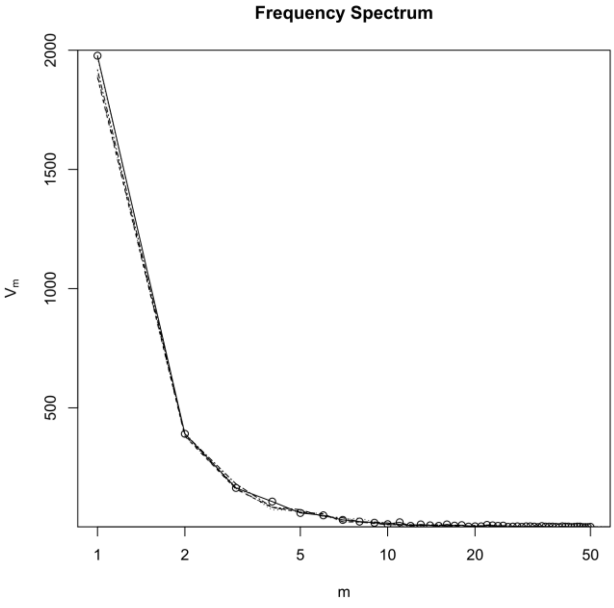

Figure 1 Frequency spectrum of words t different stages of cleaning. Picture contains four frequency spectrum lines, each with an increasing intensity of cleaning. As seen in the figure, the spectrum of word frequency remains stable throughout cleaning.

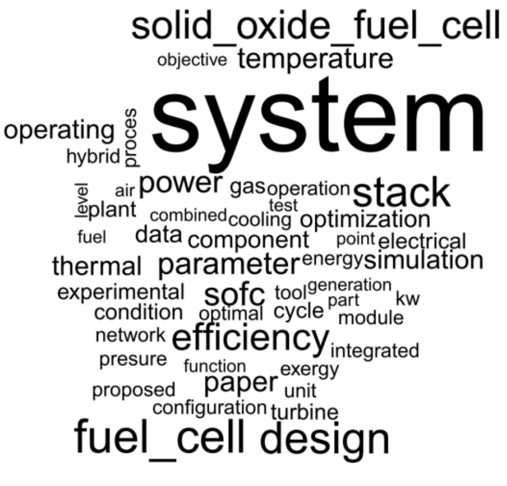

Figure 2 Example of top words in a topic after cleaning. N-grams are merged with a "_" character.