**Sequential Latent Dirichlet Allocation**

Description

LDA was constructed with the assumption of exchangeability, that means every document is theoretically exchangeable with any other. In reality, documents have in inherent order, e.g. by their timestamp. Also, some of the topic or word probability might change over time. These considerations lead to a version, where such a change is enabled. (Blei et al 2006) propose the Sequential Latent Dirichlet Allocation “LDASeq” to achieve this. Since “LDASeq” doesn’t overcome the exchangeability, which was embedded by design in LDA, the data is instead split on the time interval, e.g. one corpus for every year. In every of this by-year-corpi a separate LDA is conducted. These separate models are connected via some parameters, which are responsible for the distribution of topics and words (α and β). That means if in a previous chunk some topics and words occurred often, then in the next chunk these will also occur more often. The change of these parameters is done via Gaussian Noise.

The calculation of the posterior of this model seems to be intractable, so a variational method is used here as was in LDA. In variational methods the Kullback Liebler distance of a calculable formula to the posterior is to be minimized. That means that not the true likelihood will be calculated, but rather a lower bound. Since there are two variables, which are transported to every chunk, two new variational parameters are created here. The update of these parameters in time can be done via Kalman Filtering or Wavelet Regression.

Most relevant Parameters of training method in Gensim

* Time\_slice (list of int): number of documents in each time-slice
* num\_topics (int): number of topics
* initilalize (LdaModel): use a previously trained LDA Model to initialize LDAseq
* lda\_model (LDaModel): if in initialize “genism” was chosen, here the LDaModel can be entered
* passes (int): number of passes over the corpus for the initial LdaModel
* random\_state (int) – use a seed integer for reproducability

Shortcomings

* Data is split into diverse separate chunks and a Topic Model is built on every chunk. This is a problem if the dataset was small beforehand.
* Model enables no autoregression and no heteroscedasticity. Heteroscedasticity is the change in variance over time. Autoregression exists if observations depend on the time itself.
* There is just one set of topics for the whole collection of corpi. In reality this is unlikely that the set of topics stays the same for every separate corpus. Some topics might vanish through the years, while other come up.

Expected Results

In the example given in (Blei et al 2006) the algorithm is used on scientific data with a timespan of a century. The amount of 250 documents per year was chosen. The number of documents per year is comparable with our dataset. The critical difference is that in our dataset the distribution of words and topics change more dramatically than those in scientific papers. To adjust for the lower life expectancy of topics in our dataset, one should shorten the time intervals, in which the data is separated. This is not possible in our dataset, since then the amount of data per single corpus would be too small. So, all in all, no major improvement is expected over LDA.

Literature

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120).