Form: how will the subquestion will be answered

Subquestion 1: How accurately can LDA be applied to Twitter data?

How can topic modelling be applied on Twitter data? The first research question estimate how accurately topic modeling can be applied to Twitter data. Perplexity and coherence score will be used as evaluation on subs-set of the Burundi Twitter data-set.

In python the function: log_perplexity( ) from gensim.models.LdaMulticore will be used to calculate the perplexity. For the coherence the get_coherence( ) function from CoherenceModel will be used.

Sources:

- Blei D.M.; Ng A.Y.; Jordan M.I. Latent Dirichlet Allocation (2003)
- Stevens K.; Kegelmeyer P.; Andrzejewski D.; Buttler D. Exploring Topic Coherence over many models many topics.
- Röder M.; Both A.; Hinneburg A. Exploring the Space of Topic Coherence Measures

Subquestion 2: What is the optimal number of topics generated by LDA?

In the second sub-question, the influence the number of topics generated by LDA is measured. Also the perplexity and coherence score will used as evaluation. In many tutorials they suggest to use these methods to find the optimal number of topics. The same python functions will be used as in subquestion 1.

Subquestion 3: Can LDA distinguish between Twitter message using hash-tags as true labels

In the third sub-question the LDA output will be compared with the hash-tags which are labeled with each twitter message. Are the message with in common hash-tag assigned to the same topic group. For the this question the accuracy score will be used for evaluation. Since precision and recall necessarily depend on the notion of true classes. The hash-tags will be used as true classes. Topic models output has the same amount groups as the number of different hash-tags. A LDA topic will be assigned to a hash-tag based on the biggest number labeled hash-tags within LDA topic. From there the true positives, true negatives, false positives and false negatives can be calculated.

Source:

- A. Javed; Byung S.U.Sense-Level Semantic Clustering of Hashtags in Social Media.
- Javed A. A Hybrid Approach to Semantic Hash-tag Clustering in Social Media.