Preliminary experiments – Yue Wang Master Thesis

Topic: document similarity (for heterogeneous unstructured data)

Topic should covers:

- Investigating on different sources of document types as testing data (query document): **text and pdf document (same language vs. cross-language learning), newsfeed, twitter or microblogs**
- Evaluation on baseline model vs. trained WE (from stacked LSTM, stacked autoencoders)
- Evaluation on similarity distances
- Framework for evaluating multi-views document classification/distance

Immediate plans:

The first month is for preliminary experiments and literature review. During this first month (approx. after 2 weeks), please let me know and discuss what you have been thinking about your **own problem statement** related to the preliminary cases and/or our problem domain.

1. Implement a baseline experiments on document similarity distance, e.g. as described in the following sources: https://stanford.edu/~rjweiss/public_html/IRiSS2013/text2/notebooks/ and https://www.safaribooksonline.com/library/view/mining-the-social/9781449368180/ch04.html
2. Literatures on this subject are not limited to, but can be found in [1,2]
3. Following these literatures [1,2], implement experiments on newsfeed data set: https://github.com/mkusner/wmd
4. And implement experiments on specific-task language word embedding: https://github.com/clips/dutchembeddings

References

1. Kusner, Matt J., et al. "From Word Embeddings To Document Distances." *ICML*. Vol. 15. 2015.
2. Tulkens, Stéphan, Chris Emmery, and Walter Daelemans. "Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource." *arXiv preprint arXiv:1607.00225* (2016).

Timeline (tentative):

February - March 2017

No	Progress meeting	Description
1	Fri 10-02-2017 16.00	Report and discussion on the construction of baseline model
2	Fri 17-02-2017 16.00	Report and discussion on implementing experiments on newsfeed data sets [1]
3	Fri 24-02-2017 16.00	Results and further investigation on different query document (online web resources)
4	Fri 03-03-2017 16.00	Report and discussion on experiments and further investigation on cross-language documents (Dutch embedding vs English embedding) [2]
5	Fri 10-03-2017 16.00	tba