Supplementary material for: Unsupervised Keyphrase Extraction with Multipartite Graphs

Florian Boudin
LS2N, Université de Nantes, France
florian.boudin@univ-nantes.fr

1 Preprocessing

We use the preprocessed version of the SemEval-2010 dataset made available by (Boudin et al., 2016)[1]. For the Hulth-2003 and Marujo-2012 datasets, we preprocess each file with the Stanford CoreNLP suite (Manning et al., 2014). LDA topic distributions are computed using scikit-learn (Pedregosa et al., 2011)[2].

2 Parameter tuning

We tuned the α parameter, that controls the strength of the graph weight adjustment, on the training set of the SemEval-2010 dataset. The plot in Figure 1 shows the variation in F_1@10 score for α values ranging from 0 to 2. Best scores are achieved at $\alpha = 1.1$ and we use this value for all our experiments.

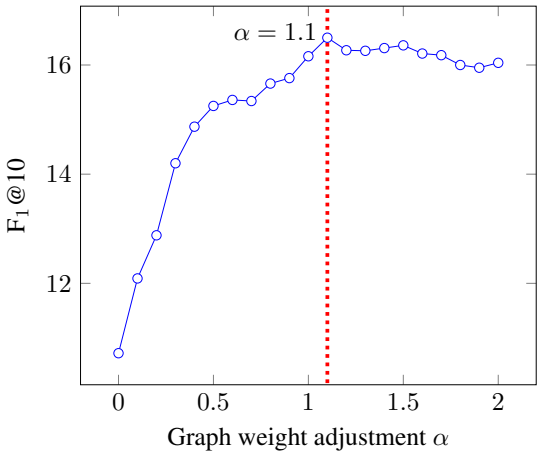

Figure 1: F_1-score computed at the top 10 extracted keyphrases for our model according to the graph weight adjustment parameter α on the training set of the SemEval-2010 dataset.

References

Florian Boudin, Hugo Mougard, and Damien Cram. 2016. How document pre-processing affects keyphrase extraction performance. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 121–128. http://aclweb.org/anthology/W16-3917.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 55–60. http://www.aclweb.org/anthology/P14-5010.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

[1] https://github.com/boudinfl/semeval-2010-pre/tree/master/test/lvl-2

[2] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html