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). 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 Parameter tuning

We tuned the $\alpha$ 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 $\alpha$ values ranging from 0 to 2. Best scores are achieved at $\alpha = 1.1$ and we use this value for all our experiments.

References


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

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