ADVANCED LEARNING FOR TEXT AND GRAPH DATA
[M2, MVA]

Maha ELBAYAD
maha.elbayad@student.ecp.fr

Lab 4 : Graph-based approaches to NLP

1 Keyword extraction

We evaluate the performances of weighted and unweighted main core extraction against the golden keywords and the two baselines (PageRank and TF-IDF) and obtain the following macro-averaged metrics:

Metric	Weighted	Unweighted	PageRank	TF-IDF
Precision	**0.603**	0.443	0.549	0.585
Recall	0.435	**0.602**	0.361	0.385
F1-score	0.408	**0.475**	0.419	0.446

Table 1: Performances - window=3

Metric	Weighted	Unweighted	PageRank	TF-IDF
Precision	**0.5851**	0.450	0.530	**0.5852**
Recall	0.453	**0.575**	0.350	0.385
F1-score	0.426	**0.466**	0.405	0.446

Table 2: Performances - window=4

The average size of weighted main cores is of 16 words against 26 of the unweighted main cores which justify the fact that the unweighted version has a better recall but a smaller precision compared to the weighted one. The F1-score taking into account the tradeoff between precision and recall favors the unweighted graph. Both versions of main core extraction outperform the PageRank and TF-IDF. When increasing the width of the co-occurrences sliding window, the unweighted model performs at the same level as the TF-IDF (independent of the window parameter).

2 Document classification

We train SVM classifiers on the WebKB' dataset that consists of academic webpages belonging to 4 classes. Each classifier takes features of a given representation [TF-IDF, TW-IDF(Normalized degree centrality) (wighted/unweighted), TW-IDF(Closeness centrality) (wighted/unweighted)]

Representation	Accuracy (w=3)	Accuracy (w=4)
TF-IDF	0.898	–
degree -unweighted	0.905	0.900
degree -weighted	0.897	0.890
closeness -unweighted	0.907	0.906
closeness -weighted	0.907	0.906

The closeness is less affected by the window width whilst the degree performs better with $w = 3$. In both tests (w=3 / w=4) the closeness yields better accuracy which is justified by the fact that it is a more global metric aggregating information from the entire graph.