A Text Feature Based Automatic Keyword Extraction Method for Single Documents

40th European Conference on Information Retrieval (ECIR 2018)
Monday 26th - Thursday 29th. Grenoble, France

Ricardo Campos[1,2]
ricardo.campos@ipt.pt

Vítor Mangaravite[2]
vima@inesctec.pt

Arian Pasquali[2]
arrp@inesctec.pt

Alípio Mário Jorge[2,3]
amjorge@fc.up.pt

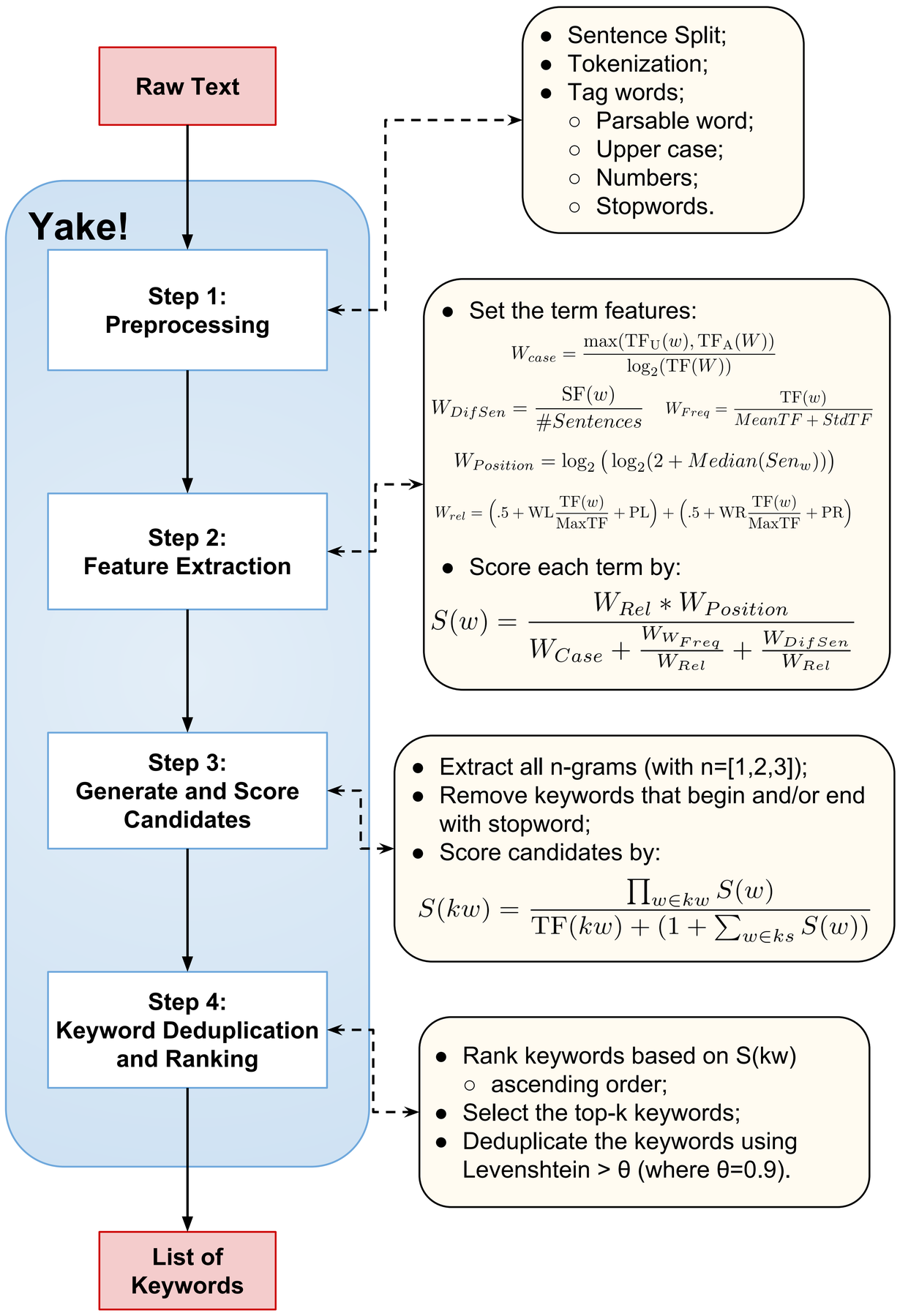
Célia Nunes[4]
celian@ubi.pt

Adam Jatowt[5]
adam@dl.kuis.kyoto-u.ac.jp

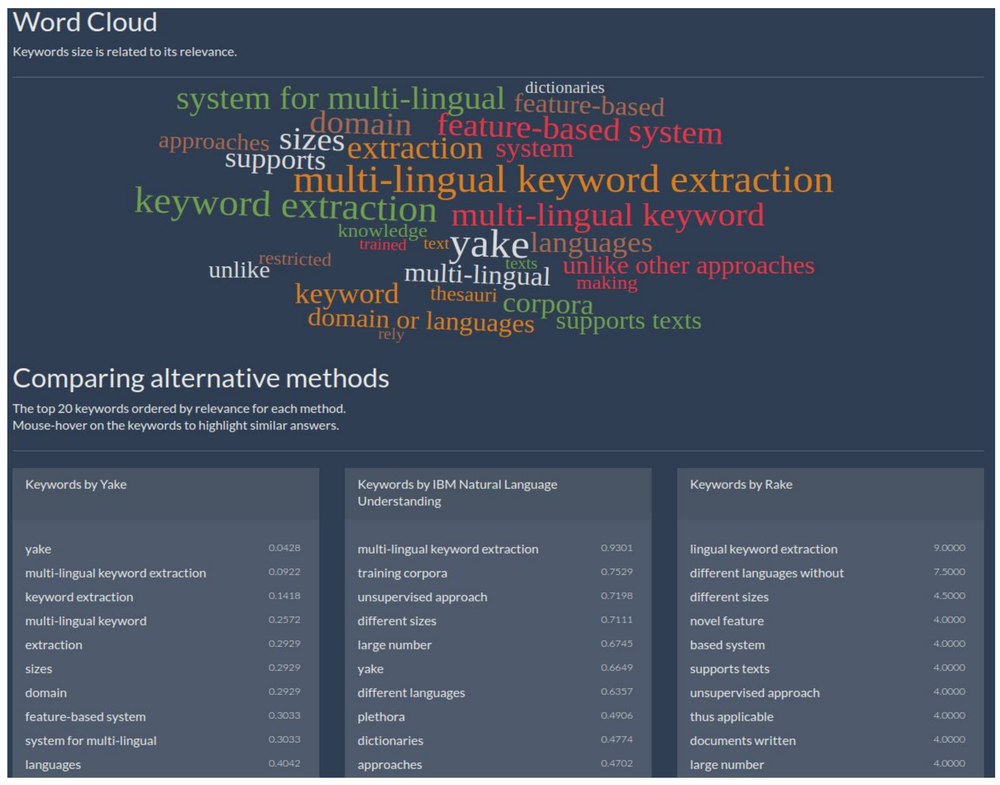
Abstract

Yake! is a novel **feature-based system** for **multi-lingual keyword extraction**, which supports texts of different **sizes**, **domain** or **languages**. Unlike other approaches, **Yake**! does not rely on dictionaries nor thesauri, neither is trained against any corpora. Instead, it follows an unsupervised approach which builds upon features extracted from the text, making it thus applicable to documents written in different **languages** without the need for further knowledge. This can be beneficial for a large number of tasks and a plethora of situations where the access to training corpora is either limited or restricted.

Keyword Extraction Process



Demo, API and Package



QR to Demo [2]

QR to API

QR to Package

Results

Method	SemEval2010			Schutz2008		
	P@10	R@10	F1@10	P@10	R@10	F1@10
YAKE!	0.153	0.103	0.123	0.217	0.058	0.091
TextRank [3]	0.101▼	0.067▼	0.081▼	0.198▼	0.052▼	0.082▼
TF.IDF [4]	0.036▼	0.023▼	0.028▼	0.100▼	0.028▼	0.043▼
SingleRank [5]	0.035▼	0.022▼	0.027▼	0.082▼	0.024▼	0.037▼
RAKE [6]	0.007▼	0.004▼	0.005▼	0.013▼	0.004▼	0.006▼

Method	500N-KPCrowd (pt)			WICC (es)		
	P@10	R@10	F1@10	P@10	R@10	F1@10
YAKE!	0.251	0.063	0.101	0.050	0.141	0.073
TextRank [3]	0.265	0.063	0.103	0.018▼	0.058▼	0.027▼
TF.IDF [4]	0.223▼	0.060▼	0.095▼	0.026▼	0.067▼	0.037▼
SingleRank [5]	0.190▼	0.054▼	0.084▼	0.017▼	0.045▼	0.024▼
RAKE [6]	0.120▼	0.038▼	0.058▼	0.004▼	0.012▼	0.006▼

Conclusions

- In this work, we propose an unsupervised lightweight approach for keyword extraction of a single document;
- YAKE! achieves better results in comparison to four state-of-the-art unsupervised keyword extraction algorithms;

Future works

- Investigate how our method performs in comparison with other datasets and other unsupervised and supervised algorithms;

[1] Campos, Ricardo, et al. (2018). *A Text Feature Based Automatic Keyword Extraction Method for Single Documents*. Proceedings of the 40th European Conference on Information Retrieval (ECIR'18), Grenoble, France. March 26 – 29.
[2] Campos, Ricardo, et al. (2018). *YAKE! Collection-independent Automatic Keyword Extractor.* Proceedings of the 40th European Conference on Information Retrieval (ECIR'18), Grenoble, France. March 26 – 29
[3] Mihalcea, Rada, and Paul Tarau. *Textrank: Bringing order into text*. Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.
[4] Hasan, Kazi Saidul, and Vincent Ng. *Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art*. Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.
[5] Wan, Xiaojun, and Jianguo Xiao. *Single Document Keyphrase Extraction Using Neighborhood Knowledge*. AAAI. Vol. 8. 2008.
[6] Rose, Stuart, et al. *Automatic keyword extraction from individual documents*. Text Mining: Applications and Theory (2010): 1-20.