UNIVERSIDADE
BEIRA INTERIOR

FACULDADE
ENGENHARIA

Departamento de
Informática

Automatic Keyword Extraction from Texts

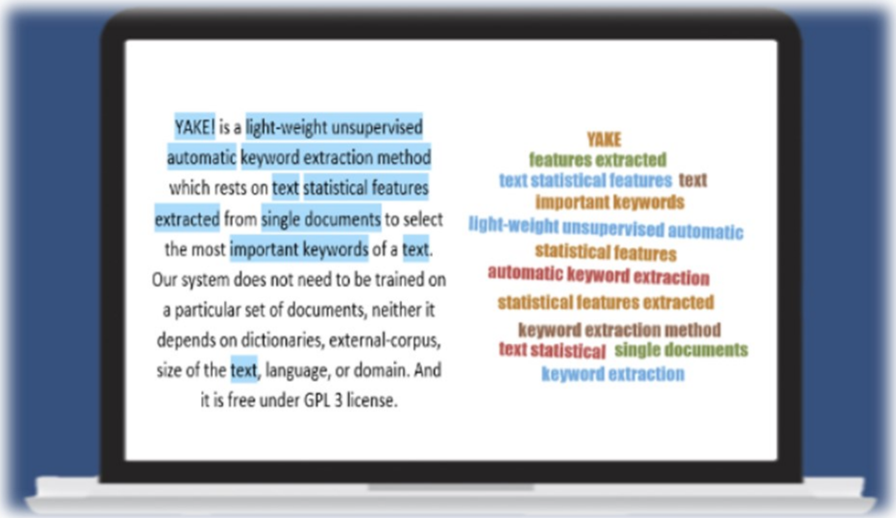
Proposta de Projeto. Lic. Eng. Informática

Orientador: Ricardo Campos (ricardo.campos@ubi.pt)

Objectives

Can you imagine how difficult would it be to analyze thousands of reports, tweets, without a keyword extraction system?

Keyword extraction (aka keyphrase extraction or detection [1]) is a natural language processing (NLP) technique that aims to automatically extract the most relevant words and phrases from a text (see Fig. 1), thus providing a concise summary of the text's content and of the main topics therein discussed. This is useful for several downstream tasks such as text analysis, social media monitoring, text summarization, text generation, text tagging, information retrieval, etc.



YAKE! is a light-weight unsupervised automatic keyword extraction method which rests on text statistical features extracted from single documents to select the most important keywords of a text. Our system does not need to be trained on a particular set of documents, neither it depends on dictionaries, external-corpus, size of the text, language, or domain. And it is free under GPL 3 license.

YAKE
features extracted
text statistical features
text
important keywords
light-weight unsupervised automatic
statistical features
automatic keyword extraction
statistical features extracted
keyword extraction method
text statistical
single documents
keyword extraction

Fig. 1: Annotated text by YAKE demo: http://yake.inesctec.pt

In this work, the student is challenged to refactor the code of an already existing keyword extraction system and to implement a few novel features. In particular, the work will rely on YAKE! a popular unsupervised algorithm [2] that has been the subject of several uses by the research community in a number of downstream tasks (e.g., text summarization, chatbots, Q&A systems, etc). A great example of its use was the creation of the general index, which used YAKE! to extract more than 19 billion keywords from over 100M documents (see INESC TEC and

Nature web articles) or its use to extract insights from Political texts (see this medium article). Also John Snow Labs, through its Spark NLP package (claimed to be the most widely used NLP library in the business sector) uses YAKE! as their portfolio solution.

The objective of this project is to refactor the code of YAKE! and incorporate/implement some new features requested by the community in the github issues of the project. In addition to this, the student should implement and make available an API of the algorithm and develop an extension of YAKE! to the different browsers available in the market.

Workplan

T1: Project setup and planning (understand existing YAKE! codebase) (2 weeks)

T2: Code Refactoring (refactor and clean up the existing code) (4 weeks)

T3: Implementation of new features (review github issues and community requests) (4 weeks)

T4: Documentation and testing (1 weeks)

T5: API development (2 weeks)

T6: Browser extension development (2 weeks)

T7: Report writing and presentation (2 weeks)

Table 1 presents the distribution of tasks for each of the 15 weeks.

Table 2: Chronology of tasks (T) per week (S).

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**
T1	▨	▨													
T2			▨	▨	▨	▨									
T3					▨	▨	▨	▨							
T4									▨						
T5										▨	▨				
T6												▨	▨		
T7														▨	▨

Technical and Academic Prerequisites

- **Proficiency in Python**: Strong programming skills in Python, as YAKE! is primarily implemented in Python.
- **Natural Language Processing (NLP) Knowledge**: A solid understanding of NLP concepts, including text preprocessing, feature extraction, and keyword extraction algorithms.
- **GitHub and Open Source Collaboration**: Familiarity with GitHub for version control and experience with open-source collaboration.

- **APIs design**: Knowledge of RESTful API design and implementation.
- **Web Development Skills**: Proficiency in web development technologies for creating browser extensions.

Expected Results

- code refactored should be made available on YAKE! github
- YAKE! github issues cleaned and new features implemented
- API of YAKE! implemented and made available
- Browser extension implemented and made available

Bibiography

[1] Papagiannopoulou, E., and Tsoumakas, G. (2019). A Review of Keyphrase Extraction. WIREs Data Mining and Knowledge Discovery, vol 10(2).

[2] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C. and Jatowt, A. (2020). YAKE! Keyword Extraction from Single Documents using Multiple Local Features. In: Information Sciences Journal. Elsevier, Vol 509, pp 257-289, ISSN 0020-0255