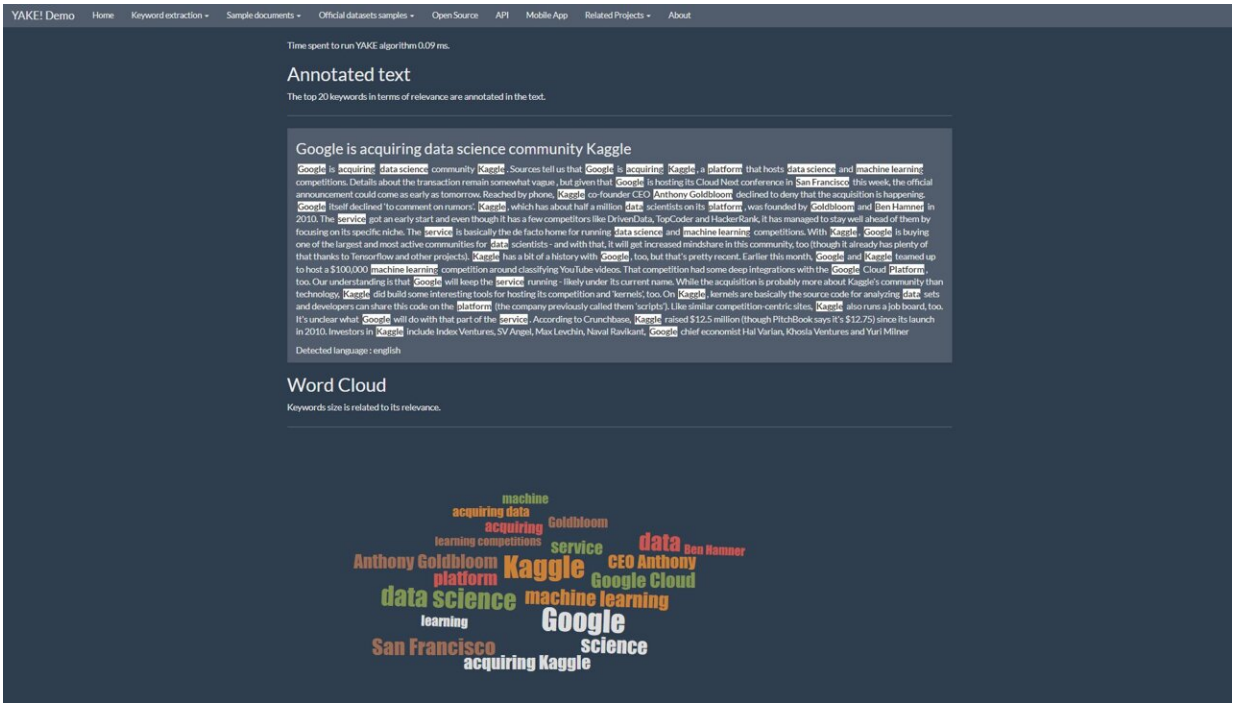
New tool can extract keywords from texts in every language about any topic

January 11 2022



It is called YAKE! ("Yet Another Keyword Extractor"), a program developed by INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, in Portugal. Its developers claim the tool can be used in texts of any size, written in any language and about any topic. YAKE! uses statistics to understand which words are

more relevant in the text, thus not needing input from other corpora of texts to learn what words are more important—like machine learning approaches usually do.

Why do we need keywords?

People might have a general idea that the amount of data produced every day is enormous. But can you really picture the quantity of data produced in one minute? For every minute of 2020, for example, Instagram users shared 65,000 photos, Twitter users posted 575,000 tweets and Google conducted 5.7 million searches. According to Siteefy, at least 175 new websites are created every minute and it is estimated Amazon publishes more than 7,500 Kindle eBooks per day. The same happens with news articles: *The Washington Post* alone publishes around 1,200 stories every day.

"The need for organizing and, more importantly, processing information, is due to the high volume of data being produced every day. A tool such as YAKE! is a precious helper in the process of automatically extracting information, by obtaining a set of relevant keywords that characterize the text itself. Doing this manually would be truly impossible," says Ricardo Campos, co-developer of YAKE!.

If you are a student, YAKE! can help you summarize texts or book chapters you need to study for your next exam. You can also benefit from using YAKE! when finding a trend on published news articles about a specific topic (such as COVID-19) or even contradictory arguments on the speeches given by a specific politician during his/her mandate. These are just some examples of what this tool could do for you, but why should you use it to extract keywords?

A new way to sort information

"Extracting keywords is a particularly complex challenge that presents relative low effectiveness/performance. YAKE! can help anyone extract keywords and sort information easily and fast," says Ricardo Campos. One of the reasons why it is so fast is the fact that it does not require previous corpora of text to work properly, unlike machine learning solutions do. "In our approach, we detect relevant keywords based on statistics extracted from the documents instead of operating on top of a document collection," he added. Furthermore, YAKE! works on the go, as a plug-and-play solution that can be used on documents of any size, language or subject.

The technology is available for free and includes a website where one can extract keywords from a text or a webpage, and an android app available on the Play Store. For developers, there is also an API that allows the integration of the technology in other tools.

The General Index & other applications

YAKE! has been used in multiple projects so far, but none came closer to the work developed for the General Index. This project aimed to catalog 107 million scientific articles, towards facilitating the search for the information they contain. The new database of 38 terabytes was launched in October and it is a giant index of 19 billion keywords extracted using YAKE! software. The collection is available under a public domain license on Internet Archive, the world's largest content preservation digital archive. However, this tool has been used in many different contexts to perform different tasks. These include summarizing educational texts for further automatic generation of comprehension questions; the generation of clarification questions in question answering systems, the detection of trending keywords on Twitter; using text mining in accident reports; generating word clouds for visually representing public opinion regarding COVID-19 on social media, and even the generation of Persian poetry from prose corpora.

Newly integrated into John Snow Labs' portfolio of open-source solutions, the most widely used natural language processing and text mining library in the business field,YAKE! is also used by the National Library of Finland, by Chartbeat Labs—textacy, and within the scope of the INESC TEC Conta-me Histórias project, included in the Portuguese web archive, arquivo.pt.

The software is currently cited or used in more than 270 articles, with more than 860 stars on Github and 141 forks, accounting for more than 1000 installations on the Android system. In 2018, it was awarded the "Best Short Paper" at the most important European conference on information retrieval, the ECIR.

In addition to Ricardo Campos, the team that developed YAKE! is composed of Alípio Jorge, Célia Nunes, Adam Jatowt, Vítor Mangaravite and Arian Pasquali.

More information: Ricardo Campos et al, YAKE! Keyword extraction from single documents using multiple local features, *Information Sciences* (2019). DOI: 10.1016/j.ins.2019.09.013

Ricardo Campos et al, A Text Feature Based Automatic Keyword Extraction Method for Single Documents, *Advances in Information Retrieval* (2018). DOI: 10.1007/978-3-319-76941-7_63

Ricardo Campos et al, YAKE! Collection-Independent Automatic Keyword Extractor, *Advances in Information Retrieval* (2018). DOI: 10.1007/978-3-319-76941-7_80

Provided by INESC Brussels HUB