

Developing a keyword extraction package in R

Written by:

Hadi Al Aithan

Abstract:

The goal of the project is to develop a program using the R programming language that can find important keywords and key phrases from a piece of text. In Python, a method called KeyBERT is available, which uses BERT embeddings and cosine similarity to find which words are most similar to the overall meaning of the document. Choosing R can be helpful when the rest of the work is already being done in R, especially for users familiar with R or when integrating keyword extraction into statistical reports. In the end, the program will be an R-based version of KeyBERT, capable of producing comparable keyword extraction results using BERT embeddings and cosine similarity.

Introduction:

Searching for specific information within large datasets has become increasingly challenging. A common solution to this problem is to organize the data using keywords and key phrases, thereby facilitating more efficient retrieval. Among the algorithms designed for keyword extraction is **KeyBERT**, which uses BERT-based embeddings to transform text into vector representations and identify semantic correlations between textual segments.

Challenges:

To implement a KeyBERT-like package in R, an embedding model is required for text representation. Unlike Python, R lacks efficient native models for this task. One possible approach is to use the **reticulate** package, which allows Python libraries to be accessed directly from R. Alternatively, the **text2vec** package can be employed with GloVe embeddings to generate word vectors; however, utilizing pretrained models typically requires external tools. Consequently, extracting high-quality word embeddings natively within R remains a challenging task.

Conclusion:

In the future, it will be necessary to develop word embedding models natively in R. The current lack of sufficient models makes it difficult to implement KeyBERT directly in this language. Further research on improving keyword extraction algorithms is also needed, and R can play an important role by providing robust statistical tools for comparing different approaches. Since KeyBERT is among the most effective keyword extraction algorithms, implementing it in R would facilitate direct comparisons with alternative methods and contribute to identifying the most efficient techniques for keyword extraction.