Survey of Text Retrieval, Processing and Analysis Toolkits

CS410 Fall 2021 Tech Review
Author: Aaron Shipway

Introduction

The advent and development of the world wide web has resulted in an explosion in the amount of unstructured text data available for retrieval and mining. Given the high degree of interest corporations have in finding new and better ways of targeting customers for advertising and the interests of researchers in the fields of information retrieval and data mining, software applications have been developed, largely in the last 20 years, to assist in the retrieving, processing, and mining of text data. Accurately and efficiently retrieving desired text data given it is generally the product of human language requires the computer to be able to "understand" what the author of a segment of text intends to communicate with that text. Natural language processing (NLP) is a field of computer science, linguistics and artificial intelligence that aims to develop methods of assisting computers in this task. This survey briefly describes four software toolkits which provide methods for text retrieval, processing, and analysis: Apache OpenNLP, Apache Lucene, Natural Language Toolkit (NLTK), and Modern Text Analysis (MeTA).

Apache OpenNLP

Initiated in 2010 and released in 2015 with the latest version (1.9.3) released in 2020, OpenNLP is, as the name suggests, an open-source toolkit for natural language processing. Its machine learning based Java toolkit contains methods for many NLP tasks which can be used to build a text processing service. Examples of commonly used processing methods: tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, language detection and coreference resolution, information grouping, speech recognition, feedback analysis, natural language generation, coreference resolution. It also includes maximum extropy and perceptron based machine learning methods. In addition to text processing there are components in the toolkit for building and evaluating models, giving it text analysis capability. It provides both an API and command line interface for flexibility.

Apache Lucene

Apache Lucene is, like OpenNLP, an open-source, Java based API toolkit; however, it's focus is to provide tools and components for developing information retrieval software, especially search applications. It was originally written in 1999 and became a top-level Apache project in 2005. The current version 8.10.1 was released in October 2021. The stated goal of Lucene is to "provide world class search capabilities" and toward that end Lucene offers: "scalable, high-performance indexing"; "powerful, accurate and efficient search algorithms"; licensing options for both commercial and open-source programs; and implementations in languages beside Java that are index-compatible (C, C++, Python, Perl, Ruby, Lisp, .NET, PHP, Delphi). Lucene's query capabilities allow a range of query customization options to the user such as queries based

on wildcards, phrases, proximity, or range. Search types include ranked, fielded, multiple-index, and update-and-search. Lucene uses a unique inverted index structure that allows efficient, scalable indexing at rates of over 150GB/hour with small RAM requirements (1MB heap).

NLTK

The Natural Language Toolkit (NLTK) is an open-source, Python-based platform for natural language processing. NLTK was developed in academia, initially released in 2001 and the current version 3.6.1 was released in April 2021. Although it was developed for academic research in computational linguistics, its user-friendly manual and comprehensive API documentation make it highly accessible to students and industry users as well and is currently one of the most popular toolkits for natural language processing. Among NLTK's features are processing tools for tokenization, stemming, tagging, parsing, semantic reasoning and classification as well as statistical analysis tools. In addition to processing and analysis tools, NLTK includes many text corpora of various kinds, some of which have user friendly interfaces. NLTK is often used for prototyping text processing applications due to its ease of use and broad capabilities.

MeTA

Modern Text Analysis (MeTA) is an open-source C++ based toolkit for information retrieval and machine learning. Its aim is to provide a "unifying framework" for researchers by providing broad and integrated text processing and analysis tools to help overcome the inefficiencies caused by the fragmentation of algorithm offerings and data formatting standards found in the open-source software community. By combining and tightly integrating the capabilities of a search engine toolkit such as Lucene and a text analysis toolkit, MeTA will ideally provide a means for researchers increase the efficiency of their experimental work. MeTA provides capability for tokenization, filtering, part-of-speech tagging, parse trees, text analysis, classification, ranking, topic models, language models, natural language processing, and search engine creation. MeTA may be used in a Python environment with metapy Python bindings.

Conclusion

The four toolkits surveyed here include the search engine toolkit Lucene, the processing and analysis toolkits OpenNLP and NLTK, and a toolkit which combines search, processing and analysis capabilities – MeTA. For search engine development, Lucene is the clear choice given its efficiency and scalability in combination with its variety of query and search options. Users focused on natural language processing may choose between OpenNLP, NLTK and possibly MeTA. NLTK is likely the best choice for a novice given the available user-friendly documentation and that it is Python based; however, it was developed for academics and thus has a variety of advanced functions related to classification and statistical analysis which, in addition to its inclusion of a variety of useful corpora, makes it a good choice for anyone doing NLP. MeTA's combination of strong search, processing and analysis capabilities makes it an ideal choice for anyone doing information retrieval and analysis research. Its metapy Python interface makes it more accessible to many novice users and its integration of a broad range of

functionality makes it ideal for researchers who don't want to be held up by platform deficiencies and incompatibilities.

References

OpenNLP

- http://opennlp.apache.org/
- https://www.tutorialspoint.com/opennlp/opennlp_overview.htm

Lucene

- https://lucene.apache.org/
- A. Sharma, *Practical Apache Lucene 8*, Apress, 2020

NLTK

- https://www.nltk.org/
- https://www.tutorialspoint.com/natural_language_toolkit/index.htm
- Bird, S., Ewan Klein, Edward Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, 2009
- Wang, M., Fanghui Hu, "The Application of NLTK Library for Python Natural Language Processing in Corpus Research", *Theory and Practice in Language Studies*, Vol. 11, No. 9, pp. 1041-1049, September 2021

MeTA

- https://meta-toolkit.org/