Tyler Wong

Professor ChengXiang Zhai

CS 410: Text Information Systems

15 November 2020

The Lemur Toolkit - Sifaka

Sifaka is an open-source text mining application that is part of the Lemur Project, developed at the University of Massachusetts, Amherst and Carnegie Mellon University. The Lemur Project started in 2000 and it is still being continually updated by their contributors. Sifaka is built using Java above Apache Lucene with a desktop GUI that supports running on Windows, Mac, and Linux. It is continually being maintained with the latest version 1.8 being released on 6/22/2020. The main features of Sifaka are that it supports full-text search, document frequency analysis, co-occurrence analysis, and the ability to export feature vectors compatible with Weka. In addition, Sifaka as well as the other software as part of the Lemur Project can be found on their SourceForge project page for download. The download includes both the source code and pre-built packages.

The primary benefit of using Sifaka is that it enables a person to quickly explore and analyze large text collections through it's GUI interface. Instead of having to go through a programming interface, you are able to quickly use Sifaka to get up and running text mining and text analytics. If you are brand new to text mining and text analytics, Sifaka provides you the ability to jump start your interest because you won't need to write any code to get started. In addition, Sifaka provides many pre-built document parsers for many different types of data and sources. They provide out of the

box support for these document parsers: plain text, reuters 21578, HTML, WARC, Twitter, simplified TREC, and Wall Street Journal.

There are two main components of Sifaka, the first is the Index application and the second is the Text Miner application. In order to actually analyze the text data, Sifaka first needs to build the index using Lucene and that is done using Sifaka's Index application. This is where you would point Sifaka to your data source and various options to index and annotate the data. One this is done, it builds the Lucene index into a folder so that you can reuse this index without having to rebuild it every time you want to analyze the text. This is really important because indexing the text data can take a very long time especially if your computer is not up to the task. The second component, the Text Miner application is where you are able to view and query the index that the first component created. In addition, this is where you are able to generate feature vectors, calculate co-occurrence, and frequency without writing any code at all; this is all done through the GUI still. Since Sifaka is open source you are also able to add new additions to the code to supplement the existing features. For example, you are able to add a new document indexer so that you can index the data that is important to your use case.

In conclusion, I think Sifaka is a great application for beginners to get into text mining and text analytics without having to write any code. Sifaka hasn't gained widespread attention though so there may be bugs and not much support if you run into problems or have questions. So if you were to use Sifaka I would recommend supplementing it with gaining knowledge in text mining and text analytics so that you are able to understand what Sifaka is doing behind the scenes with your text data.

References

- https://www.lemurproject.org/sifaka.php
- https://www.lemurproject.org/sifaka_tutorials/properties.html
- https://www.lemurproject.org/sifaka_tutorials/quickstart.html
- https://www.lemurproject.org/about.php
- https://arxiv.org/pdf/1810.02907.pdf