Google Style Rank Based Searching for Investment Funds with Thousands of Data Attributes

Mr. Fakruddin Mohammed
fam281@g.harvard.edu

Motivation

Searching (text based) for information has become essential part in the era of digital world. We all know how the search should work but the reality is implementing a good search engine is extremely challenging task. Therefore, in this short tutorial I wanted to explore the search technologies.

Problem Statement

Nowadays every business have a Digital Front End providing informational services to the end users. When the user is looking for information, the ability to give the most relevant information is key ingredient to the success of the business who purely rely on digital data services. There are number of ways one can add text based search features to the application such as Relational Database queries or No-SQL based data queries or Remotely hosted CGI scripts. But the issues with these approaches is that it doesn't use contextual information from what the user is asking for, it doesn't give the results sorted by relevance and the free text based search implementation is extremely difficult task, often simplified or compromised to make implementation easier therefore, it doesn't retrieve the results accurately what the user is actually looking for. Therefore, in this short article the study explores the alternate ways to implement the search engine features using the investment funds data (mutual funds, exchange traded funds and hedge funds data). The reason for choosing the financial instruments data is due to the fact that a given investment product can be described in using thousands of attributes, therefore, it challenges the implementation. The data is web scraped from the publicly available websites and where data is not available, initialised with random data and normalised to the OpenFunds standards.

Technology

The following software stack is used to build the device.

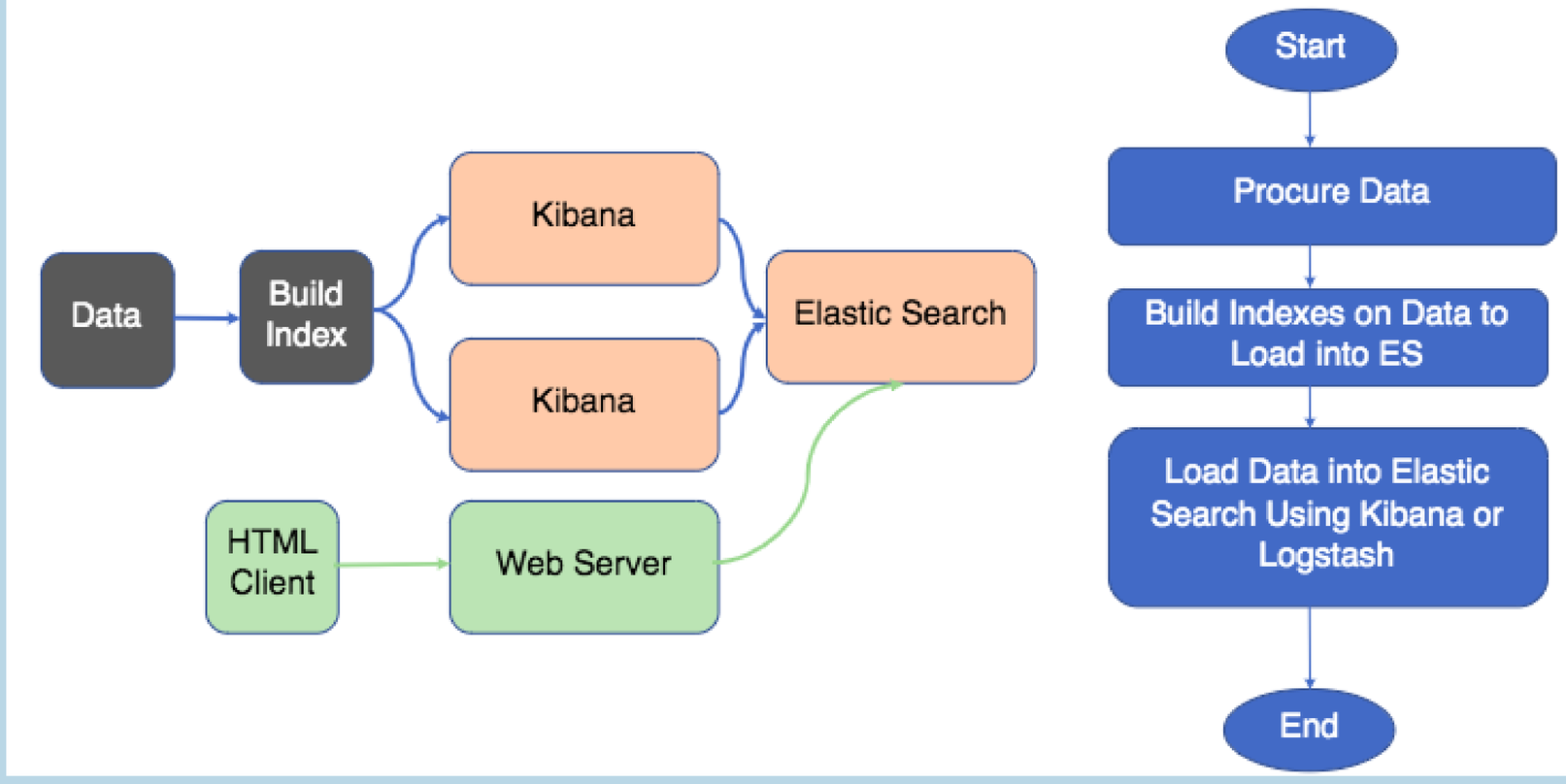
- ELK Stack (ElasticSearch, Logstash Kibana)
- HTML, CSS
- Java Script
- Python
- Apache
- Excel/VBA

Hardware

The experiment is conducted on the following hardware platform.

- MacOSX Laptop (4 CPU, 8GB RAM, 256 SSD Hard disk)
- Python Simple HTTP Server

Architecture

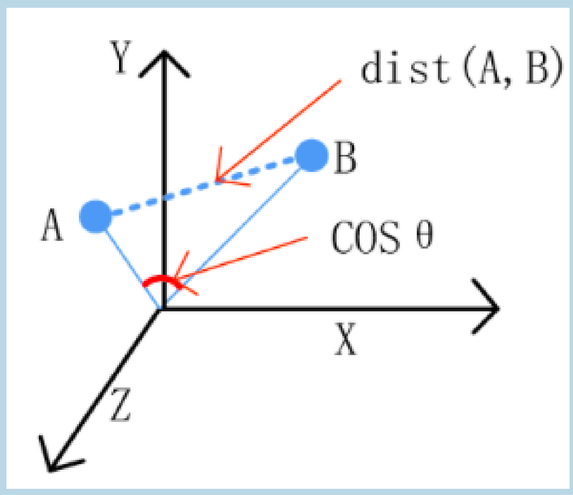


Approach, Data Flow & Demo, Current Status and Next Steps

Approach For google style free-text search, there are two approaches are available.

- **Index Based**: Build a index for every word in the document (equivalent to row in a database or entity) and apply $\frac{TF}{IDF}$ to calculate the rank. In the above equation **TF** (term frequency) refers how many times the given search keywords have appeared in the document and **IDF**(inverted document frequency) refers to how many times the search key words have appeared in the rest of the documents.
- **Natural Language Processing based Word Embeddings**: This technique refers to generating N-Dimensional word embeddings and then calculating the cosine similarity $\frac{A.B}{\|(A).\|(B)}$ to find the relevance and for ranking scores.

For the purpose of this study, index based search is implemented using the fund data scraped from the publicly available websites on the Elastic Search Engine stack. The data flow as described in the flowchart above consists of the following steps:



- Collate cleanse the data
- Build the word index for every keyword in the document
- Load into the eleasticsearch engine using Kibana console or Logstash
- Use Kibana console to test the search and ranking scores.

Current Status Using the fund data scraped from the public websites and further the data is normalised to the OpenFunds [3] data attributes format is loaded into the elasticsearch engine. Please refer to the YouTube video for short demo and My Github respository for the resources & detailed step-by-step instructions if you would like to replicate this experiment.

Next Steps Add the voice to the search engine where the human voice is translated into text and return the search results to the user.

References

[1] Open Funds Website: `https://www.openfunds.org/fieldmenu/fields/`, this website has the data attributes dictionary used by the industry to describe the fund data.

[2] Lifewire Website: `https://www.lifewire.com/searching-your-site-3466200/`, this website gives brief backgroun on website search engine features (2017).

[3] Elastic IO Website: `https://www.elastic.co/`, this website has the link to all elastic search engine components documentation.

YouTube URLs Github Repository `https://github.com/mohammed-fakruddin/Google-Based-Free-Text-Search-Engine-for-Investment-Funds`