NLP Based Linguistic Search Engine for Fund Objectives
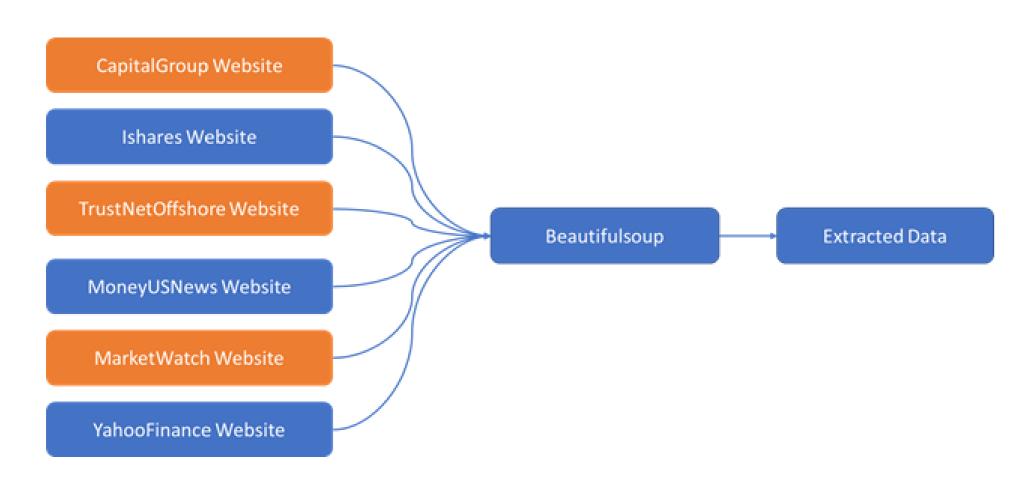
Fakruddin Mohammed

Harvard Extension School, Harvard University in Cambridge, Massachusetts, USA

Objective

The business objective is to provide the large and growing number of average investors with a sophisticated search engine which can help bridge the gap between the legal language of a prospectus and the vocabulary of retail investors. To achieve the objectives, a linguistic search engine was built using advanced NLP text similarity algorithms to match an investor's simple text query to the language of the prospectus, allowing someone without intimate knowledge of the investing industry to find mutual funds matching their risk appetite and objectives.

Data Technology

Data Objectives of mutual funds are publicly available as part of their prospectuses. Using the Beautifulsoup web-scraping library, the fund objectives for 3000+ funds are extracted from various finance and investment related websites.

Technology Most of the implementation work for this Proof of Concept (POC) was conducted on the following hardware: Laptop with 8GB RAM, 8th Generation i7 Processor, no GPU. Finally, the entire application was successfully tested by deploying on to AWS EC2 (8GB RAM/m2.large) cloud instance for hosting the website and NLP linguistic models. The following software and libraries are used in this project for building the linguistic search engine:

Technology
Python Flask Framework
NLTK
Spacy
Gensim
Tensorflow 1.14, Tensorflow Hub
Scikit Learn
Matplotlib
Beautifulsoup 4 (extraction of fund objectives)
PRAW (extraction of investor objectives from Reddit)

Models In this study the focus is on finding semantic similarity between investor objectives and fund objectives using corpus-based algorithms. The following NLP models are used to evaluate semantic similarity between investor statements and fund objectives and the results are compared.
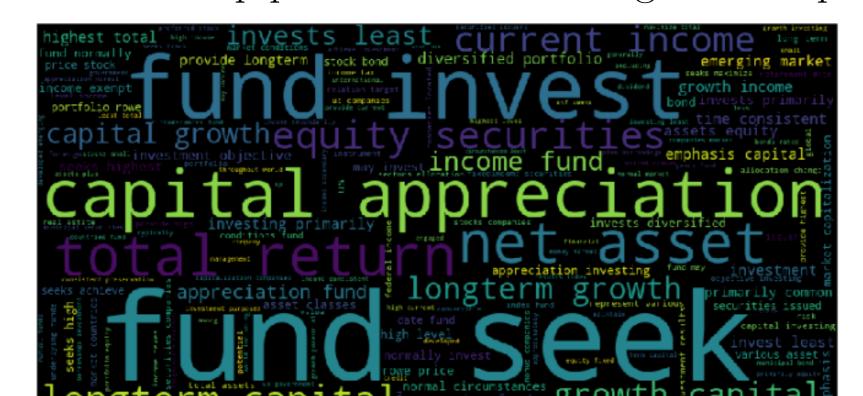
Models
Word2Vec
TF-IDF
TF-IDF/LSI
ELMO
Universal Sentence Encoder
FastText

The end-to-end NLP data pipeline is shown in Figure on top right hand side.

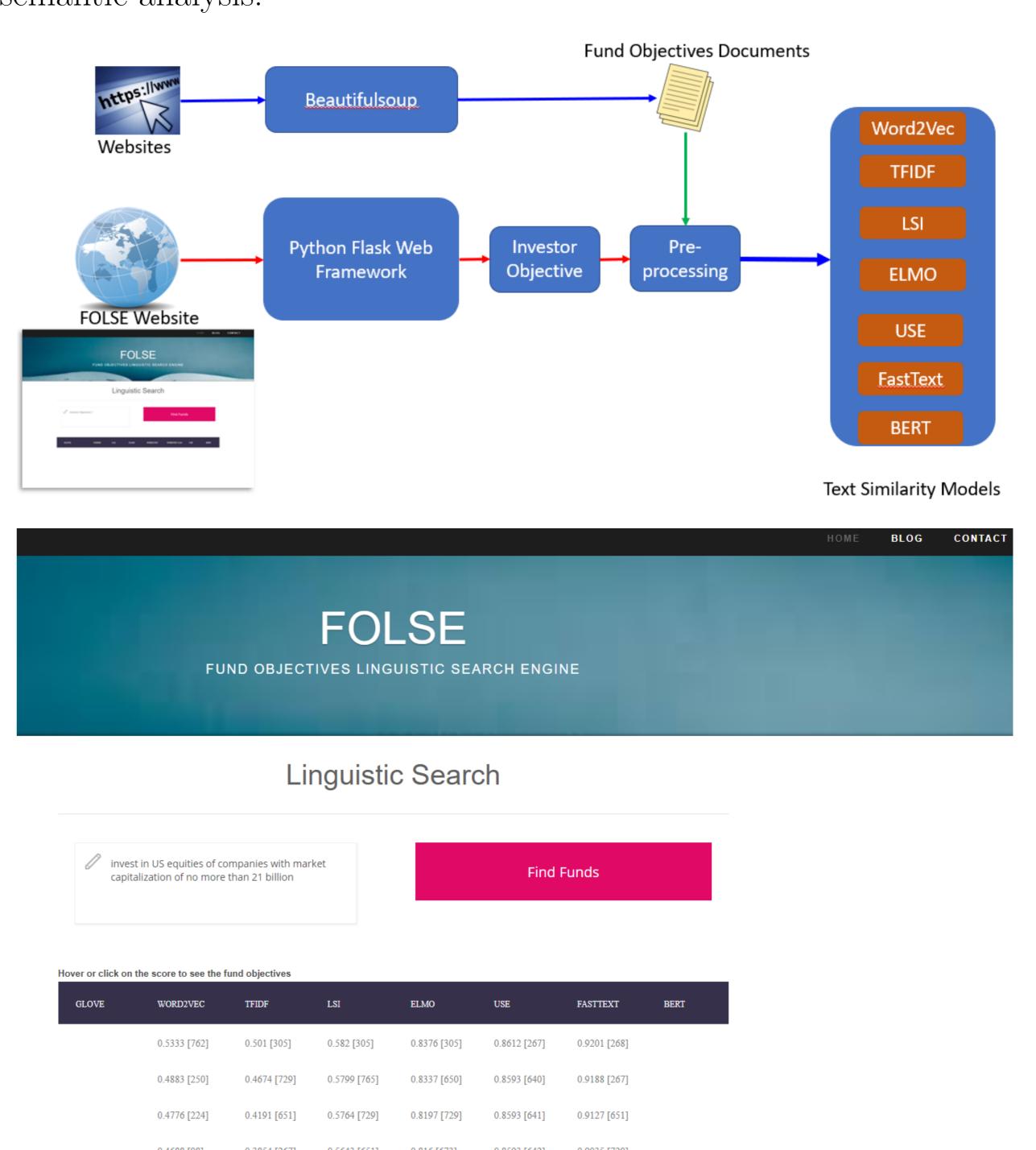## Approach

The Modelling approach at high level is described below.

- The fund objectives that were web-scraped from 6 websites are cleansed by running through a standard NLP pre-processing pipeline. The results of this NLP pre-processing pipeline is shown in the form of word cloud on right hand side (bottom).
- Hyper dimensional Word embeddings are generated for number of NLP models specified in the table left below.
- The investor objective query string received from the FOLSE website at run-time is also pre-processed using the same pipeline before feeding the result to the six different models.
- In a loop, lexical similarity scores are calculated between the investor objective and the funds objectives. The top four funds based on their objectives' similarity with the user input are returned to the FOLSE website as recommendations.

Results Conclusion

ELMO, USE (Universal Sentence Encoder) and Fast Text NLP model performance is much superior than Word2Vec, TF-IDF and LSI, especially, is the investor fund objectives query string is longer and verbally complex. For simple query sentences, TF-IDF is performance is similar to mode advanced models such as ELMO/USE/FastText.

Next Steps - Aim to improve models' performance with regard to understanding the context of numerical information, such as 'at least 5 billions', 'at most 1 billion', 'over 40%', '60 to 80%' etc. Pure text similarity may not be sufficient here and it may be necessary to dig deeper into various forms of semantic analysis.

GLOVE	WORD2VEC	TFIDF	LSI	ELMO	USE	FASTTEXT	BERT
	0.5333 [782]	0.501 [305]	0.582 [305]	0.8376 [305]	0.8612 [287]	0.9201 [288]	
	0.4883 [250]	0.4674 [729]	0.5799 [785]	0.8337 [610]	0.8593 [840]	0.9188 [307]	
	0.4776 [226]	0.4191 [651]	0.5764 [729]	0.8197 [729]	0.8593 [641]	0.9127 [651]	
	0.4688 [98]	0.3854 [267]	0.5643 [651]	0.816 [673]	0.8593 [842]	0.9035 [729]	

Contacts & Links

For implementation details, source code and other useful information, please visit to my GitHub link below. A short video of the Linguistic Search Engine is also available on YouTube (link below).

- Github:https://github.com/mohammed-fakruddin/NLPLinguisticSearchEngineForFundIndustry
- YouTube Video:https://youtu.be/spkuaf-rd7g