



Increasing the Efficiency of Text Analysis Programs Using a Multi-Tiered Approach

By Daniel O'Brien and Advisor Michael Gildein

Abstract

Natural Language Processing is an increasingly researched field that has a few major barricades holding back its progress. One of these barricades is the intensive computational costs used by its processes. The goal for this project is to explore potential efficiencies that can be gained by implementing a multi-tiered filter based analysis system. A text-analysis script was written and designed in order to prove this hypothesis. The resulting data indicates that a multi-tier approach can be successfully implemented for an increase in efficiency.

Technologies Used

- Python
 - Programming Language
- Jupyter Notebook
 - Integrated Development Environment
- Github
 - Source Code Management
- Doc2Vec
 - Natural Language Processing Python package
- Pandas
 - Data manipulation Python package
- Nltk
 - Natural Language Tokenizer Python package
- Sklearn
 - Vector Analysis Python package
- BeautifulSoup
 - Web scraper python package

Corpora

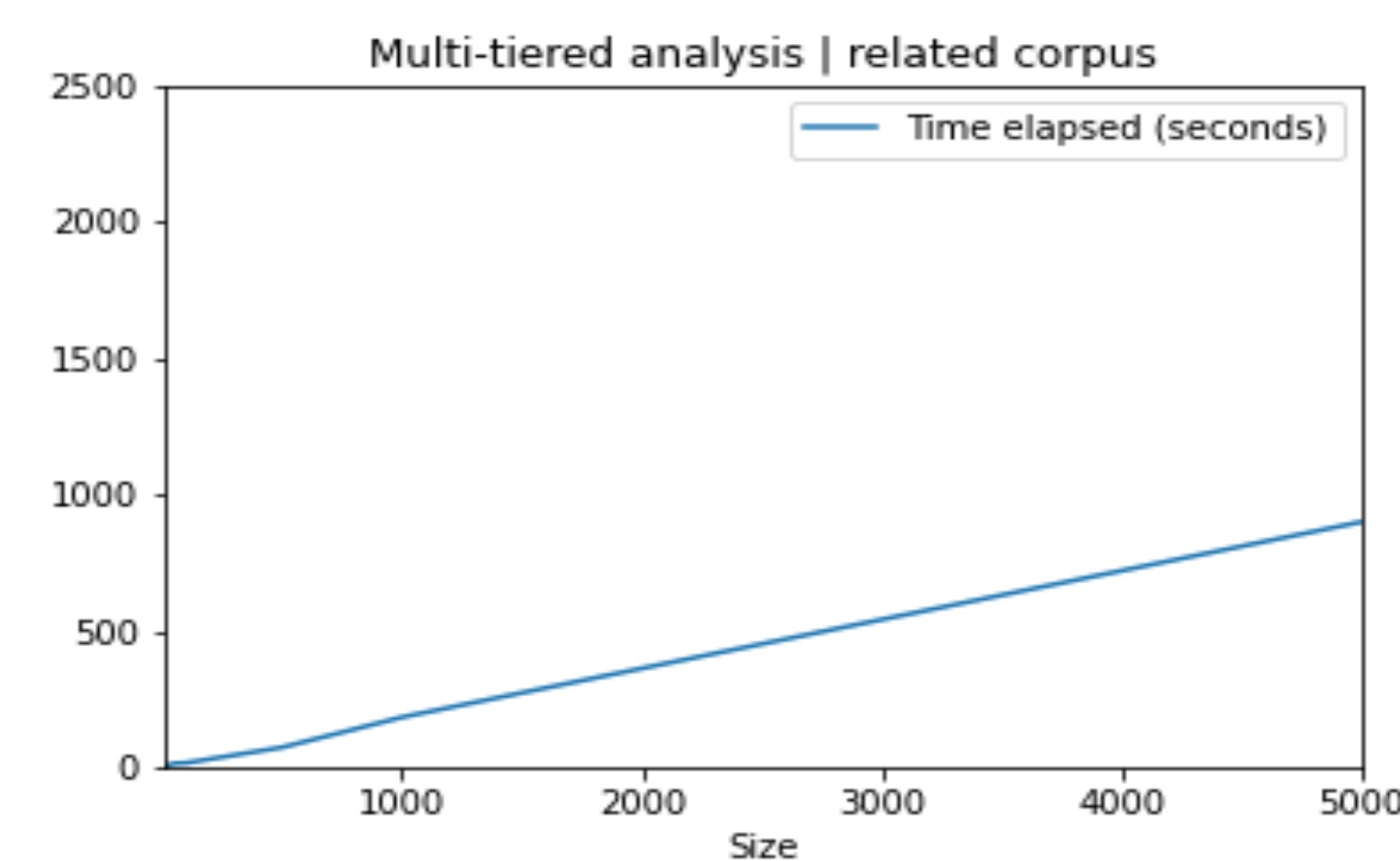
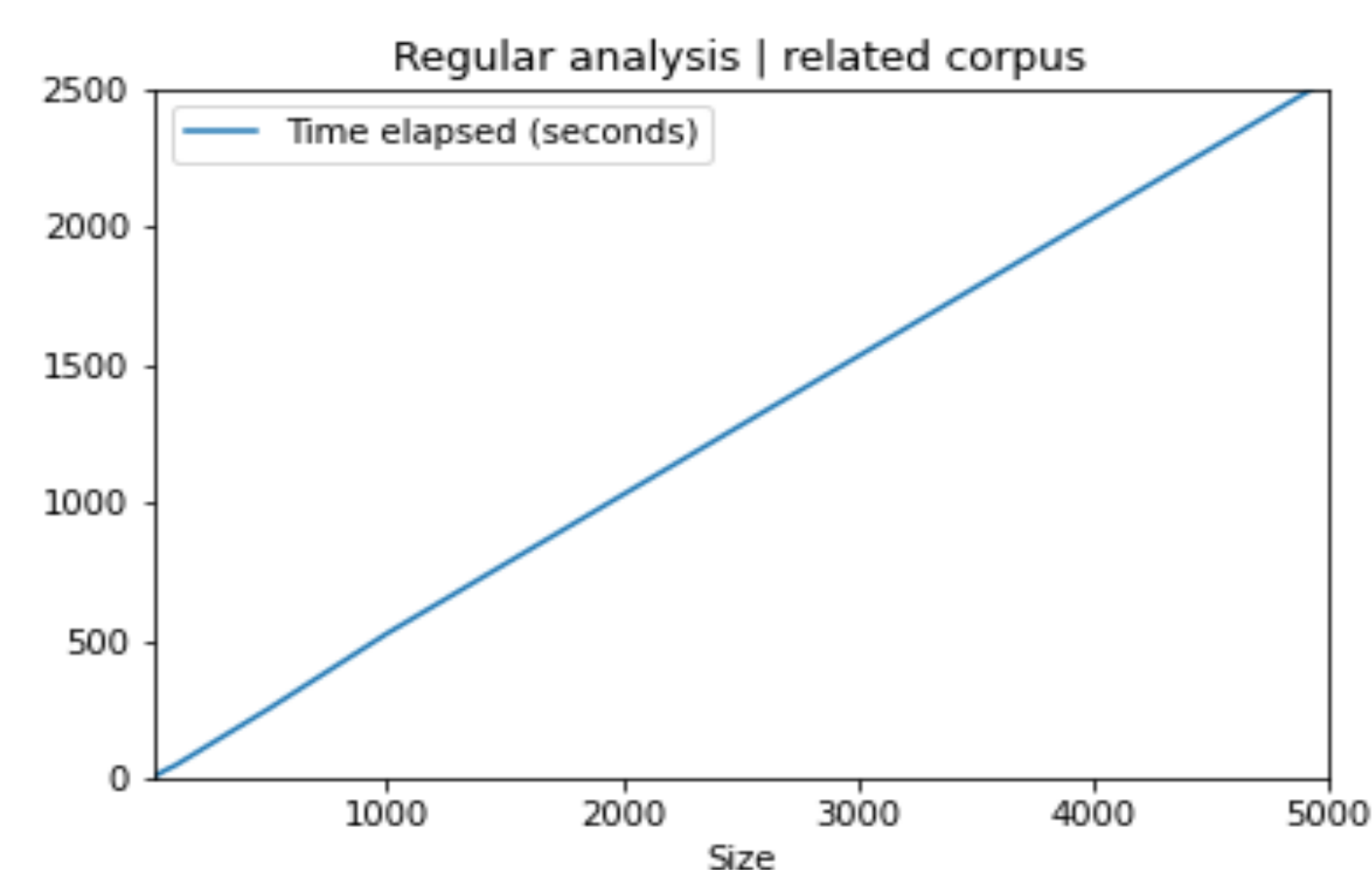
- **Related**
Corpus is filled with articles related to given seed. Worst case scenario
- **Fifty Fifty**
Corpus is half filled with related articles, half filled with random articles. Average case scenario.
- **Random**
Corpus is filled with entirely random articles. Best case scenario.

Real World Applications

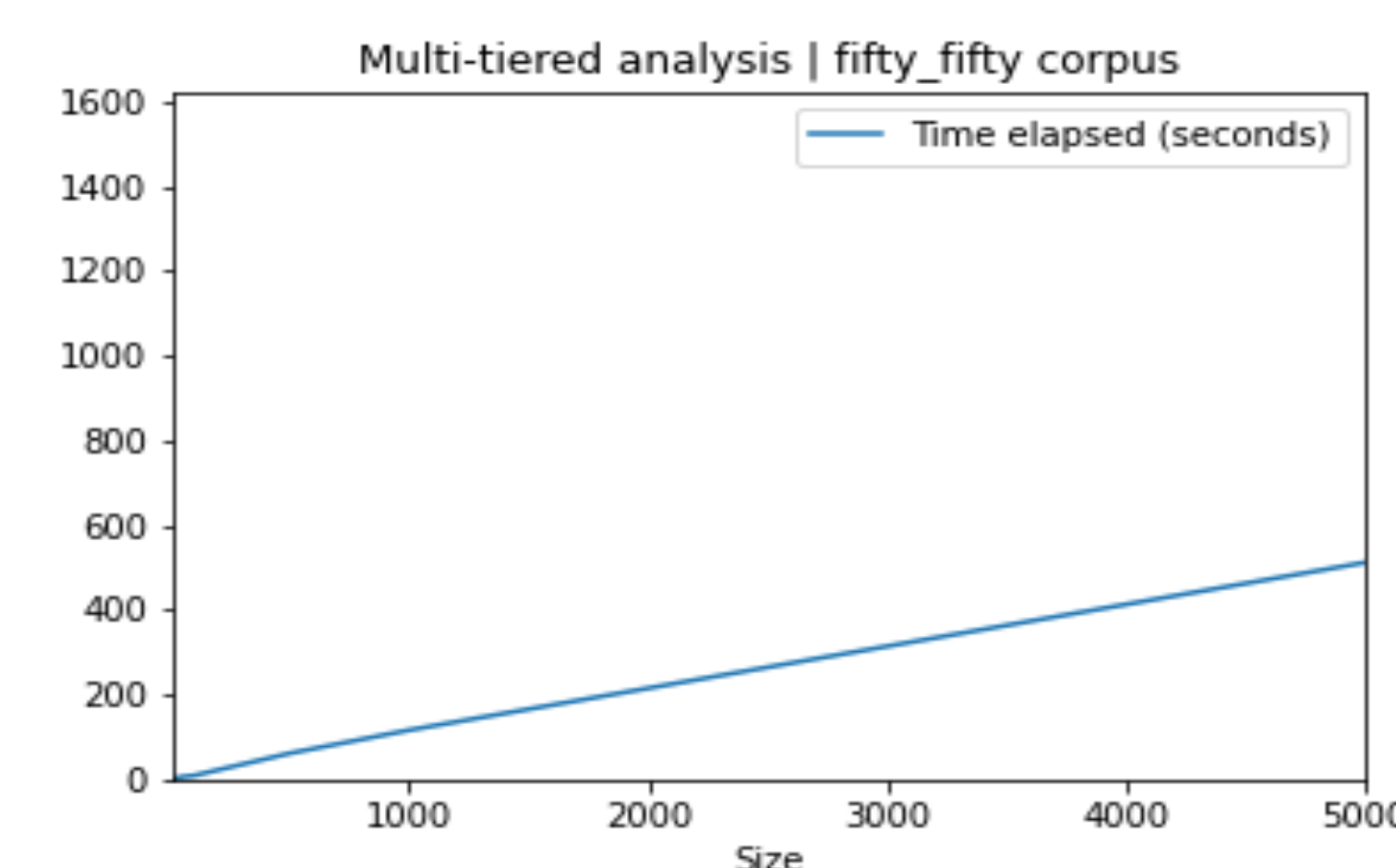
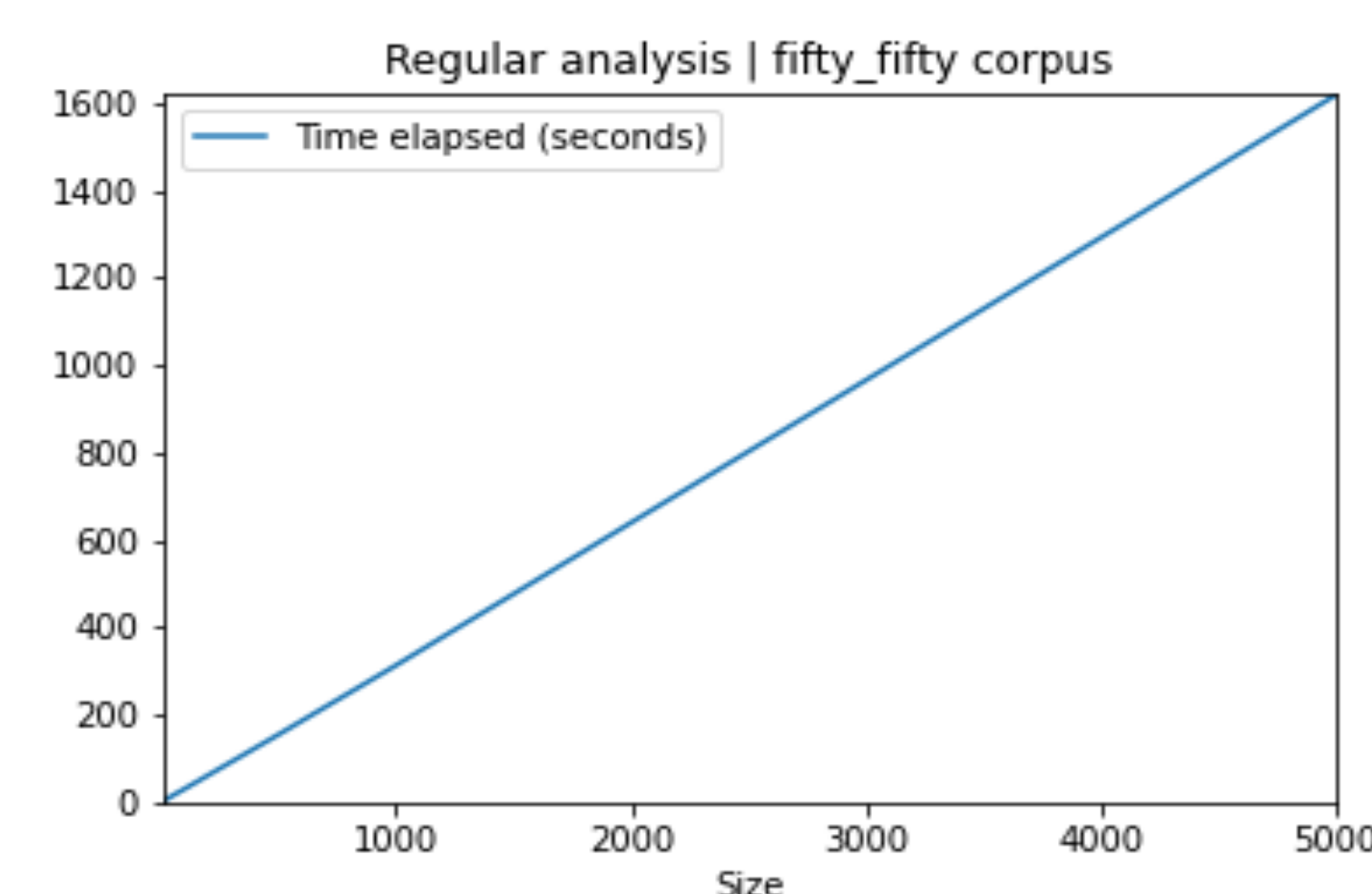
- **Patent Enforcement**
 - Patents can be compared to product descriptions quickly and efficiently allowing companies to sift through large volumes of potential infringements.
- **Credit Transfers**
 - Enable universities to efficiently compare the course descriptions and make a guided decision as to whether or not the credits should transfer.

Results

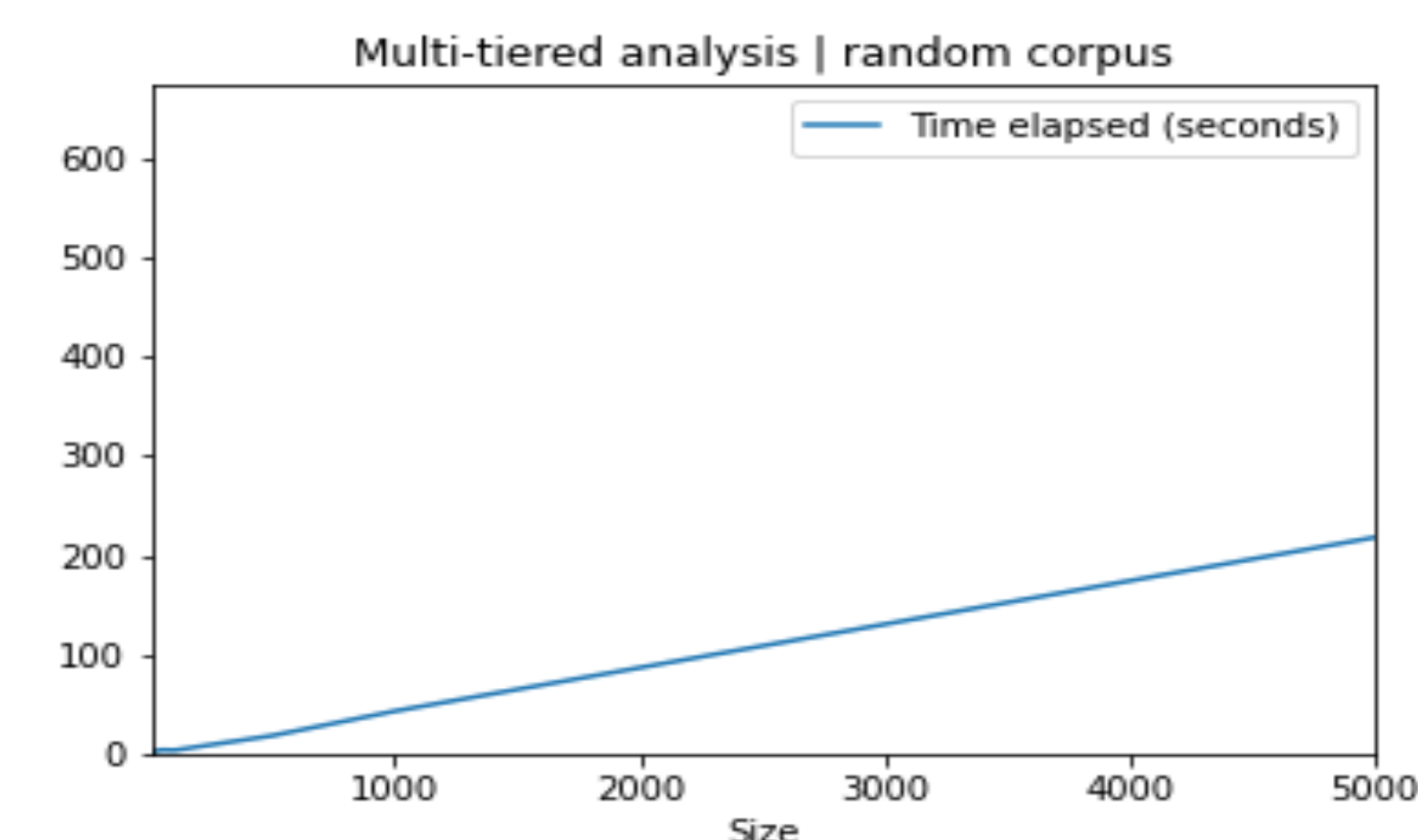
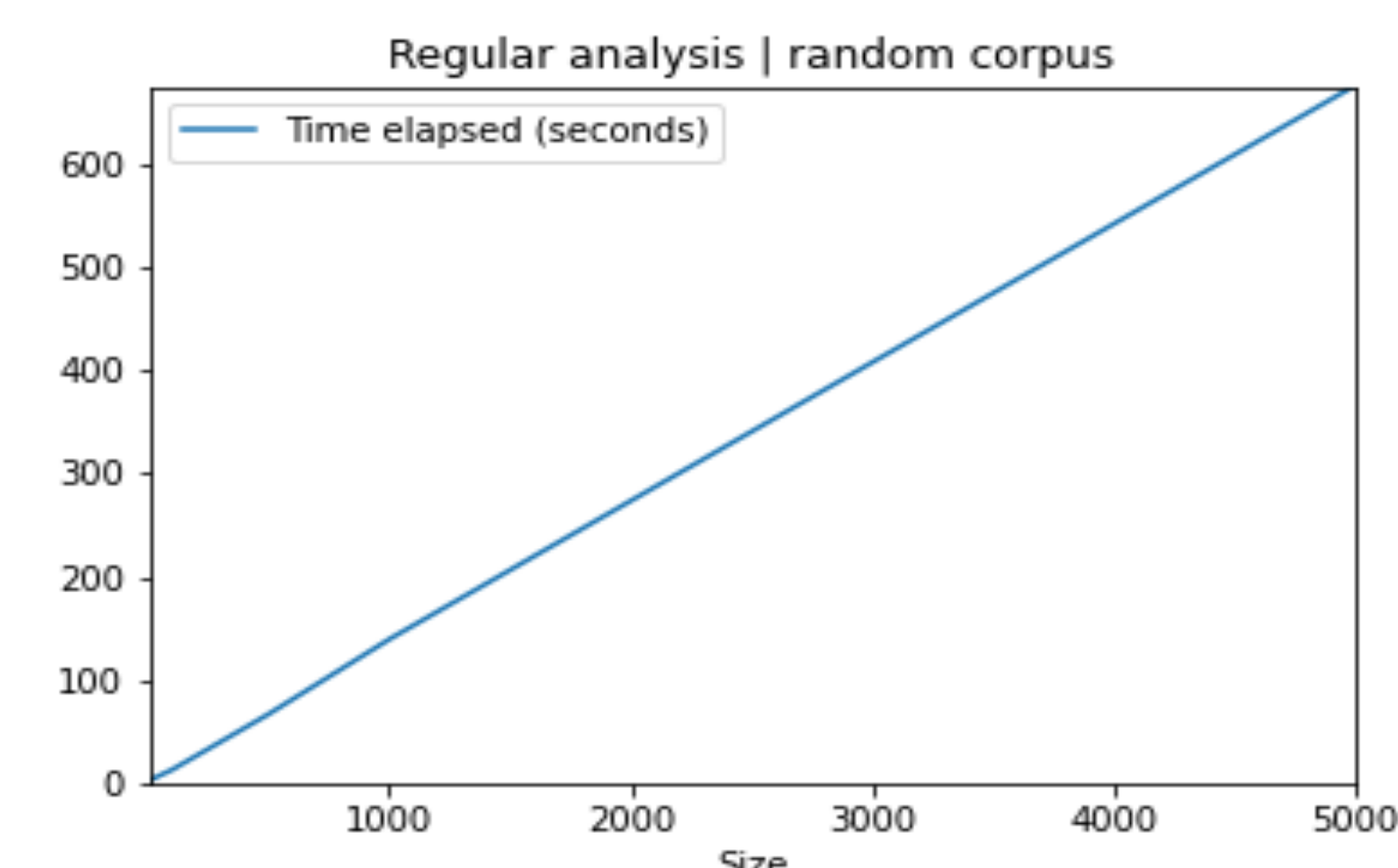
Related



Fifty Fifty



Random



Conclusion

This data shows that it is possible to boost efficiency in natural language processing applications using a multi-tiered approach. It also shows that this approach works best on broad sets of documents that span multiple different topics. However, it still proves to be effective for boosting efficiency on more focused data sets. This means that the approach has a wide range of potential applications in the Data Science Industry. Considerations should be made to implement a similar system in any application involved in natural language processing and text similarity analysis.