

Mining Big Data - Final Report

Investigating Paragraph Vectors

a1632538 Zachary Forman

a1646930 James Caddy

October 30, 2016

Introduction

Contributions

This work makes the following contributions:

- Replication of a relatively recent paper [1] that makes some contested claims [2].
- Comparison of `doc2vec`'s suitability as a document similarity measure as compared to Jaccard distance.
- Application of the `doc2vec` algorithm to a novel domain – source code.

Background

Word Vectors

Paragraph Vectors

Document Similarity

Data Collection

Data Sources

We used three sources of data. Firstly, we use Maas' IMDB set [3] for verification and English language documents which can be found at <http://ai.stanford.edu/~amaas/data/sentiment/>. Secondly, we used The Wikimedia Foundation's English Wikipedia data dump [4] as a source for English language documents. We downloaded the full wiki dump from <https://dumps.wikimedia.org/enwiki/20160820/>. Lastly, we used The Chromium Project's Chromium repository [5] as a source for source code documents. We cloned the repository at revision 4fe31bb06cf458234d7017950a8b2b82427487c8.

Data Scale

The IMDB dataset contains a total of 100,000 files. Half are labelled as either a positive or negative reviews, and the remainder are unlabelled. It is 346MB in size and contains 23 million words, from a vocabulary of 90 thousand words.

The English Wikipedia Dataset is a single XML file, 13GB compressed and 55GB uncompressed. It contains 16 million pages, however this includes redirects and meta-pages such as categories and portals, and therefore only 5 million of these are useful as documents. Excluding markup, this dataset contains approximately 4 billion words.

The Chromium codebase is 2.6GB in size, and contains 245 thousand files. Of these, 23 thousand are C and C++ files, which contain 4.8 million lines of code and 608 thousand lines of comments.

Data Format

The fundamental requirements for the `doc2vec` [1] algorithm to work are that each document can be associated with the words in the document, and that words in each document are just a series of words, i.e. no excess punctuation or markup.

This requirement can easily be met by storing our set of documents as a newline separated list of space separated tokens. The following example shows how a simple document set might be transformed into this file format:

```
{ "My cow was sick, but now it is better.", "Apples are nice.", "Amazing results!" }
```

Convert all text to lowercase and remove all punctuation:

```
{ "my cow was sick but now it is better", "apples are nice", "amazing results" }
```

Concatenate documents, separated by newlines:

```
my cow was sick but now it is better
apples are nice
amazing results
```

This format is simple to produce, simple to parse, and - given that the `doc2vec` algorithm needs to process every word anyway - is efficient to read. Perhaps more importantly, especially for large datasets like the Chromium dataset, this storage format has no storage requirements in excess of just the raw text in the files.

Data Processing

To transform the data into the format previously described, some data processing is required. All programs that we used to transform the datasets are available at <https://github.com/mbd-doc2vec-team/mbd-doc2vec/tree/master/data-processing>.

IMDB Dataset

The IMDB dataset needed minimal processing. The bulk of the work had already been performed by the authors of the dataset. Beyond what had already been done, punctuation, symbols and elements of markup were removed, and each document concatenated into a file as discussed in the previous section.

Wikipedia Dataset

The Wikipedia dataset required extensive pre-processing, with each document needing to be parsed from the XML, and the plain text extracted from the wiki markup. Additionally, “fake” documents (e.g. portals, category pages) were removed. Each wiki page was then tokenized and stored in a file as previously stated.

Chromium Dataset

The Chromium codebase required extensive pre-processing as well, with each C++ file being tokenized and having unicode characters (e.g. `̀`) removed. Unlike the other two datasets, because punctuation and uppercase/lowercase distinctions are semantically significant in code, we retained punctuation and the original casing in tokens. Each comment and string literal was then tokenized as well, with punctuation being removed from these tokens only, and the resulting tokens being stored as discussed in the previous section. For an example of this process, the document

```
// Frobulates the bar, in an efficient manner.
void Frobulate(Bar &bar);
```

would be represented by a line in the file as follows

```
Frobulates the bar in an efficient manner void Frobulate ( Bar & bar ) ;
```

Experimental Procedure

IMDB Dataset

Wikipedia Dataset

Chromium Dataset

Results

IMDB Dataset

Wikipedia Dataset

Chromium Dataset

Conclusion

Future Work

References

- [1] Quoc V Le and Tomas Mikolov. “Distributed Representations of Sentences and Documents.” In: *ICML*. Vol. 14. 2014, pp. 1188–1196.
- [2] *Google Groups*. URL: <https://groups.google.com/d/msg/word2vec-toolkit/q49firnoqro/bp--14e4unwj>.
- [3] Andrew L. Maas et al. “Learning Word Vectors for Sentiment Analysis”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 142–150. URL: <http://www.aclweb.org/anthology/P11-1015>.
- [4] Meta. *Data dump torrents — Meta, discussion about Wikimedia projects*. [Online; accessed 9-August-2016]. 2016. URL: https://meta.wikimedia.org/w/index.php?title=Data_dump_torrents.
- [5] *Git repositories on chromium*. URL: <https://chromium.googlesource.com/>.