Mining Big Data Project Proposal

Investigating Paragraph Vectors

a1632538 Zachary Forman
a1646930 James Caddy

August 10, 2016

Topic

We seek to investigate the use of the dimensionality reduction technique of *Paragraph Vectors* [1], also known as `doc2vec` [2] for use in document similarity algorithms and understanding source code.
We seek to achieve three goals:

1. Replicate the results claimed by Le and Mikolov in their 2014 paper [1].
2. Compare `doc2vec` backed with k-nearest-neighbours with standard approaches to determining document similarity [3].
3. Apply `doc2vec` to source code, which has never been done before, and explore how it distributes code segments.

Data Sources

Le and Mikolov's original paper shows the most pronounced results on Maas' IMDB dataset [4]. We therefore seek to use this dataset to reproduce Le and Mikolov's claims. It consists of 100000 reviews, where 25000 of them are labelled for training, 25000 of them are labelled for test and 50000 are unlabelled, but usable for unsupervised training. For English language documents, we plan to use a standard datasource used to train and evaluate both `word2vec` and `doc2vec` [2, 5], the data dump of all English Wikipedia articles [6]. This dump is around 50GB uncompressed, consisting of approximately 5 million articles and 3 billion words. For source code, we plan to use the Chromium repository [7]. This repository contains over 7 million lines of C++ source code, and over 1.5 million lines of comments.

Goals

- To reproduce Le and Mikolov's results, we expect to see 92% test accuracy [1] on the sentiment analysis task for the IMDB dataset.
- From the Wikipedia datasource, we intend to perform the following experiments:
  - Use t-sne [8] to visualize Wikipedia articles, and see if patterns based on e.g. article category can be seen.
  - Compare the similarities found by the w-shingles algorithm with min-hashing to the similarities found by our `doc2vec` based approach.
- From the Chromium datasource, we intend to perform the following experiments:
  - Use t-sne [8] to visualize code snippets, and see if the resulting pattern has human relevant meaning.
  - Compare the document vectors for code snippets and their associated comment blocks, and see if they are closely related.

Contributions

This research project will have the following contributions:

- Replication of a relatively recent paper that has some contested claims [9].
- Implementation and analysis of a novel method based on the `doc2vec` algorithm for analysing document similarity.
- Application of the `doc2vec` algorithm to a novel domain – source code.

References

[1] Quoc V Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents." In: *ICML*. Vol. 14. 2014, pp. 1188–1196.

[2] Jey Han Lau and Timothy Baldwin. "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation". In: *arXiv preprint arXiv:1607.05368* (2016).

[3] Andrei Z Broder. "On the resemblance and containment of documents". In: *Compression and Complexity of Sequences 1997. Proceedings*. IEEE. 1997, pp. 21–29.

[4] 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.

[5] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[6] 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.

[7] *Git repositories on chromium*. URL: https://chromium.googlesource.com/.

[8] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605.

[9] *Google Groups*. URL: https://groups.google.com/d/msg/word2vec-toolkit/q49firnoqro/bp--14e4unwj.