Word embeddings for semantic similarity: comparing LDA with Word2vec

Luandrie Potgieter[1] and Alta de Waal[1,2]

[1] Department of Statistics, University of Pretoria
[2] Center for Artificial Intelligence Research (CAIR)

1 Introduction

Distributional semantics is a subfield in Natural Language Processing (NLP) that studies methods to derive meaning, and semantic representations for text. These representations can be thought of as statistical distributions of words assuming that it characterises semantic behaviour. These statistical distributions are often referred to as embeddings, or lower dimensional representations of the text. The two main categories of embeddings are count-based and prediction-based methods. Count-based methods most often result in global word embeddings as it utilizes the frequency of words in documents. Prediction-based embeddings on the other hand are often referred to as local embeddings as it takes into account a window of words adjacent to the word of interest. Once a corpus is represented by its semantic distribution (which is in a vector space), all kinds of similarity measures can be applied. This again leads to other NLP applications such as collaborative filtering, aspect-based sentiment analysis, intent classification for chatbots and machine translation.

In this research, we investigate the appropriateness of Latent Dirichlet Allocation (LDA) [1] and word2vec [4] embeddings as semantic representations of a corpus. Once the semantic representation of a corpus is obtained, the distances between the documents and query documents are obtained by means of distance measures such as cosine, Euclidean or Jensen-Shannon. We expect short distances between documents with similar semantic representations and longer distances between documents with different semantic representations.

2 Experimental setup

In our experiment, we choose the 20 Newsgroups dataset which contains 20000 documents that are partitioned across 20 different newsgroups [3]. Examples of the newsgroup topics are motorcycles, religion, computer software, etc. Our experimental workflow is as follows:

1. Train the word embeddings (LDA and word2vec) on the complete dataset. The output of this step is a vector space representation of the corpus.
2. Choose documents of one newsgroup, say motorcycles. Split this newsgroup into a 80%, 20% split. The 80% subset is referred to as the reference set and the 20% is referred to as the query set.

3. Index the 80% reference set and the 20% query set (separately) on the vector space representation obtained in step 1.
4. Calculate the similarity between the reference and query set. For LDA, we use the Jensen-Shannon distance [2], and for word2vec we use soft cosine [5].
5. Create an alternative query set from a different newsgroup, say religion. Also index this query set on the vector space representation obtained in step 1.
6. Calculate the similarity between the reference and the alternative query set.

Figure 1 gives an illustration of initial results. In this experiment, an LDA semantic representation was obtained. The 'motorcycles' newsgroup was used as the reference set. The 'politics' newsgroup was chosen as alternative query set. Slightly shorter distances can be seen for motorcycle-motorcycle comparisons (green), where motorcycle-politics distances tend more towards longer distances. Future work include experimentation on word2vec and additional datasets.

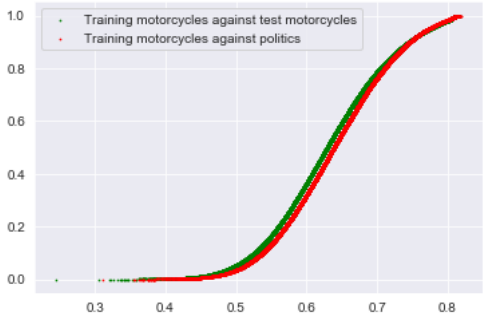

Fig. 1. Cumulative Jensen-Shannon distances between motorcycle-motorcycle and motorcycle-politics corpora.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**, 30 (2003)
2. Fuglede, B., Topsoe, F.: Jensen-Shannon divergence and Hilbert space embedding. In: International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings. pp. 30–30. IEEE, Chicago, Illinois, USA (2004). https://doi.org/10.1109/ISIT.2004.1365067
3. Karishma Borkar, Nutan Dhande: Efficient Text Classification of 20 Newsgroup Dataset using Classification Algorithm | International Journal IJRITCC. International Journal on Recent and Innovation Trends in Computing and Communication **5**(6) (Jun 2017)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs] (Jan 2013), http://arxiv.org/abs/1301.3781, arXiv: 1301.3781

5. Sidorov, G., Gelbukh, A., Gomez-Adorno, H., Pinto, D.: Soft similarity and soft cosine measure: similarity of features in vector space model. Computación y Sistemas **18**(3) (Sep 2014). https://doi.org/10.13053/cys-18-3-2043