Word Embeddings – Random (Ecostate) vs Trained Models

Vijay Mallidi

19635574

A project proposal submitted for 300597 – Master Project 1

in partial fulfilment of the requirements for the degree of

Master's in Data science

Supervisor: Laurence Park

School of Computing, Engineering and Mathematics

Western Sydney University

April 2020

1. Background

In the earlier days most of the Language Processing Systems uses hand coded sets of rules such as grammar rules or heuristic rules. Since late 1980's and mid 1990's most of the natural language processing systems depend on machine learning where the system automatically learns the rules for Language processing through analysis of large sets of documents. Although this approach offers simplicity and robustness, it has many limitations such as the amount of relevant data for speech recognition. Introduction of word embeddings starts the concept that the words can be represented as the dimensional vectors, where all the words with similar meaning stays close to each other in the vector space. According to this approach words can have multiple degrees of similarity. Results obtained by Learning high dimensional embeddings from a large data are the most accurate.

Although this technique provides the best results, the scalability of the training method makes it challenging. Then the Random Ecostate approach comes into existence which may provide similar results while using a very less data as compared to the previous trained model approach.

Many researchers have done the work on any one of these two approaches. However, there is no solid evidence that favours the one approach over the other, so I intend to compare the accuracy of these two techniques. Although Random ecostate isn't trained, it might be as good as the Trained models because the work in high dimensional spaces, Random Projections provide a good method for dimension reduction with very little loss in accuracy.

2. Objective

To compare the difference in accuracy of Random (Ecostate) system and Trained models for creating word embeddings.

3. Hypothesis/question

As our objective is to compare the difference in accuracy of random (Ecostate) system and trained for creating word embeddings several questions arise to achieve this, such as

- Data used to compare the accuracy
- Tests to compare the accuracy
- Parameters on which these tests are performed

4. Methodology

4.1 Literature

One of the greatest challenges in Artificial Intelligence and Machine learning is Understanding text and Natural Language. So, any type of developments in this domain are considered very crucial. Understanding the meaning and usage of words is very important in increasing the efficiency of the Machine learning. Last 2 decades has seen a lot of changes and developments in understanding words and some of the noticeable developments in understanding the similarity of meaning of the words are Introduction to Latent Semantic Analysis by *Landauer, T. K., Foltz, P. W., & Laham, D. 1997*, Probabilistic Latent Semantic Analysis by *Hofmann, T., 2013*, and Latent Dirichlet Allocation by *Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003*.

The Authors of the paper on **Introduction to Latent Semantic Analysis** Landauer, T. K., Foltz, P. W., & Laham, D describes Latent Semantic Analysis as a method that can be used to extract and represent the contextual-usage meaning of the words by statistical computations applied to a large corpus of text. The author describes greatly about the similarity of the meaning of the words by aggregating all the word contexts that does and does not appear. Latent Semantic Analysis (LSA) processes the large sample of words combined together to form meaningful sentences and passages, and represents the words used in these sentences and passages as points in a very high dimensional Semantic space. Latent Semantic Analysis's similarity estimates are derived from a powerful mathematical analysis which are capable of dealing with even more deeper relations. Out of many limitations of LSA some of them are that LSA doesn't use the order of the words, doesn't have any logic or relations. Even with the many limitations, LSA manages to get the decent results in getting the similarity of the meaning of the words.

The Authors of the paper on Unsupervised method called **Probabilistic Latent Semantic Analysis (PLSA)** Hofmann, T describes Probabilistic Latent Semantic Analysis as a novel statistical technique for the analysis of two-mode and co-occurrence data. This has applications in information retrieval, NLP, machine learning, and in related areas. The Probabilistic Latent Semantic Analysis is based on a mixture decomposition derived from a statistical latent class model as compared to standard Latent Semantic Analysis that uses linear algebra and singular value Decomposition. The statistical model used in Probabilistic Latent Semantic Analysis is called aspect model which is the model variable of co-occurrence of data which in our case is usually words. The authors have mentioned the advantages of the PLSA over the Standard Latent Semantic Analysis.

The Authors of the paper on **Latent Dirichlet Allocation** Blei, D.M., Ng, A.Y. and Jordan, M.I describes LDA as a model for collections of discrete data that provides full generative probabilistic semantics for documents. In LDA documents are modelled using a Dirichlet random variable which can be describes as a probability distribution on a latent low-dimensional topic space. The distribution of the words of a document which hasn't been seen yet is treated as a continuous mixture over all the documents space and a discrete mixture over all possible topics. This paper mostly concentrates on finding the correct document during the search using the words rather than just similarity between the words.

4.2 Data Used

Now, our project to do the comparison for the two approaches the data required is very large. So, I choose to work with one of the most popular social networking sites, Twitter. Getting Tweet data is very easy and very understandable to most of the audience.

Twitter data have five data objects that comprises of different fields:

I. **Tweet** has the fields id, text, attachments, author_id, created_at, entities, geo, in_reply_to_user_id, lang, possibly_sensitive, referenced_tweets, source, public_metrics, withheld.
II. **Media** has the fields media_key, type, duration_ms, height, preview_image_url, url, width.
III. **Poll** has the fields id, options, duration_minutes, end_datetime, voting_status.
IV. **Place** has the fields id, full_name, contained_within, country, country_code, geo, name, place_type.
V. **User** has the fields id, name, username, created_at, description, entities, location, pinned_tweet_id, profile_image_url, protected, url, verified, withheld.

For the trained models, first we need to divide the twitter data into 2 parts as training data set and testing dataset. Training data set is used to train the model we are going to use, and testing data set is used to get the output. Whereas for random ecostate we just run the testing dataset through the random ecostate to obtain the results.

4.3 Testing the accuracy

Now to test the accuracy of these two techniques we use an evaluation Toolkit for Universal Sentence Representations called SentEval. SentEval performs various tests as shown in the below Table 1. All the sentences manually classified with values from 1 to 5.

Name	Task
MR	Sentiment(Movies)
CR	Product Reviews
SUBJ	Subjectivity/Objectivity
MPQA	Opinion Polarity
TREC	Question-Type
SST-2	Sentiment(Movies)
SST-5	Sentiment(Movies)

Table 1 – Various Tests performed in SentEval

But since our output is represented in vectors for all the words, we will have to convert the words stored in vectors to a meaningful sentences that we have used in SentEval. Then the outputs of both the techniques can be compared with the sentences in SentEval to get the accuracy of both the techniques.

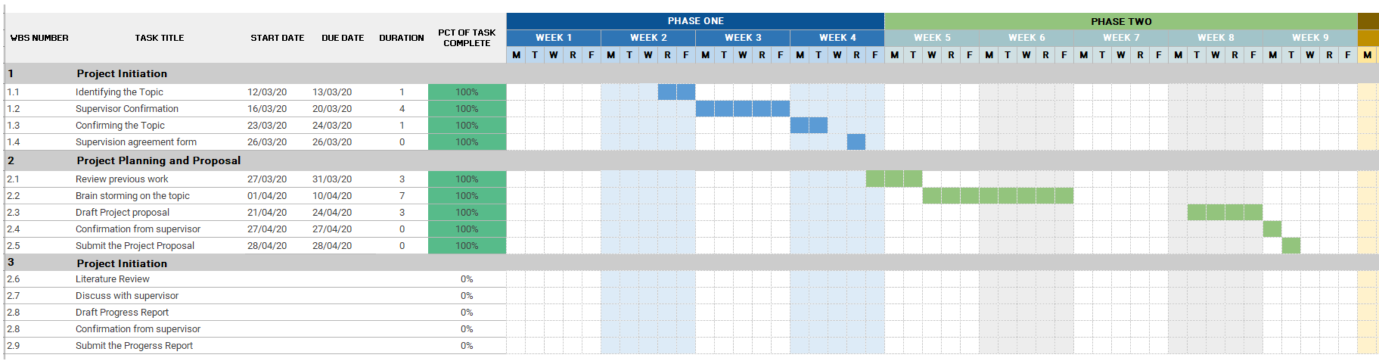
5. Expected Outcomes

At the end of this project I would get the better understanding of word embeddings and the two different approaches Trained models and Random (Ecostate) models. Also, in this project we will be using an evaluation Toolkit for Universal Sentence Representations called SentEval to test the accuracy of both the techniques which is a new method of testing the accuracy for word embeddings.

After the comparison of both the techniques I expect that the trained models will be more accurate. If the accuracy of Randomly assigned model (Ecostate) is close to what we have from trained models it would be very efficient option to choose Randomly assigned model for word embeddings.

6. Program of Work

The plan for the project is represented in the Gantt chart below:



7. References

Achlioptas, D., 2001, May. Database-friendly random projections. In Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 274-281).

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2002. Latent dirichlet allocation. In Advances in neural information processing systems (pp. 601-608).

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

Conneau, A. and Kiela, D., 2018. Senteval: An evaluation toolkit for universal sentence representations. arXiv preprint arXiv:1803.05449.

Hofmann, T., 2013. Probabilistic latent semantic analysis. arXiv preprint arXiv:1301.6705.

Landauer, T.K., Foltz, P.W. and Laham, D., 1998. An introduction to latent semantic analysis. Discourse processes, 25(2-3), pp.259-284.

Landauer, T.K. and Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), p.211.

Li, P., Hastie, T.J. and Church, K.W., 2006, August. Very sparse random projections. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 287-296).

Wieting, J. and Kiela, D., 2019. No training required: Exploring random encoders for sentence classification. arXiv preprint arXiv:1901.10444.