Natural Language Processing
A study of Language Models using word2vec

Isabel B. Amaro, Adriano Veloso

[1]Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil

{isabel.amaro,adrianov}@dcc.ufmg.br

Abstract. *Clustering is a Data Mining technique capable of grouping data with some non trivial similarity. With the advance of technology and data generation, clustering algorithms are required to provide useful information. Therefore, Data Mining researchers are constantly proposing new clustering techniques and algorithms. This survey overviews some state-of-the-art clustering algorithms, organizes and presents a temporal analysis in order to support the studies of its readers.*

1. Introduction

With the popularization of social networks, a massive amount of data is constantly being generated in the last few years [Cambria and White 2014]. Natural Language Processing is a Computer Science field that studies the best ways to retrieve, process and generate from representations of human language [Cambria and White 2014]. To work with human language, Language Models are used in NLP. A well known Language Model is word2vec, which represents each word from a corpus as a vector. Each word in the space is related to a context, which are the other words that are around it.

This work starts explaining two implementations of word2vec: Continuous Bag-Of-Words and Skip-Gram. Then, it analyses these two implementations' analogy after training with different sizes of corpus and context.

2. word2vec

Word2vec represents words in a vector space. To do this representation, word2vec can use Continuous Bag-Of-Words Model or Skip-Gram Model. The Continuous Bag-Of-Words uses continuous distributes representation of the context to predict the current word. Thus, as standard bag-of-words model, the other of words in context doesn't change the prediction [Mikolov et al. 2013]. On the other hand, in Skip-Gram, the context is the one to be predicted given the current word. The Skip-Gram model is capable get better results increasing the range of the context, but it is related to more computational cost [Mikolov et al. 2013].

3. Trained models

There were used three Wikipedia corpus with different sizes to train both CBOW and Skip-Gram language models, with 3 different window sizes (context). The word2vec code was also provided by Google.

4. Result and analysis

To do so

decrease decreasing fly flying flies 0.03

Table 1: CBOW results

Vocab. Size	Window	Best Distance	Worst Distance	Average Distance	Accuracy
39071	4	0.0	0.67	0.1	0.25
39071	8	0.0	0.77	0.1	0.27
39071	16	0.0	0.72	0.11	0.28
57014	4	0.0	0.68	0.07	0.41
57014	8	0.0	0.6	0.08	0.43
57014	16	0.0	0.65	0.09	0.42
71291	4	0.0	0.63	0.07	0.5
71291	8	0.0	0.65	0.07	0.54
71291	16	0.0	0.72	0.08	0.51

Table 2: Skip-Gram results

Vocab. Size	Window	Best Distance	Worst Distance	Average Distance	Accuracy
39071	4	0.0	0.5	0.1	0.2
39071	8	0.0	0.52	0.1	0.27
39071	16	0.0	0.63	0.11	0.31
57014	4	0.0	0.47	0.08	0.36
57014	8	0.0	0.49	0.08	0.43
57014	16	0.0	0.49	0.09	0.43
71291	4	0.0	0.51	0.07	0.48
71291	8	0.0	0.47	0.07	0.52
71291	16	0.0	0.51	0.08	0.5

5. Conclusion

References

Cambria, E. and White, B. (2014). Jumping NLP curves: A review of natural language processing research [review article]. *IEEE Comp. Int. Mag.*, 9(2):48–57.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

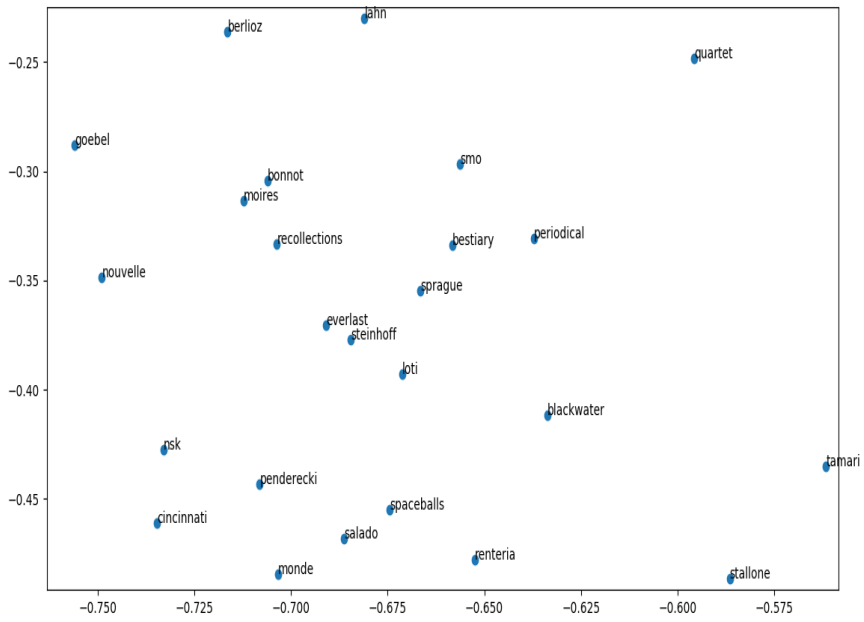

Figure 1: Word embeddings from Skip-Gram model with maximum vocabulary size and window size 8