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 appearing of social networks, a massive amount of data is constantly being generated weekly 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 (vector) 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 [?]. 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 [?].

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

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.

Table 1: Some random CBOW results

Corpus	Window	Input	Expected	Predicted	Distance
Small	4	decrease decreasing fly	flying	flies	0.03
Small	4	thinking thought sitting	sat	said	0.23
Small	4	slow slower hard	harder	harder	0.0
Small	4	young youngest bad	worst	biggest	0.06
Small	4	efficient efficiently quiet	quietly	impulsive	0.23
Small	8	sing singing see	seeing	topics	0.44
Small	8	falling fell walking	walked	divers	0.1
Small	8	responsible irresponsible ethical	unethical	irrational	0.19
Small	8	efficient inefficient responsible	irresponsible	alleged	0.4
Small	8	eagle eagles building	buildings	buildings	0.0
Small	16	vanish vanishes find	finds	get	0.11
Small	16	slowing slowed describing	described	praising	0.13
Small	16	sudden suddenly obvious	obviously	asks	0.12
Small	16	brothers sisters dad	mom	scrooge	0.09
Small	16	running ran enhancing	enhanced	instigated	0.16
Medium	4	enhancing enhanced knowing	knew	know	0.12
Medium	4	dancing danced singing	sang	accompaniment	0.08
Medium	4	enhance enhancing read	reading	aloud	0.26
Medium	4	running ran implementing	implemented	cooperating	0.03
Medium	4	write writes search	searches	searching	0.15
Medium	8	dog dogs eye	eyes	retina	0.09
Medium	8	man men cow	cows	tipping	0.07
Medium	8	easy easiest quick	quickest	toss	0.09
Medium	8	generating generated taking	took	took	0.0
Medium	8	certain uncertain convenient	inconvenient	secure	0.19
Medium	16	banana bananas color	colors	cmyk	0.03
Medium	16	seeing saw knowing	knew	dumb	0.140.5ex] heig
Medium	16	reading read decreasing	decreased	decrease	0.03
Medium	16	slow slowing say	saying	think	0.1
Medium	16	jumping jumped swimming	swam	stables	0.27