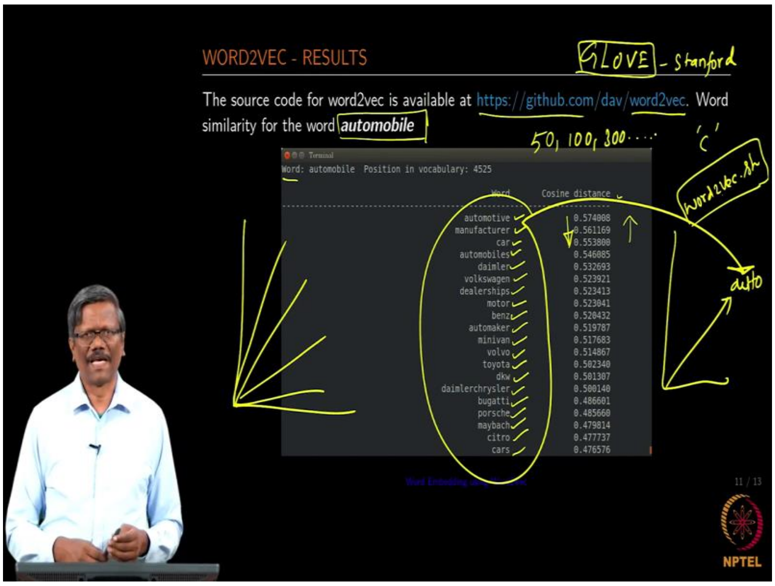
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute

Lecture – 50
Discussion on the results obtained from word2vec

(Refer Slide Time: 00:14)



WORD2VEC - RESULTS

The source code for word2vec is available at https://github.com/dav/word2vec. Word similarity for the word ***automobile***

```
Word: automobile  Position in vocabulary: 4525

                                              Word       Cosine distance
------------------------------------------------------------------------
                                        automotive              0.574008
                                      manufacturer              0.561169
                                               car              0.553800
                                       automobiles              0.546085
                                           daimler              0.532693
                                        volkswagen              0.523921
                                       dealerships              0.523413
                                             motor              0.523041
                                             benz               0.520432
                                         automaker              0.519787
                                           minivan              0.517683
                                             volvo              0.514867
                                            toyota              0.502340
                                               dkw              0.501307
                                   daimlerchrysler              0.500140
                                           bugatti              0.486601
                                           porsche              0.485660
                                           maybach              0.479814
                                             citro              0.477737
                                              cars              0.476576
```

And, then what Google also had done is it is a good thing that they have done, they have developed a sample program and implemented the application in C. There is one person by name McCormick who has commented each and every line of what Google has presented and that would also help you if you do not understand C very well, it will help you understand the concepts. And then how they really optimize the weights, how they use this sub-sampling mechanism, negative sampling, and also a hierarchical model in the softmax values.

So, I like you to go to this website, download this, you do not have to really do anything if you have a C compiler in your desktop you just have to run the Make File and it will do every job for you. It will download about 1 billion words and then train the neural network for those words and it is very simple to use. So, once you finish it you will have executable, there are scripts available for you to run so, both for windows MAC environment, as well as Linux environment.

And then try to run some small example there is a word2vec dot sh if you are using a Linux box or MAC or I think it is a bat file in the windows. So, you will find an executable script in this name and when you run the script it will ask you to input some values ok. So, it will ask you to give me the word ok. So, when you provide the word here what I have done is I have given automobile as the word and then it tries to find out all the words that are somehow related to this word.

How does it find I think that is entire exercise is all about that right. The word is identified by the company it keeps, remember that quote from Firth ok. So, that means, this word has a context which is available across the entire 1 billion words that it downloaded. Based on the context the relationships are established and the words are found closer to each other using a cosine distance.

So, in this case, I think I am not sure about the vocabulary size, it is pretty huge though ok. So, when you do the cosine distance you have to find the cosine distance for the entire one how many hour vocabularies that you have and then list them in the order ok. So, here we have the descending order of importance. So, you look at the words that it had picked up automotive ok, then manufacturer. So many places it would have encountered the context related to the automobile consisting of the manufacturer, the name of the manufacturer all that right and then it found that a car is related to this, the plural is associated here, the name of the company is also associated with this and then dealerships.

So, this word again is associated with you are able to appreciate the word vector model right. So, it is not just about the meaning of the word alone, it is about how it really works along with the other words in the entire corpus. And then motor, Benz, automaker, minivan, Volvo, Toyota, dkw, DaimlerChrysler, Bugatti, Porsche, Maybach, and cars. So, how wonderful it is right.

So, without really using any dictionary that is available only based on the data that is provided to these systems, it is able to really relate a given word with the other word just based on the context that it formed right and this context is what we can call as a pattern that is available and the system has identified the patterns which are close to the word automobile here. So, you remember when we spoke about the dimensionality reduction this exactly it is. Now, I can place all of these words in one dimension. So, all of these could be in this dimension. So, instead of keeping them in so many different dimensionalities right we have reduced the dimensionality using this model.

So, please read this. This is the core of the machine learning part in natural language processing and many many applications are developed based on this alright. There is another important implementation that is available called Glove. So, this is from Stanford. So, you can also take a look at this and then see how this is implemented ok. Then maybe sometime later I will also talk about what is the difference between the word2vec and the Glove model and again you know if you go to the websites of glove they have provided a lot of things for you as well.

So, you can also instead of creating your word vector using the corpus which is a very time-consuming process on a small desktop, you can download the word vectors as well. So, you know word vectors are available with 50 elements, 100 elements or 300 elements I am not sure what else you have there. You can use those word vectors also for your application directly.

So, the output of what we have done so far is going to be the word vectors for all the vocabulary that we have. So, that means, every word now has got in the relationship among all the words wherever possible and instead of creating that you can also use the readymade word vector that is available from Glove or from the word2vec and these are all available for free and you can use it in your applications.

See, one of the applications that I demonstrated briefly that for sentiment analysis I use the word vectors from Glove. This we will be using in the subsequent lectures from now onwards. The meaning of the word vectors all those things, now you have some ideas and to extend the ideas now we will take it to how do I really encode a sentence ok, how do I encode a paragraph, how do I identify important phrases in a given paragraph, how do I use the sentences that I have encoded so that I would be able to translate from one language to another.

So, all involve the basics of word vectors try to read as much as possible in this part. This is the fundamentals of modern natural language processing.