## HU Extension Assignment 12 E63 Big Data Analytics

### Handed out: 04/22/2017 Due by 11:59AM EST on Sunday, 04/29/2017

All my comments and code are in blue color

**Problem 1.** Install and compile Word2Vec C executables. Train CBOW model and create 200 dimensional embedding of Word Vectors. Demonstrate that you could run analogical reasoning when searching for country’s favorite food starting with Japan and sushi. Find favorite food for 5 different countries.

Firstly download Google’s word2vec source code and compile it

|  |
| --- |
| https://storage.googleapis.com/google-code-archive-source/v2/code.google.com/word2vec/source-archive.zip  sudo apt-get unzip  unzip source-archive.zip  make word2vec |

Then run the required script:

|  |
| --- |
| chmod 777 demo-word.sh  sudo ./demo-word.sh |

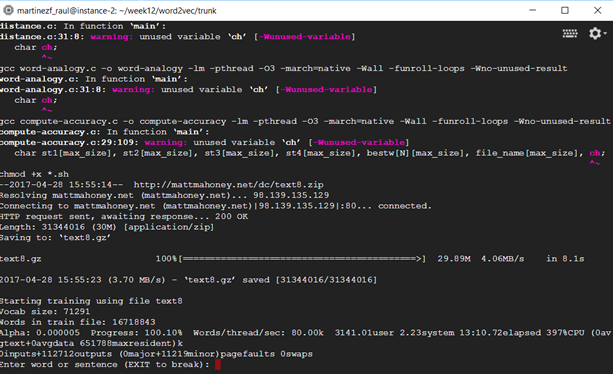


Figure 1 - Google's word2vec execution

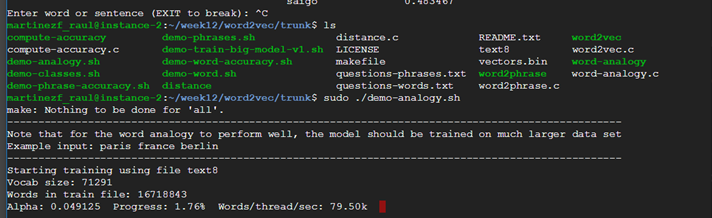


Figure 2 - Word2vec training process execution

According to problem’s requirements the following 5 a is to “b as c is to…” as executed, using “japan sushi word”

***japan sushi spain***

leche 0.488866

dulce 0.485657

salsa 0.484177

filo 0.473820

empanadas 0.459266

First one for Spanish, note that 4 out of 5 words are in spanish. First one is Spanish word for milk. We drink milk at Spain but the are some other more popular meals. However it is worth mentioning all words are related to food.

On the other hand, when executing the second sentence, two first occurences are not meals, but last names

***japan sushi portugal***

henriques 0.465154

bartolomeu 0.461889

filo 0.442283

faro 0.424407

moniz 0.419019

For third try:

***japan sushi france***

we get:

steak 0.463837

filo 0.462571

azur 0.460133

shallots 0.456433

sausage 0.455278

sauces 0.454213

Those are meals as well but not the most popular ones, it’s true that sauces are very common in French cuisine.

Fourth string:

***japan sushi uk***

cake 0.463504

jam 0.442010

cheesy 0.426536

gravy 0.422892

topping 0.414728

breakfast 0.410494

In this case result represent probably most popular meals in at least England, more than UK, including also a meal time like breakfast that it’s definitely very popular there

Last one is China

***japan sushi china***

choy 0.527990

stew 0.506387

garnish 0.499344

wonton 0.491529

pickled 0.486861

First result, choy is fine, since bok choy is very popular in China, or at least in Asiatic food, same for wonton.

Report imporobable results as well as good results. Use scripts provided with original Google C code. (**50%)**

**Problem 2.** Install and run genism Python Word2Vec API. Find out what are the most probable words you will obtain when you start with an emperor add a woman and subtract a man. Use this tutorial as a guide <https://rare-technologies.com/word2vec-tutorial/>

**(50%)**

First install, genism python library

|  |
| --- |
| Sudo pip install --upgrade genism |

Then create vector model from the first problem’s corpus

|  |
| --- |
| # import modules & set up logging  import gensim, logging  logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)  #load model, model was created in problem 1  model = gensim.models.KeyedVectors.load\_word2vec\_format('./word2vec/trunk/vectors.bin', binary=True)  #what requested in the problem  model.most\_similar(positive=['emperor', 'woman'], negative=['man'], topn=1) |

At the result is hundred percent correct, since we get empress:

|  |
| --- |
| [(u'empress', 0.5768526792526245)] |

Please, describe every step of your work and present all intermediate and final results in a Word document. Please, copy past text version of all essential command and snippets of results into the Word document with explanations of the purpose of those commands. We cannot retype text that is in JPG images. Please, always submit a separate copy of the original, working scripts and/or class files you used. Sometimes we need to run your code and retyping is too costly. Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE DO NOT EMBED files into your MS Word document. For issues and comments visit the class Discussion Board.

If you are working with Jupyter Notebooks please provide clear and full comments for all of important steps or changes you are making. Please provide the notebook itself (ipynb file) and the PDF version of the file. Canvas cannot read ipynb file and if you do not provide an MS Doc or PDF version of your work, you will be penalized.

It is not acceptable that you describe your solution of any of these problems on Piazza.