## HU Extension Assignment 08 E63 Big Data Analytics

### Handed out: 10/20/2017 Due by 4:00 PM EST on Saturday, 10/28/2017

If you are familiar with NLP API-s in languages other than Python or Python NLP API-s other than NLTK please be free to solve these problems using technology of your choice.

**Problem 1. Use the text of the Universal Declaration of Human Rights (UDHR). Create a table for 5 languages in which you will collect statistics about the languages used. Place in that table the number of words in each language in UDHR, number of unique words, average length of words, number of sentences contained in UDHR and average number of words per sentence. Create a distribution of sentence lengths for each language. Plot those (non-cumulative) distributions on one diagram.**

**(25%)**

See p1.ipyb for a closer look

**Problem 2. Identify 10 most frequently used words longer than 7 characters in the entire corpus of Inaugural Addresses. Do not identify 10 words for every speech but rather 10 words for the entire corpus. Which among those words has the largest number of synonyms? List all synonyms for those 10 words. Which one of those 10 words has the largest number of hyponyms? List all hyponyms of those 10 most frequently used “long” words. The purpose of this problem is to familiarize you with WordNet and concepts of synonyms and hyponyms.**

**(25%)**

**Your literature for Problems 1 and 2 are chapters 1 and 2 of Natural Language Processing with Python book by Steven Bird et al.**

See p2.ipyb for a closer look

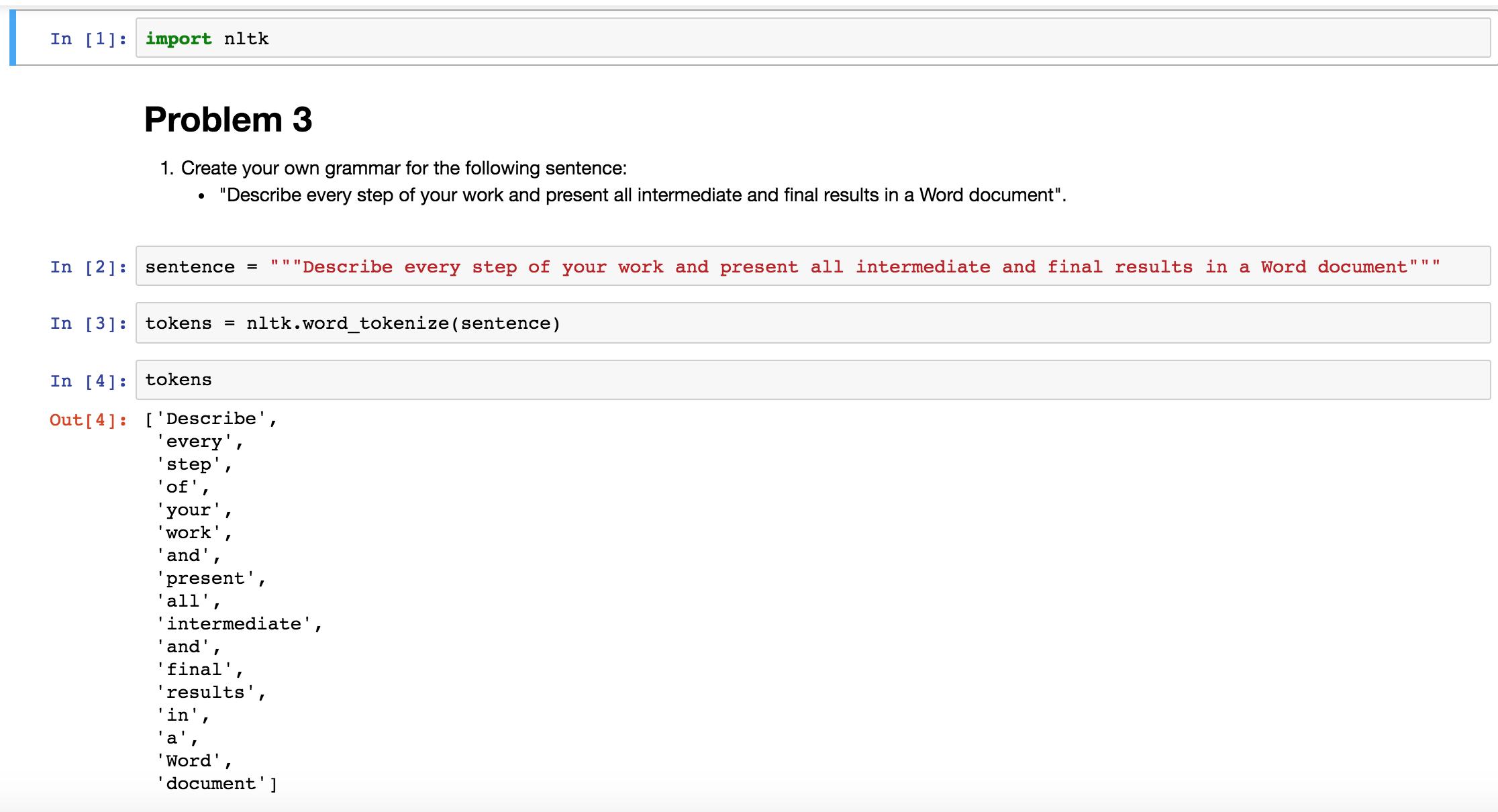
**Problem 3. Create your own grammar for the following sentence:**

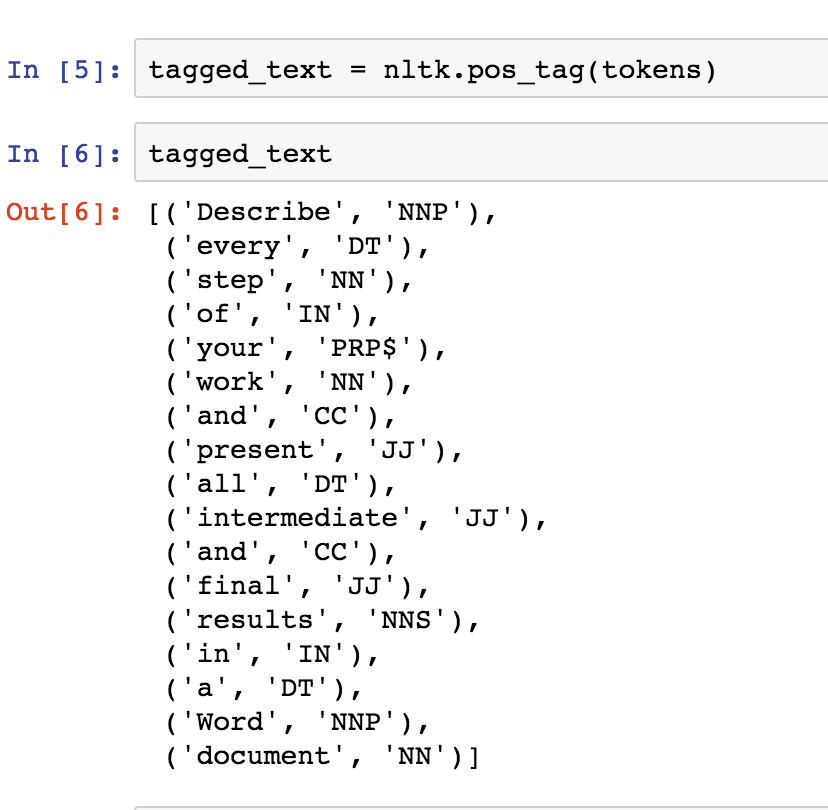
**“Describe every step of your work and present all intermediate and final results in a Word document”.**

**(10%)**

**Your literature for Problem 3 is chapter 8 of Natural Language Processing with Python book by Steven Bird et al.**

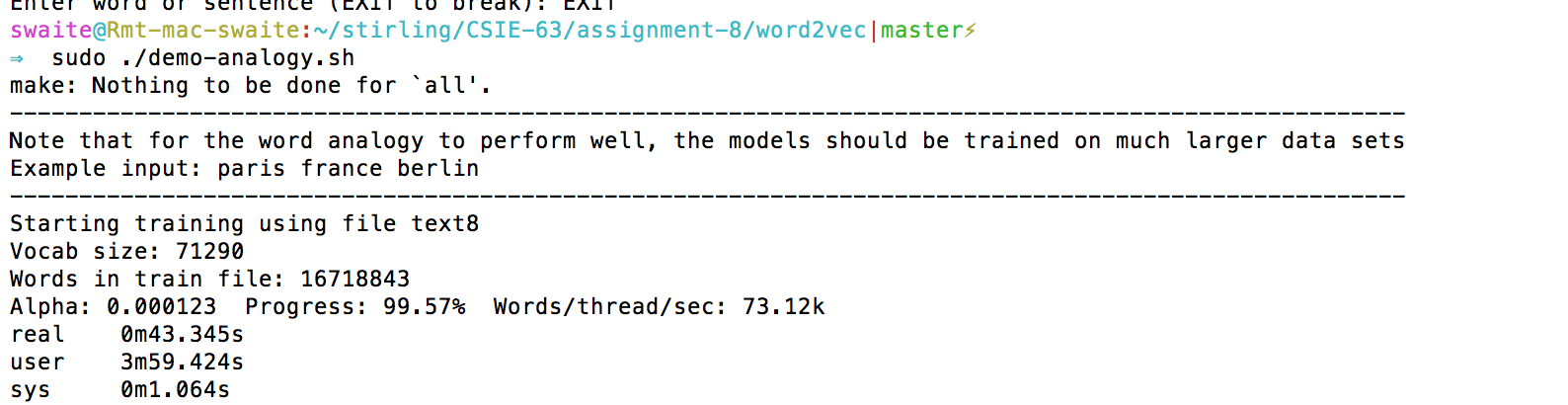
See p3.ipyb for a closer look





**Problem 4. 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. Note that words might have to be in lower case. Find favorite food for 5 different countries. Report improbable results as well as good results. Use scripts provided with original Google C code.**

**(20%)**

****

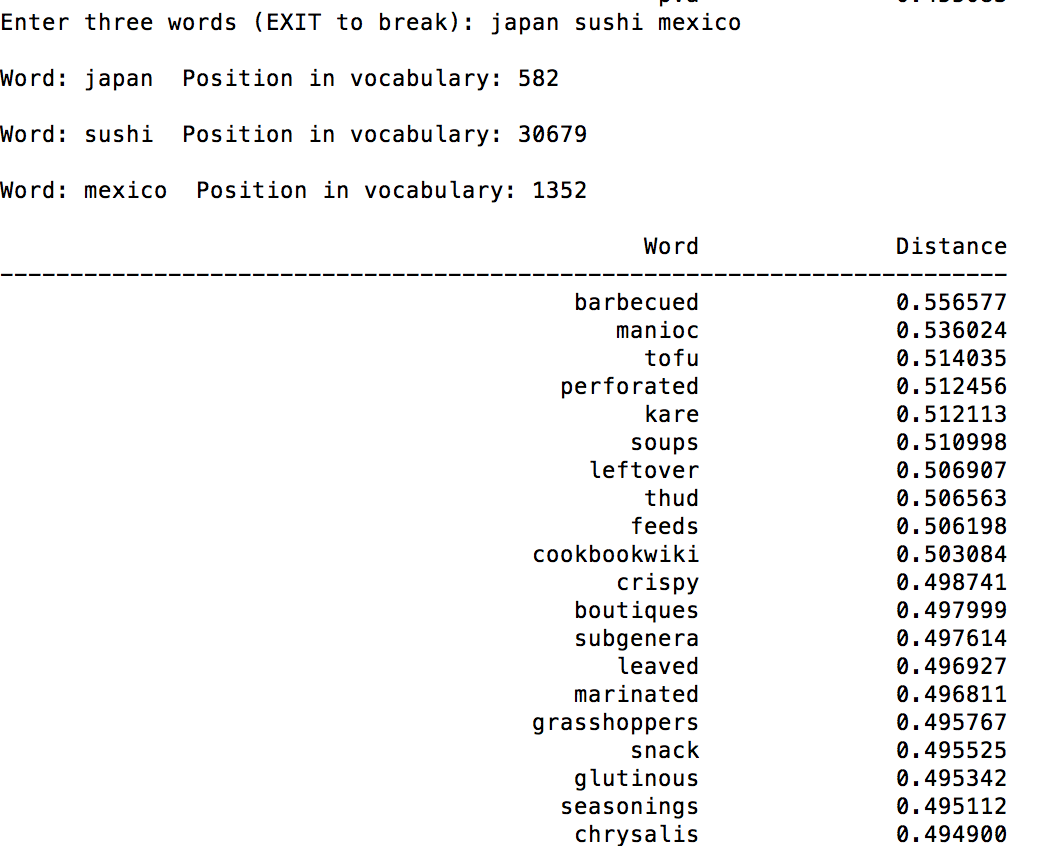
**japan sushi thailand**

|  |
| --- |
| Enter three words (EXIT to break): **japan sushi thailand**  Word: japan Position in vocabulary: 582  Word: sushi Position in vocabulary: 30679  Word: thailand Position in vocabulary: 5640  Word Distance  crab 0.561080  bento 0.559273  kimchi 0.541000  mochi 0.540748  crepe 0.536611 |

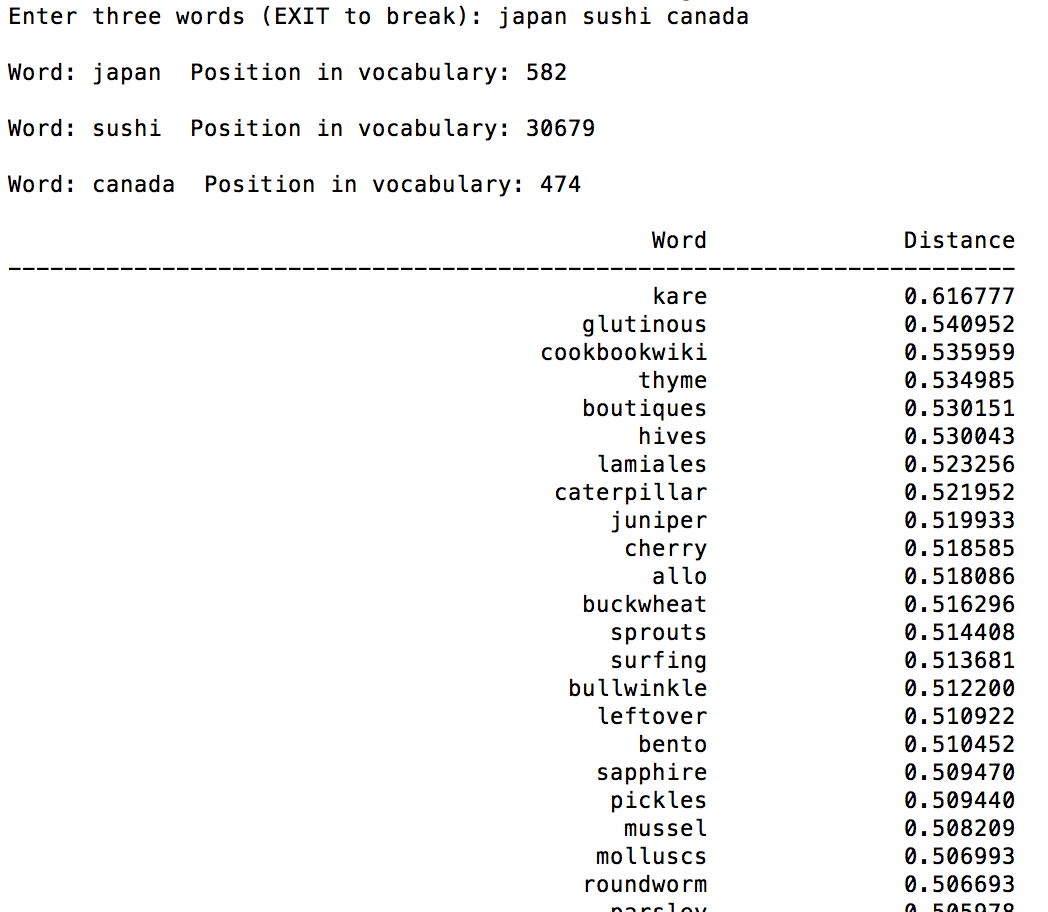
**japan sushi italy**

|  |
| --- |
| Enter three words (EXIT to break): japan sushi italy  Word: japan Position in vocabulary: 582  Word: sushi Position in vocabulary: 30679  Word: italy Position in vocabulary: 843  Word Distance  -----------------------------------------------------------------  strawberries 0.500201  kelp 0.500120  sprouts 0.495927  omelette 0.493959  mussel 0.491622  pies 0.489951  cranberries 0.474655 |

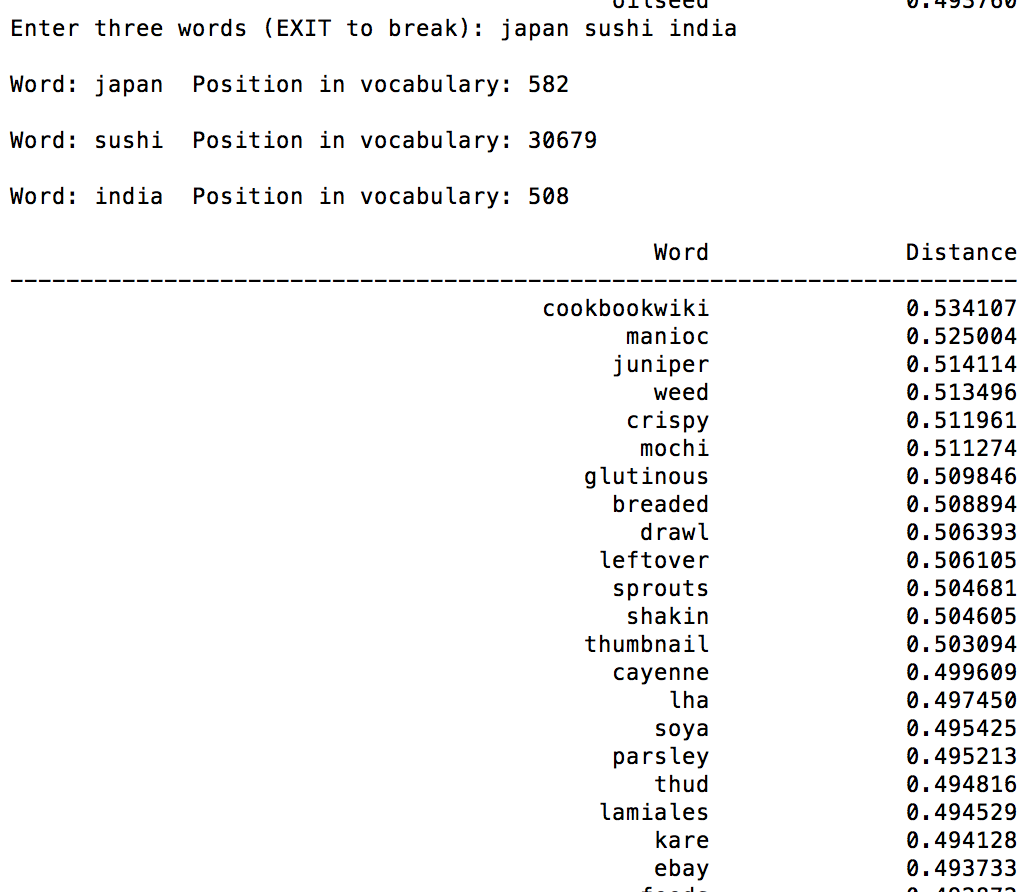
japan sushi mexico



**japan sushi canada**

****

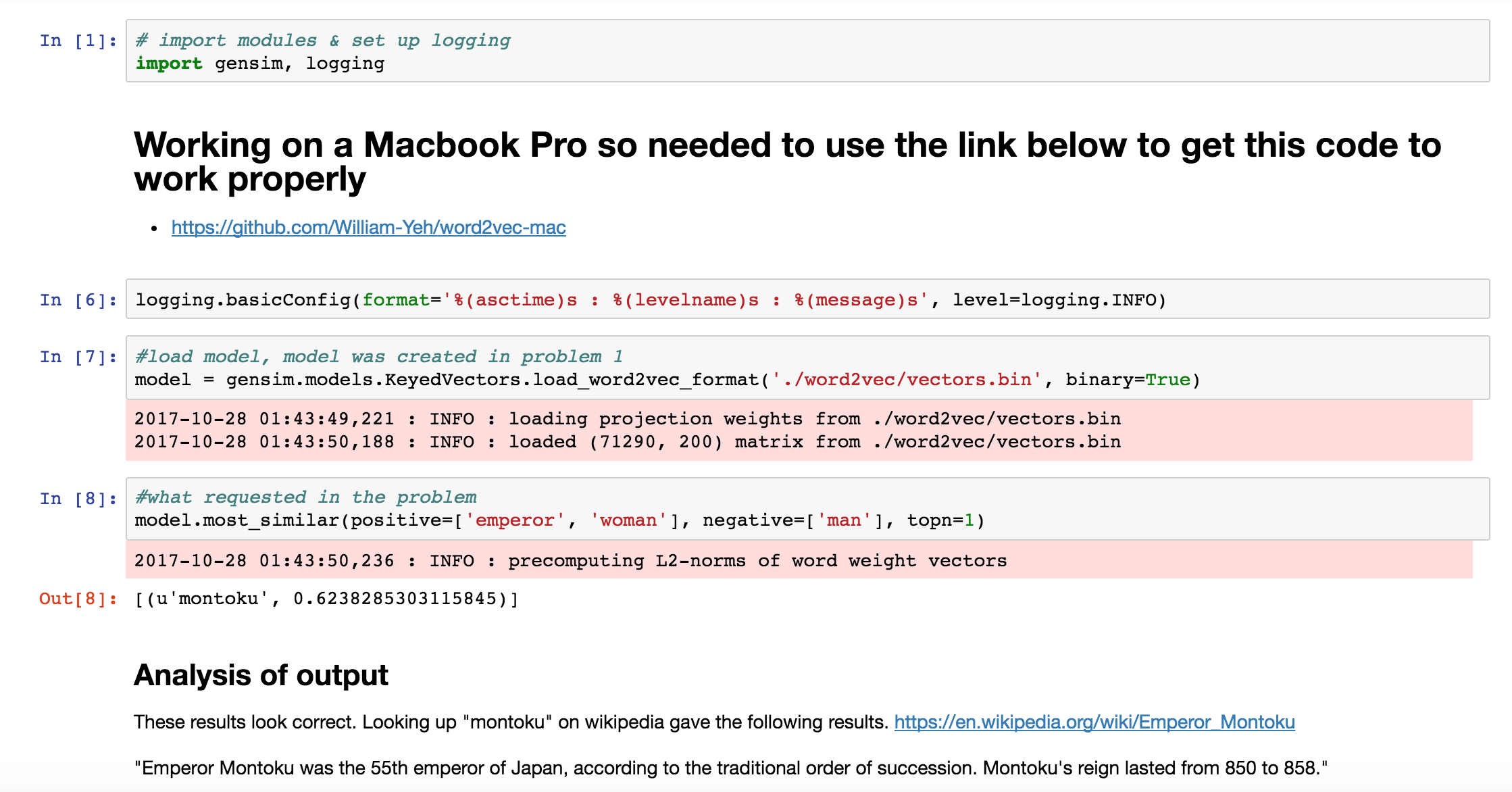
**japan sushi india**

****

**Problem 5. Install and run Genism Python Word2Vec API. Find 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/**](https://rare-technologies.com/word2vec-tutorial/)

**(20%)**

See p5.ipyb for a closer look



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 use some other language other than Python in your daily work with NLP, please be free to use that language and a framework of your choice to do this assignment.