# CSE256_Assignment2

April 19, 2022

# 1 CSE 256: Statistical NLP UCSD Assignment 2

## 1.1 Exploring Word Vectors (12.5 points + 1 bonus point)

## 1.2 Vignesh Nanda Kumar PID: A59010704

### 1.2.1 Due 11:59pm, Monday April 18, 2022

Before you start, make sure you read the README.txt in the same directory as this notebook.

**Notes:** Please make sure to save the notebook as you go along. Submission Instructions are located at the bottom of the notebook.

[1]:

```
# All Import Statements Defined Here
# Note: Do not add to this list.
# ----------------

import sys
assert sys.version_info[0]==3
assert sys.version_info[1] >= 5

from gensim.models import KeyedVectors
from gensim.test.utils import datapath
import pprint
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [10, 5]
import nltk
nltk.download('reuters')
from nltk.corpus import reuters
import numpy as np
import random
import scipy as sp
from sklearn.decomposition import TruncatedSVD
from sklearn.decomposition import PCA

START_TOKEN = '<START>'
END_TOKEN = '<END>'

np.random.seed(0)
random.seed(0)
```

```
# ----------------
```

```
[nltk_data] Downloading package reuters to /Users/vignesh/nltk_data...
[nltk_data]   Package reuters is already up-to-date!
```

## 1.3 Word Vectors

Word Vectors are often used as a fundamental component for downstream NLP tasks, e.g. question answering, text generation, translation, etc., so it is important to build some intuitions as to their strengths and weaknesses. Here, you will explore word vectors derived from *Word2Vec*.

**Note on Terminology:** The terms “word vectors” and “word embeddings” are often used interchangeably. The term “embedding” refers to the fact that we are encoding aspects of a word’s meaning in a lower dimensional space. As Wikipedia states, “*conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with a much lower dimension*”.

## 1.4 Word Vectors

We shall explore the embeddings produced by word2vec. Please revisit the class notes and lecture slides for more details on the word2vec algorithm. Paper 1 review due May 4th, involves reading the word2vec paper, reading it now might help you with this assignment.

Run the following cells to load the word2vec vectors into memory. **Note**: If this is your first time to run these cells, i.e. download the embedding model, it will take a couple minutes to run. If you’ve run these cells before, rerunning them will load the model without redownloading it, which will take about 1 to 2 minutes. In *Colab*, the embeddings are downloaded to the server everytime you restart the notebook). For this reason, you may prefer to work on your local machine where the download only happens once.

[2]:

```
def load_embedding_model():
    """ Load Word2Vec Vectors
        Return:
            wv_from_bin: All the embeddings
    """
    import gensim.downloader as api
    wv_from_bin = api.load("word2vec-google-news-300")

    print("Loaded vocab size %i" % len(wv_from_bin.vocab.keys()))
    return wv_from_bin
```

[3]:

```
# -----------------------------------
# Run Cell to Load Word Vectors
# Note: This will take a couple minutes
# -----------------------------------
wv_from_bin = load_embedding_model()
```

```
Loaded vocab size 3000000
```

**Note: If you are receiving a “reset by peer” error, rerun the cell to restart the download.**

### 1.4.1 Plot function

Let’s define a plot function that reduces the vectors from 300-dimensions to 2-dimensions, and visualises them.

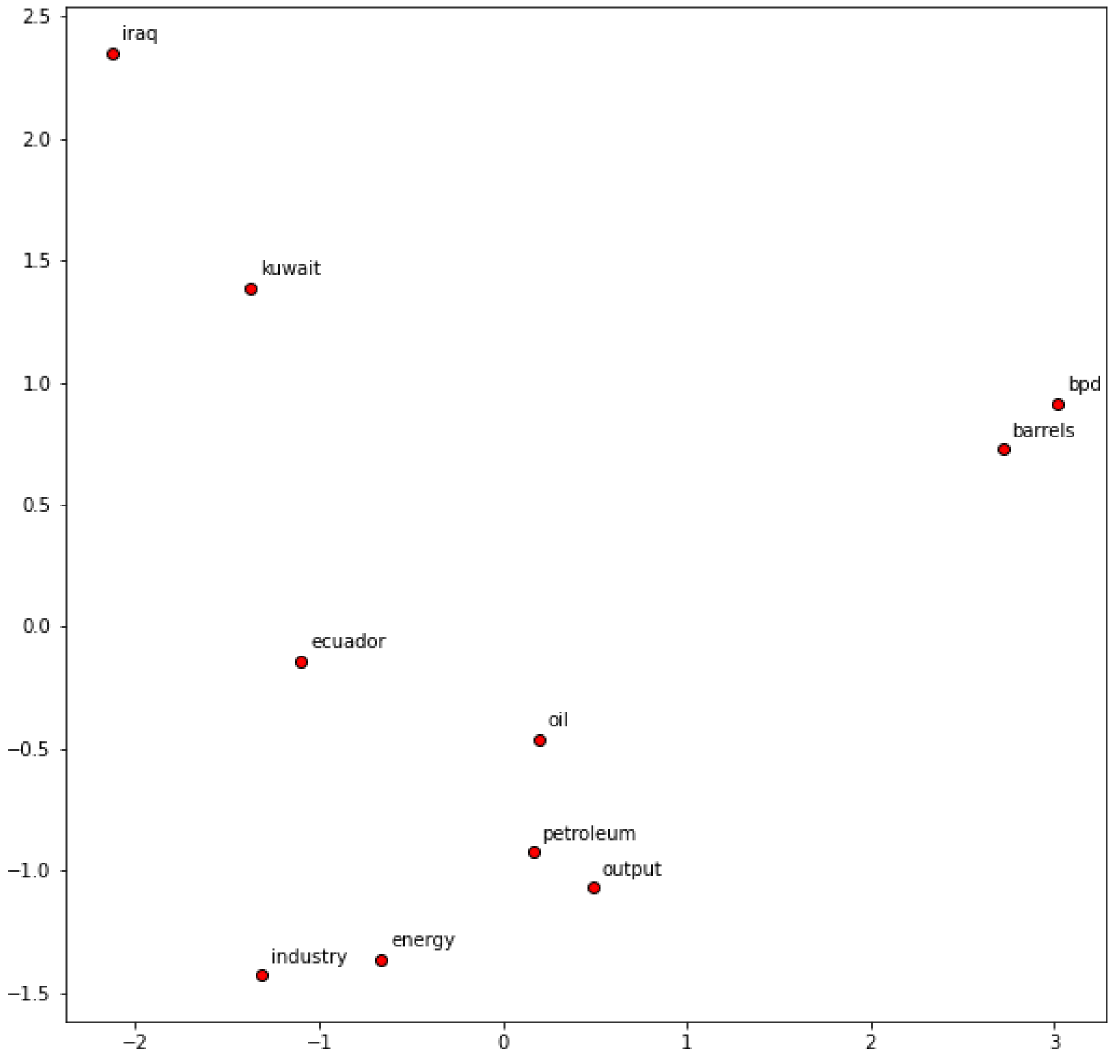
[4]:
```
def display_pca_scatterplot(model, words=None, sample=0):
    if words == None:
        if sample > 0:
            words = np.random.choice(list(model.key_to_index.keys()), sample)
        else:
            words = [ word for word in model.vocab ]

    word_vectors = np.array([model[w] for w in words])

    twodim = PCA().fit_transform(word_vectors)[:,:2]

    plt.figure(figsize=(10,10))
    plt.scatter(twodim[:,0], twodim[:,1], edgecolors='k', c='r')
    for word, (x,y) in zip(words, twodim):
        plt.text(x+0.05, y+0.05, word)
```

### 1.4.2 Question 1: Word2vec Plot Analysis [written] (2 points)

Run the cell below to plot the 2D GloVe embeddings for `['barrels', 'bpd', 'ecuador', 'energy', 'industry', 'kuwait', 'oil', 'output', 'petroleum', 'iraq']`.

What clusters together in 2-dimensional embedding space? What doesn’t cluster together that you think should have?

[5]:
```
words = ['barrels', 'bpd', 'ecuador', 'energy', 'industry', 'kuwait', 'oil',␣
 ↪'output', 'petroleum', 'iraq']
display_pca_scatterplot(wv_from_bin, words)
```



**Oil, petroleum, output is one cluster, energy and industry is another cluster relating to the topic of fuel industry. Barrels, bpd also cluster together. Kuwait, Equador, Iraq doesn't cluster together but should've because all of them are countries. Also Iraq and Kuwait could've been more closer to oil given that they are major oil producing nations so there could be sentences which have these countries occuring in context of oil. Also, barrels and bpd (barrels per day could've been closer to petroleum since they also relate to the petroleum industry. But it could be possible that barrels was used in a context other than the petroleum industry and hence is not closer to petroleum.**

### 1.4.3 Cosine Similarity

Now that we have word vectors, we need a way to quantify the similarity between individual words, according to these vectors. One such metric is cosine-similarity. We will be using this to find words that are "close" and "far" from one another.

The Cosine Similarity $s$ between two vectors $p$ and $q$ is defined as:

$$s = \frac{p \cdot q}{||p||||q||}, \text{ where } s \in [-1, 1]$$

### 1.4.4 Question 2: Words with Multiple Meanings (2 points) [code + written]

Polysemes and homonyms are words that have more than one meaning (see this wiki page to learn more about the difference between polysemes and homonyms ). Find a word with *at least two different meanings* such that the top-10 most similar words (according to cosine similarity) contain related words from *both* meanings. For example, "leaves" has both "go_away" and "a_structure_of_a_plant" meaning in the top 10, and "rock" has both "music" and "stone". You will probably need to try several polysemous or homonymic words before you find one.

Please state the word you discover and the multiple meanings that occur in the top 10. Why do you think many of the polysemous or homonymic words you tried didn't work (i.e. the top-10 most similar words only contain **one** of the meanings of the words)?

**Note**: You should use the `wv_from_bin.most_similar(word)` function to get the top 10 similar words. This function ranks all other words in the vocabulary with respect to their cosine similarity to the given word. For further assistance, please check the **GenSim documentation**.

[6]:
```
# ------------------
# Write your implementation here.
word = "bass"
wv_from_bin.most_similar(word)

# ------------------
```

[6]:
```
[('crappie', 0.6551623344421387),
 ('largemouth', 0.6519780158996582),
 ('largemouths', 0.6363433599472046),
 ('largemouth_bass', 0.6293675899505615),
 ('striper', 0.6191805601119995),
 ('stripers', 0.6170703768730164),
 ('smallmouth', 0.6161339282989502),
 ('Spotted_bass', 0.6154202222824097),
 ('acoustic_bass', 0.613075852394104),
 ('upright_bass', 0.6106366515159607)]
```

**The word is bass which is a type of fish and also a type of singing voice. The top 10 related words contains words from both meanings: fish (largemouth_bass, smallmouth are types of bass fish), and acoustic bass (related to music and also a type of guitar). Many of the words didn't work because it could be possible that the corpus on which the model was trained did not contain the word used context of its second meaning. Or the word occurred in context of its first meaning much more and in context of its second meaning much less (so the first meaning related words are closer in cosine distance and come in top 10 of the word, while the second meaning related words would be far off)**

### 1.4.5 Question 3: Analogies with Word Vectors [written] (2 points)

Word vectors have been shown to *sometimes* exhibit the ability to solve analogies.

As an example, for the analogy “man : king :: woman : x” (read: man is to king as woman is to x), what is x?

In the cell below, we show you how to use word vectors to find x using the `most_similar` function from the **GenSim documentation**. The function finds words that are most similar to the words in the `positive` list and most dissimilar from the words in the `negative` list (while omitting the input words, which are often the most similar; see this paper). The answer to the analogy will have the highest cosine similarity (largest returned numerical value).

[7]:

```
# Run this cell to answer the analogy -- man : king :: woman : x
pprint.pprint(wv_from_bin.most_similar(positive=['woman', 'king'],␣
 ↪negative=['man']))
```

```
[('queen', 0.7118192911148071),
 ('monarch', 0.6189674139022827),
 ('princess', 0.5902431011199951),
 ('crown_prince', 0.5499460697174072),
 ('prince', 0.5377321243286133),
 ('kings', 0.5236844420433044),
 ('Queen_Consort', 0.5235945582389832),
 ('queens', 0.518113374710083),
 ('sultan', 0.5098593235015869),
 ('monarchy', 0.5087411999702454)]
```

Let $m$, $k$, $w$, and $x$ denote the word vectors for `man`, `king`, `woman`, and the answer, respectively. Using **only** vectors $m$, $k$, $w$, and the vector arithmetic operators $+$ and $-$ in your answer, what is the expression in which we are maximizing cosine similarity with $x$?

**x = (k - m + w)**

### 1.4.6 Question 4: Finding Analogies [code + written] (1 point)

Find an example of analogy that holds according to these vectors (i.e. the intended word is ranked top). In your solution please state the full analogy in the form x:y :: a:b. If you believe the analogy you came up might not be obvious to the TAs, explain why the analogy holds in one or two sentences.

**Note**: You may have to try many analogies to find one that works!

[8]:

```
    # ------------------
    # Write your implementation here.
    pprint.pprint(wv_from_bin.most_similar(positive=['Japan', 'Paris'],␣
 ↪negative=['France']))


    # ------------------
```

```
[('Tokyo', 0.8142861127853394),
 ('Toyko', 0.659669816493988),
 ('Osaka', 0.6350962519645691),
 ('Nagoya', 0.6258591413497925),
 ('Seoul', 0.6054927706718445),
 ('Japanese', 0.5919331312179565),
 ('Yokohama', 0.5900902152061462),
 ('Osaka_Japan', 0.585975170135498),
 ('Takamatsu', 0.57918381690979),
 ('Fukuoka', 0.5664029121398926)]
```

**France:Paris :: Japan:Tokyo . The analogy is of country : capital**

### 1.4.7 Question 5: Incorrect Analogy [code + written] (2 point)

Find an example of analogy that does *not* hold according to these vectors. In your solution, state the intended analogy in the form x:y :: a:b, and state the (incorrect) value of b according to the word vectors.

```
[9]:    # ------------------
        # Write your implementation here.
        pprint.pprint(wv_from_bin.most_similar(positive=['car', 'sea'],␣
      ↪negative=['ship']))

        # ------------------
```

```
[('cars', 0.4880220890045166),
 ('Car', 0.46311694383621216),
 ('Mazda_MX5', 0.46152472496032715),
 ('SUV', 0.4557592272758484),
 ('Volkswagon_Polo', 0.4556606709957123),
 ('Ford_Focus', 0.44932806491851807),
 ('Honda_Civic', 0.44522327184677124),
 ('motorbike', 0.443489134311676),
 ('vehicle', 0.44239693880081177),
 ('minivan', 0.43566223978996277)]
```

**ship:sea :: car:road is the intended analogy (vehicle:where the vehicle moves). The incorrect value is “cars”.**

### 1.4.8 Question 6: Guided Analysis of Bias in Word Vectors [written] (2 point)

It’s important to be cognizant of the biases (gender, race, sexual orientation etc.) implicit in our word embeddings. Bias can be dangerous because it can reinforce stereotypes through applications that employ these models.

Run the cell below, to examine (a) which terms are most similar to “woman” and “worker” and most dissimilar to “man”, and (b) which terms are most similar to “man” and “worker” and most dissimilar to “woman”. Point out the difference between the list of female-associated words and the list of male-associated words, and explain how it is reflecting gender bias.

```
[10]: # Run this cell
      # Here `positive` indicates the list of words to be similar to and `negative`␣
       ↪indicates the list of words to be
      # most dissimilar from.
      pprint.pprint(wv_from_bin.most_similar(positive=['woman', 'worker'],␣
       ↪negative=['man']))
      print()
      pprint.pprint(wv_from_bin.most_similar(positive=['man', 'worker'],␣
       ↪negative=['woman']))
```

```
[('workers', 0.6582455635070801),
 ('employee', 0.5805293321609497),
 ('nurse', 0.5249921679496765),
 ('receptionist', 0.5142490267753601),
 ('migrant_worker', 0.5001609325408936),
 ('Worker', 0.4979269802570343),
 ('housewife', 0.48609834909439087),
 ('registered_nurse', 0.4846190810203552),
 ('laborer', 0.48437267541885376),
 ('coworker', 0.48212406039237976)]

[('workers', 0.5590360164642334),
 ('laborer', 0.54481041431427),
 ('foreman', 0.5192232131958008),
 ('Worker', 0.5161596536636353),
 ('employee', 0.5094279050827026),
 ('electrician', 0.49481213092803955),
 ('janitor', 0.48718899488449097),
 ('bricklayer', 0.4825313091278076),
 ('carpenter', 0.47498998045921326),
 ('workman', 0.4642517566680908)]
```

**The female associated words are nurse, receptionist, housewife, coworker while the male associated words are foreman, electrician, janitor, carpenter which shows the extreme gender stereotype being brought out by the model. For example for female the related word is coworker but for male foreman is a related word. This is reflecting gender bias that only males can supervise and direct other workers. Also, occupations like nurse, receptionist are coming in female related words while electrician, carpenter are coming in male related words. These occupations should be gender neutral but instead the model is reflecting gender bias through these related words for man and woman.**

### 1.4.9 Question 7: Independent Analysis of Bias in Word Vectors [code + written] ( 1.5 point)

Use the `most_similar` function to find another case where some bias is exhibited by the vectors. Please briefly explain the example of bias that you discover.

[11]:

```
    # ------------------
    # Write your implementation here.
    # pprint.pprint(wv_from_bin.most_similar(positive=['raj', 'white'],␣
↪negative=['vincent']))
    pprint.pprint(wv_from_bin.most_similar(positive=['african', 'worker'],␣
↪negative=['american']))
    pprint.pprint(wv_from_bin.most_similar(positive=['american', 'worker'],␣
↪negative=['african']))
    # ------------------
```

```
[('workers', 0.555210292339325),
 ('laborer', 0.48098012804985046),
 ('migrant_worker', 0.45900607109069824),
 ('employee', 0.45884913206100464),
 ('Worker', 0.44189125299453735),
 ('housemaid', 0.42111337184906006),
 ('mineworker', 0.4171217978000641),
 ('Makhosi', 0.41370075941085815),
 ('Bogopane_Zulu', 0.41295868158340454),
 ('housekeeper', 0.41201385855674744)]
[('workers', 0.5497925877571106),
 ('employee', 0.5275042057037354),
 ('Worker', 0.4733121693134308),
 ('laborer', 0.4481498897075653),
 ('technician', 0.4288828372955322),
 ('janitor', 0.4153149724006653),
 ('forklift_operator', 0.40671056509017944),
 ('electrician', 0.406697541475296),
 ('supervisor', 0.40284863114356995),
 ('em_ployee', 0.39987462759017944)]
```

**The analogy brings out the racial bias or occupational bias present in word embeddings. For Africans, the first set has words like laborer, migrant worker, housemaid, mine worker, house keeper. For american the related words are employee, laborer, technician, janitor, operator, electrician, supervisor. Here the racial bias can be clearly seen between the occupation results for both the communities.**

### 1.4.10 Question 8: Thinking About Bias [written] (Bonus: 1 point)

Give one explanation of how bias gets into the word vectors. What is an experiment that you could do to test for or to measure this source of bias?

**Bias can get into the model from the training data. The word embeddings are generated from the given text data, so if the data is biased towards a particular gender or race or religion or any other type of bias, then the word embedding model will also be biased.(Discussed the experiment with another student)An experiment that can be done to measure the source of bias can be: first find out the different groups pertaining to a particular type of bias. For example lets say we take up different religions as groups. Now we can use the intrinsic evaluation technique of analogies to see the distance between 2 religion words (R1, R2) and a common word C. The distance between R1, C should be similar to the distance between R2 and C. The difference between the distances can be used to evaluate the bias present in the model.**

## 2 Submission Instructions

1. Click the Save button at the top of the Jupyter Notebook.
2. Select Cell -> All Output -> Clear. This will clear all the outputs from all cells (but will keep the content of all cells).
3. Select Cell -> Run All. This will run all the cells in order, and will take several minutes.
4. Once you've rerun everything, select File -> Download as -> PDF via LaTeX (If you have trouble using "PDF via LaTex", you can also save the webpage as pdf. Make sure all your solutions especially the coding parts are displayed in the pdf, it's okay if the provided codes get cut off because lines are not wrapped in code cells).
5. Look at the PDF file and make sure all your solutions are there, displayed correctly. The PDF is the only thing your graders will see!
6. Submit your PDF on Gradescope.

**Acknowledgements** This assignment is based on an assignment developed by Chris Manning

```
[ ]:
```