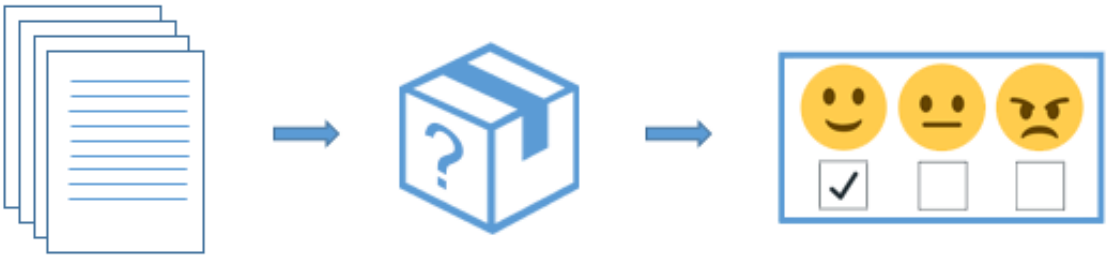
GitHub: https://github.com/MYUSER/MYPROJECT/ (https://github.com/MYUSER/MYPROJECT/)

Welcome to your assignment this week!

To better understand bias and discrimination in AI, in this assignment, we will look at a Natural Language Processing use case.

Natural Language Pocessing (NLP) is a branch of Artificial Intelligence (AI) that helps computers to understand, to interpret and to manipulate natural (i.e. human) language. Imagine NLP-powered machines as black boxes that are capable of understanding and evaluating the context of the input documents (i.e. collection of words), outputting meaningful results that depend on the task the machine is designed for.



Documents are fed into magic NLP model capable to get, for instance, the sentiment of the original content

Just like any other machine learning algorithm, biased data results in biased outcomes. And just like any other algorithm, results debiasing is painfully annoying, to the point that it might be simpler to unbias the society itself.

The big deal: word embeddings

Words must be represented as **numeric vectors** in order to be fed into machine learning algorithms. One of the most powerful (and popular) ways to do it is through **Word Embeddings**. In word embedding models, each word in a given language is assigned to a high-dimensional vector, such that **the geometry of the vectors captures relations between the words**.

Because word embeddings are very computionally expensive to train, most ML practitioners will load a pre-trained set of embeddings.

After this assignment you will be able to:

- Load pre-trained word vectors, and measure similarity using cosine similarity
- Use word embeddings to solve word analogy problems such as Man is to Woman as King is to __.
- Modify word embeddings to reduce their gender bias

Let's get started! Run the following cell to install all the packages you will need.

In [ ]:
```
#!pip install numpy
#!pip install keras
#!pip install tensorflow
```

Run the following cell to load the packages you will need.

In [4]:
```
import numpy as np
from w2v_utils import *
```

```
Using TensorFlow backend.
```

In [ ]:
```

```

Next, lets load the word vectors. For this assignment, we will use 50-dimensional GloVe vectors to represent words. Run the following cell to load the `word_to_vec_map` .

In [5]:
```
words, word_to_vec_map = read_glove_vecs('data/glove.6B.50d.txt')
```

You've loaded:

- `words` : set of words in the vocabulary.
- `word_to_vec_map` : dictionary mapping words to their GloVe vector representation.

In [6]:
```
print("Example of words: ",list(words)[:10])
print("Vector for word 'person' = ", word_to_vec_map.get('person'))
```

```
Example of words:  ['5-ht2a', 'implicitly', 'decriminalisation', '
saliers', '94.24', 'unending', 'euro237', 'renovators', '56-3', '8
.7']
Vector for word 'person' =  [ 0.61734    0.40035    0.067786  -0.3
4263    2.0647     0.60844
  0.32558    0.3869     0.36906    0.16553    0.0065053 -0.075674
  0.57099    0.17314    1.0142    -0.49581   -0.38152    0.49255
 -0.16737   -0.33948   -0.44405    0.77543    0.20935    0.6007
  0.86649   -1.8923    -0.37901   -0.28044    0.64214   -0.23549
  2.9358    -0.086004  -0.14327   -0.50161    0.25291   -0.065446
  0.60768    0.13984    0.018135  -0.34877    0.039985   0.07943
  0.39318    1.0562    -0.23624   -0.4194    -0.35332   -0.15234
  0.62158    0.79257  ]
```

GloVe vectors provide much more useful information about the meaning of individual words. Lets now see how you can use GloVe vectors to decide how similar two words are.

Cosine similarity

To measure how similar two words are, we need a way to measure the degree of similarity between two embedding vectors for the two words. Given two vectors u and v, cosine similarity is defined as follows:
\begin{equation} \text{CosineSimilarity(u, v)} = \frac {u . v} {||u||_2 ||v||_2} = cos(\theta) \tag{1} \end{equation}

where $u.v$ is the dot product (or inner product) of two vectors, $||u||_2$ is the norm (or length) of the vector u, and θ is the angle between u and v. This similarity depends on the angle between u and v. If u and v are very similar, their cosine similarity will be close to 1; if they are dissimilar, the cosine similarity will take a smaller value.

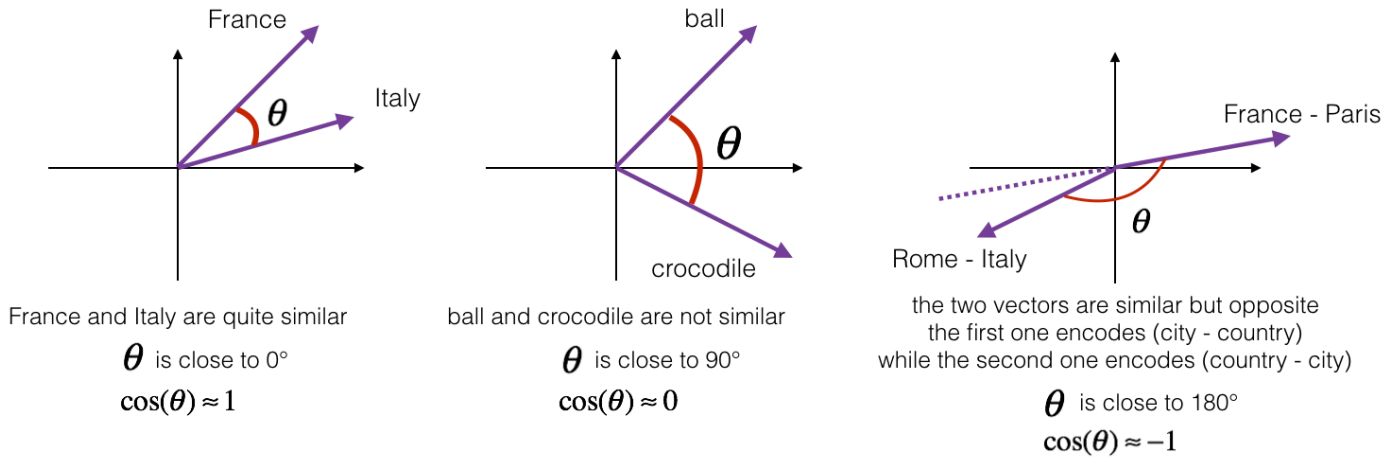


Figure 1: The cosine of the angle between two vectors is a measure of how similar they are

---

Task 1: Implement the function `cosine_similarity()` to evaluate similarity between word vectors.

Reminder: The norm of u is defined as $||u||_2 = \sqrt{\sum_{i=1}^{n} u_i^2}$

In [10]:
```
## Task 1
# cosine_similarity

def cosine_similarity(u, v):
    """
    Cosine similarity reflects the degree of similariy between u an
d v

    Arguments:
        u -- a word vector of shape (n,)
        v -- a word vector of shape (n,)

    Returns:
        cosine_similarity -- the cosine similarity between u and v
defined by the formula above.
    """
    ## START YOU CODE HERE

    distance = 0.0

    dot = u.dot(v)
    norm_u = np.linalg.norm(u)
    norm_v = np.linalg.norm(v)
    cosine_similarity = dot/(norm_u * norm_v)

    return cosine_similarity


    ## END
```

Task 2: Implement `most_similar_word` which returns the most similar word to a word.

In [11]:
```
## Task 2
# GRADED FUNCTION: most_similar_word

def most_similar_word(word, word_to_vec_map):
    """
    Most similar word return the most similar word to the word u.

    Arguments:
        word -- a word, string
        word_to_vec_map -- dictionary that maps words to their corr
esponding vectors.

    Returns:
    best_word --  the most similar word to u as measured by cosine
similarity
    """
    ## START YOU CODE HERE
    e_a = word_to_vec_map[word]


    words = word_to_vec_map.keys()
    max_cosine_sim = 0
    best_word = None

    for i in words:
        cosine_sim = cosine_similarity(e_a, word_to_vec_map[i])

        if cosine_sim > max_cosine_sim and cosine_sim < 0.99999:
            max_cosine_sim = cosine_sim
            best_word = i

    return best_word


    ## END
```

Answer the questions below:

TASK 3: Write a code the answer the following questions:

What is the similarity between the words brother and friend?

In [12]:
```
cosine_similarity(word_to_vec_map["brother"], word_to_vec_map["friend"])
```

Out[12]: 0.8713178668124657

What is the similarity between the words computer and kid?

In [13]:
```
cosine_similarity(word_to_vec_map["computer"], word_to_vec_map["kid"])
```

Out[13]: 0.43800166210363856

What is the similarity between the words V1=(france - paris) and V2=(rome - italy)?

In [14]:
```
cosine_similarity(word_to_vec_map["france"] - word_to_vec_map["pairs"],
                  word_to_vec_map["rome"] - word_to_vec_map['italy'])
```

Out[14]: -0.1658124714422703

What is the most similar word to computer?

In [15]:
```
most_similar_word("computer",word_to_vec_map)
```

Out[15]: 'computers'

What is the most similar word to australia?

In [16]:
```
most_similar_word("australia",word_to_vec_map)
```

Out[16]: 'zealand'

What is the most similar word to python?

In [17]:
```
most_similar_word("python",word_to_vec_map)
```

Out[17]: 'reticulated'

Playing around the cosine similarity of other inputs will give you a better sense of how word vectors behave.

Word analogy task

In the word analogy task, we complete the sentence "*a* is to *b* as *c* is to **____**". An example is '*man* is to *woman* as *king* is to *queen*' . In detail, we are trying to find a word *d*, such that the associated word vectors e_a, e_b, e_c, e_d are related in the following manner: $e_b - e_a \approx e_d - e_c$. We will measure the similarity between $e_b - e_a$ and $e_d - e_c$ using cosine similarity.

Task 4: Complete the code below to be able to perform word analogies!

In [18]:
```
## Task 4
# GRADED FUNCTION: complete_analogy

def complete_analogy(word_a, word_b, word_c, word_to_vec_map):
    """
    Performs the word analogy task as explained above: a is to b as c is to ____.

    Arguments:
    word_a -- a word, string
    word_b -- a word, string
    word_c -- a word, string
    word_to_vec_map -- dictionary that maps words to their corresponding vectors.

    Returns:
    best_word --  the word such that v_b - v_a is close to v_best_word - v_c, as measured by cosine similarity
    """
    ## START YOU CODE HERE

    a, b, c = word_to_vec_map[word_a],word_to_vec_map[word_b],word_to_vec_map[word_c]


    words = word_to_vec_map.keys()
    max_cosine_sim = -100
    best_word = None


    for i in words:
        cosine_sim = cosine_similarity(b - a, word_to_vec_map[i]-c)

        if cosine_sim > max_cosine_sim:
            max_cosine_sim = cosine_sim
            best_word = i

    return best_word
    ##ENFD
```

Run the cell below to test your code, this may take 1-2 minutes.

In [19]:
```
triads_to_try = [('italy', 'italian', 'spain'), ('india', 'delhi',
'japan'), ('man', 'woman', 'boy'), ('small', 'smaller', 'big')]
for triad in triads_to_try:
    print ('{} -> {} :: {} -> {}'.format( *triad, complete_analogy(
*triad,word_to_vec_map)))
```

```
/Users/Howard/opt/anaconda3/lib/python3.7/site-packages/ipykernel_
launcher.py:22: RuntimeWarning: invalid value encountered in doubl
e_scalars

italy -> italian :: spain -> spanish
india -> delhi :: japan -> tokyo
man -> woman :: boy -> girl
small -> smaller :: big -> competitors
```

Once you get the correct expected output, please feel free to modify the input cells above to test your own analogies. Try to find some other analogy pairs that do work, but also find some where the algorithm doesn't give the right answer: For example, you can try small->smaller as big->?.

Debiasing word vectors

In the following exercise, you will examine gender biases that can be reflected in a word embedding, and explore algorithms for reducing the bias. In addition to learning about the topic of debiasing, this exercise will also help hone your intuition about what word vectors are doing. This section involves a bit of linear algebra, though you can probably complete it even without being expert in linear algebra, and we encourage you to give it a shot.

Lets first see how the GloVe word embeddings relate to gender. You will first compute a vector $g = e_{woman}-e_{man}$, where e_{woman} represents the word vector corresponding to the word *woman*, and e_{man} corresponds to the word vector corresponding to the word *man*. The resulting vector g roughly encodes the concept of "gender". (You might get a more accurate representation if you compute $g_1 = e_{mother}-e_{father}$, $g_2 = e_{girl}-e_{boy}$, etc. and average over them. But just using $e_{woman}-e_{man}$ will give good enough results for now.)

Task 5: Compute the bias vector using woman - man

In [20]:
```
## START YOU CODE HERE

g = word_to_vec_map['woman'] - word_to_vec_map['man']

## END

print(g)
```

```
[-0.087144    0.2182     -0.40986    -0.03922    -0.1032      0.94
165
 -0.06042     0.32988     0.46144    -0.35962     0.31102    -0.86
824
  0.96006     0.01073     0.24337     0.08193    -1.02722    -0.21
122
  0.695044   -0.00222     0.29106     0.5053     -0.099454    0.40
445
  0.30181     0.1355     -0.0606     -0.07131    -0.19245    -0.06
115
 -0.3204      0.07165    -0.13337    -0.25068714 -0.14293    -0.22
4957
 -0.149       0.048882    0.12191    -0.27362    -0.165476   -0.20
426
  0.54376    -0.271425   -0.10245    -0.32108     0.2516     -0.33
455
 -0.04371     0.01258   ]
```

Now, you will consider the cosine similarity of different words with g. Consider what a positive value of similarity means vs a negative cosine similarity.

Task 6: Compute and print the similarity between g and the words in name_list

---

In [21]:
```
print ('List of names and their similarities with constructed vecto
r:')
name_list = ['john', 'marie', 'sophie', 'ronaldo', 'priya', 'rahul'
, 'danielle', 'reza', 'katy', 'yasmin']

## START YOU CODE HERE
for i in name_list:

    print (i, cosine_similarity(word_to_vec_map[i], g))

## END
```

```
List of names and their similarities with constructed vector:
john -0.23163356145973724
marie 0.315597935396073
sophie 0.31868789859418784
ronaldo -0.31244796850329437
priya 0.176320418390094
rahul -0.16915471039231722
danielle 0.24393299216283895
reza -0.07930429672199552
katy 0.2831068659572615
yasmin 0.23313857767928758
```

TASK 7: What do you observe?

Boy has a negative values where girl has a positive one

Task 8: Compute and print the similarity between g and the words in word_list:

In [22]:
```
print('Other words and their similarities:')
word_list = ['lipstick', 'guns', 'science', 'arts', 'literature', '
warrior','doctor', 'tree', 'receptionist',
             'technology',  'fashion', 'teacher', 'engineer', 'pilo
t', 'computer', 'singer']

## START YOU CODE HERE
for i in word_list:
    print (i, cosine_similarity(word_to_vec_map[i], g))

## END
```

```
Other words and their similarities:
lipstick 0.2769191625638267
guns -0.1888485567898898
science -0.060829065409296994
arts 0.008189312385880328
literature 0.06472504433459927
warrior -0.2092016464112529
doctor 0.11895289410935041
tree -0.07089399175478091
receptionist 0.33077941750593737
technology -0.131937324475543
fashion 0.03563894625772699
teacher 0.17920923431825664
engineer -0.08039280494524072
pilot 0.001076449899191679
computer -0.10330358873850498
singer 0.1850051813649629
```

TASK 9: What do you observe?

A word more likely a man would we used or involved has a negative value where a word more likely for a woman has positive value

---

We'll see below how to reduce the bias of these vectors, using an algorithm due to Boliukbasi et al., 2016 (https://arxiv.org/abs/1607.06520). Note that some word pairs such as "actor"/"actress" or "grandmother"/"grandfather" should remain gender specific, while other words such as "receptionist" or "technology" should be neutralized, i.e. not be gender-related. You will have to treat these two type of words differently when debiasing.

Neutralize bias for non-gender specific words

The figure below should help you visualize what neutralizing does. If you're using a 50-dimensional word embedding, the 50 dimensional space can be split into two parts: The bias-direction g, and the remaining 49 dimensions, which we'll call $g_{\perp}$. In linear algebra, we say that the 49 dimensional $g_{\perp}$ is perpendicular (or "orthogonal") to g, meaning it is at 90 degrees to g. The neutralization step takes a vector such as $e_{receptionist}$ and zeros out the component in the direction of g, giving us $e_{receptionist}^{debiased}$.

Even though $g_{\perp}$ is 49 dimensional, given the limitations of what we can draw on a screen, we illustrate it using a 1 dimensional axis below.

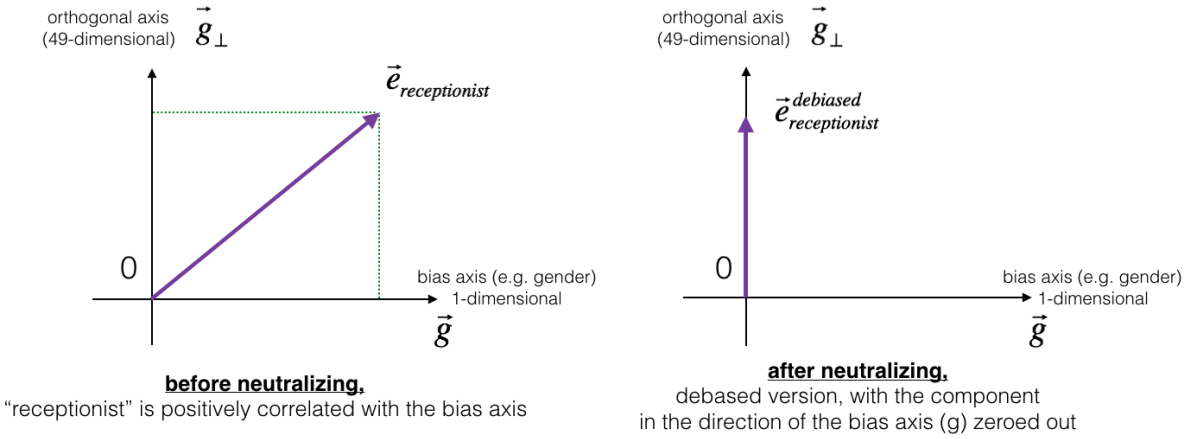


Figure 2: The word vector for "receptionist" represented before and after applying the neutralize operation.

TASK 10: Implement `neutralize()` to remove the bias of words such as "receptionist" or "scientist". Given an input embedding e, you can use the following formulas to compute $e^{debiased}$: \begin{equation} e^{bias_component} = \frac{e \cdot g}{||g||_2^2} * g\tag{2} \end{equation}\begin{equation} e^{debiased} = e - e^{bias_component}\tag{3} \end{equation}

If you are an expert in linear algebra, you may recognize $e^{bias\_component}$ as the projection of e onto the direction g. If you're not an expert in linear algebra, don't worry about this.

In [23]:
```
# TASK 10
# GRADED neutralize

def neutralize(word, g, word_to_vec_map):
    """
    Removes the bias of "word" by projecting it on the space orthog
onal to the bias axis.
    This function ensures that gender neutral words are zero in the
gender subspace.

    Arguments:
        word -- string indicating the word to debias
        g -- numpy-array of shape (50,), corresponding to the bias
axis (such as gender)
        word_to_vec_map -- dictionary mapping words to their corres
ponding vectors.

    Returns:
        e_debiased -- neutralized word vector representation of the
input "word"
    """
    ## START YOU CODE HERE

    e = word_to_vec_map[word]
    e_biascomponent = e.dot(g)/(np.linalg.norm(g)**2)*g
    e_debiased = e - e_biascomponent

    return e_debiased

    ## END
```

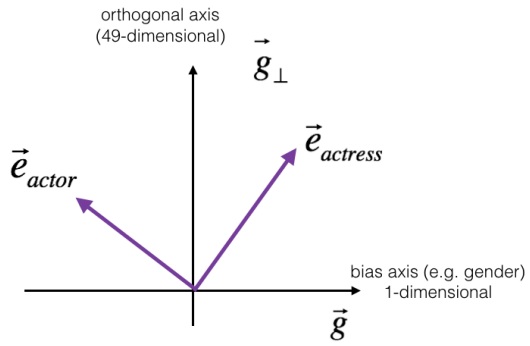
In [24]:
```
e = "receptionist"
print("cosine similarity between " + e + " and g, before neutralizi
ng: ", cosine_similarity(word_to_vec_map["receptionist"], g))

e_debiased = neutralize("receptionist", g, word_to_vec_map)
print("cosine similarity between " + e + " and g, after neutralizin
g: ", cosine_similarity(e_debiased, g))
```

```
cosine similarity between receptionist and g, before neutralizing:
0.33077941750593737
cosine similarity between receptionist and g, after neutralizing:
-5.841032332243514e-18
```

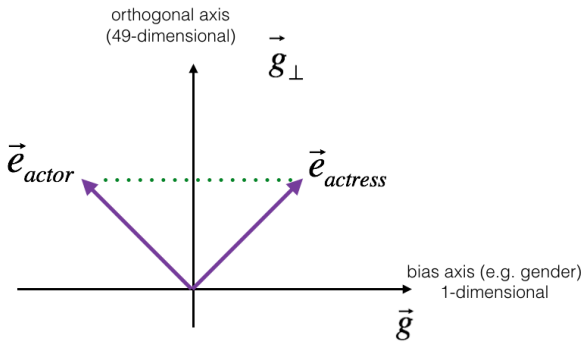
Equalization algorithm for gender-specific words

Next, lets see how debiasing can also be applied to word pairs such as "actress" and "actor." Equalization is applied to pairs of words that you might want to have differ only through the gender property. As a concrete example, suppose that "actress" is closer to "babysit" than "actor." By applying neutralizing to "babysit" we can reduce the gender-stereotype associated with babysitting. But this still does not guarantee that "actor" and "actress" are equidistant from "babysit." The equalization algorithm takes care of this.

The key idea behind equalization is to make sure that a particular pair of words are equi-distant from the 49-dimensional $g_\perp$. The equalization step also ensures that the two equalized steps are now the same distance from $e_{receptionist}^{debiased}$, or from any other word that has been neutralized. In pictures, this is how equalization works:



before equalizing,
"actress" and "actor" differ in many ways beyond the direction of $\vec{g}$



after equalizing,
"actress" and "actor" differ only in the direction of $\vec{g}$, and further are equal in distance from $\vec{g}_\perp$

The derivation of the linear algebra to do this is a bit more complex. (See Bolukbasi et al., 2016 for details.) But the key equations are:
\begin{equation} \mu = \frac{e_{w1} + e_{w2}}{2}\tag{4} \end{equation}\begin{equation} \mu_{B} = \frac {\mu \cdot bias_axis}{||bias_axis||_2^2} *bias_axis \tag{5} \end{equation}\begin{equation} \mu_{\perp} = \mu - \mu_{B} \tag{6} \end{equation}\begin{equation} e_{w1B} = \frac {e_{w1} \cdot bias_axis} {||bias_axis||_2^2} *bias_axis \tag{7} \end{equation}\begin{equation} e_{w2B} = \frac {e_{w2} \cdot bias_axis}{||bias_axis||_2^2} *bias_axis \tag{8} \end{equation}\begin{equation} e_{w1B}^{corrected} = \sqrt{ |{1 - ||\mu_{\perp} ||^2_2} |} * \frac{e_{\text{w1B}} - \mu_B} {||(e_{w1} - \mu_{\perp}) - \mu_B||} \tag{9} \end{equation}\begin{equation} e_{w2B}^{corrected} = \sqrt{ |{1 - ||\mu_{\perp} ||^2_2} |} * \frac{e_{\text{w2B}} - \mu_B} {||(e_{w2} - \mu_{\perp}) - \mu_B||} \tag{10} \end{equation}\begin{equation} e_1 = e_{w1B}^{corrected} + \mu_{\perp} \tag{11} \end{equation}\begin{equation} e_2 = e_{w2B}^{corrected} + \mu_{\perp} \tag{12} \end{equation}

TASK 11: Implement the function below. Use the equations above to get the final equalized version of the pair of words. Good luck!

In [26]:
```
# TASK 11
# GRADED equalize

def equalize(pair, bias_axis, word_to_vec_map):
    """
    Debias gender specific words by following the equalize method d
escribed in the figure above.

    Arguments:
    pair -- pair of strings of gender specific words to debias, e.g
. ("actress", "actor")
    bias_axis -- numpy-array of shape (50,), vector corresponding t
o the bias axis, e.g. gender
    word_to_vec_map -- dictionary mapping words to their correspond
ing vectors

    Returns
    e_1 -- word vector corresponding to the first word
    e_2 -- word vector corresponding to the second word
    """
    ## START YOU CODE HERE

    w1, w2 = pair
    e_w1, e_w2 = word_to_vec_map[w1],word_to_vec_map[w2]

    mu = (e_w1 + e_w2)/2

    mu_B = mu.dot(bias_axis)/(np.linalg.norm(bias_axis)**2) * bias_
axis
    mu_orth = mu - mu_B

    e_w1B = e_w1.dot(bias_axis)/(np.linalg.norm(bias_axis)**2) * bi
as_axis
    e_w2B = e_w2.dot(bias_axis)/(np.linalg.norm(bias_axis)**2) * bi
as_axis

    corrected_e_w1B = np.sqrt(np.abs(1-np.linalg.norm(mu_orth)**2))
*((e_w1B-mu_B)/(np.abs(e_w1-mu_orth-mu_B)))

    corrected_e_w2B = np.sqrt(np.abs(1-np.linalg.norm(mu_orth)**2))
*((e_w2B-mu_B)/(np.abs(e_w2-mu_orth-mu_B)))

    e1 = corrected_e_w1B + mu_orth
    e2 = corrected_e_w2B + mu_orth

    return e1, e2

    ##END
```

In [28]:
```
print("cosine similarities before equalizing:")
print("cosine_similarity(word_to_vec_map[\"man\"], gender) = ", cos
ine_similarity(word_to_vec_map["man"], g))
print("cosine_similarity(word_to_vec_map[\"woman\"], gender) = ", c
osine_similarity(word_to_vec_map["woman"], g))
print()
e1, e2 = equalize(("man", "woman"), g, word_to_vec_map)
print("cosine similarities after equalizing:")
print("cosine_similarity(e1, gender) = ", cosine_similarity(e1, g))
print("cosine_similarity(e2, gender) = ", cosine_similarity(e2, g))
```

```
cosine similarities before equalizing:
cosine_similarity(word_to_vec_map["man"], gender) =  -0.1171109576
533683
cosine_similarity(word_to_vec_map["woman"], gender) =  0.356666188
4627037

cosine similarities after equalizing:
cosine_similarity(e1, gender) =  -0.7165727525843933
cosine_similarity(e2, gender) =  0.7396596474928908
```

TASK 12: What do you observe?

Same as above, man has negative value where woman has positive value

---

Please feel free to play with the input words in the cell above, to apply equalization to other pairs of words.

These debiasing algorithms are very helpful for reducing bias, but are not perfect and do not eliminate all traces of bias. For example, one weakness of this implementation was that the bias direction g was defined using only the pair of words *woman* and *man*. As discussed earlier, if g were defined by computing $g_1 = e_{woman} - e_{man}$; $g_2 = e_{mother} - e_{father}$; $g_3 = e_{girl} - e_{boy}$; and so on and averaging over them, you would obtain a better estimate of the "gender" dimension in the 50 dimensional word embedding space. Feel free to play with such variants as well.

Congratulations!

You've come to the end of this assignment, and have seen a lot of the ways that word vectors can be used as well as modified. Here are the main points you should remember:

- Cosine similarity a good way to compare similarity between pairs of word vectors. (Though L2 distance works too.)
- For NLP applications, using a pre-trained set of word vectors from the internet is often a good way to get started.
- Bias in data is an important problem.
- Neutralize and equalize allow to reduce bias in the data.

Congratulations on finishing this notebook!

References

- The debiasing algorithm is from Bolukbasi et al., 2016, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf)
- The GloVe word embeddings were due to Jeffrey Pennington, Richard Socher, and Christopher D. Manning. (https://nlp.stanford.edu/projects/glove/ (https://nlp.stanford.edu/projects/glove/))