Name: Chakrya Ros

Section 5: Bias (10 points)

1. What is one way that NLP systems are biased? Give an example grounded in a specific task, using a specific model that we covered in this class, then explain why the bias that you've identified might occur (use model specifics) and explain why it is a difficult problem to solve?

One way that NPL systems are biased is allocation bias that the system unfairly allocates resources to certain group over the other. Gender bias is the one of allocation bias that NPL systems is difficult to solve. For example, the word man can be used to describe a male human being, but it can also be used as a verb in 'man the station" or an exclamation in "oh man!". Machine translation, word embedding, and sentiments analysis are the model that I have identified some bias. In homework4, we trained word embeddings using Spooky Authors dataset and our own dataset. The outputs are different. Below is the screenshot that we found the 10 similar word to 'man' in both datasets.

```
#Skoopy Authors Dataset using CBOW Model
wb = CBOW_Model()
cbow_model = wb.train('skoopy.csv')

#get the vocabulary from model
vocabs = list(cbow_model.wv.vocab)
print("vocabulary size of Skoopy dataset : ", len(vocabs))
#get encoding of word of 100 dimemsion
word_vectors = cbow_model.wv['love']

#words to display on the graph of the similar words in this list
words = ['dinner', 'happiness', 'man', 'sat', 'illness', 'day', 'home', 'two']

#find the most similar word
w = cbow_model.wv.most_similar(['man'])
wl = cbow_model.wv.most_similar(['dinner'])
print(w)

#similarity between two differen word
print("CBOW model find similarity between 'man' and 'woman': ",
      cbow_model.wv.similarity(wl="man", w2="woman"))

# tsne_plot_visualize("Skoopy Authors Dataset Using CBOW Model", words, cbow_model)
```

```
vocabulary size of Skoopy dataset :  15662
[('swede', 0.9102968573570251), ('usages', 0.9080865979194641), ('gouty', 0.8851004838943481), ('misanthrope', 0.8823
884129524231), ('hovers', 0.877611517906189), ('old', 0.8650044202804565), ('bugs', 0.8648256063461304), ('sensibilit
ies', 0.8625751733779907), ('woman', 0.8624029159545898), ('credited', 0.861343264579773)]
CBOW model find similarity between 'man' and 'woman':  0.86240304
```

```
vocabulary size of Emma dataset :  7827
Most similar word to 'man' [('woman', 0.941717803478241), ('fastidious', 0.9289137125015259), ('young', 0.87422823905
94482), ('incomprehensible', 0.8662596344947815), ('bestride', 0.8376628756523132), ('concise', 0.8369568586349487),
('thoughtless', 0.835969090461731), ('intreat', 0.8333361148834229), ('rated', 0.8331584930419922), ('artful', 0.8321
38180732727)]
```

It's difficult to solve because the limitation of dataset, different datasets would get different outputs. Most NLP systems are trained and tested on text sampled from natural sources like news, blogs, Twitter, etc. So the system still get error.

Citations:

- Lecture 41, 42
- https://medium.com/dair-ai/examining-gender-and-race-bias-in-sentiment-analysis-systems-b04b269a653
- https://fisher.wharton.upenn.edu/wp-content/uploads/2019/06/Thesisi_Sosnick.pdf
-