NLP Problem Set 2

Rutvik Pansare, Kula Maganti, Tyler Arnett

Q 2.2

```
In [45]: model = gensim.models.Word2Vec(toksents,min_count=1,size=100, window=8, workers=4)
         print("Done")

         Done

In [41]: words=list(model.wv.vocab)

In [42]: model.wv['doctor']
         model.wv.similarity('doctor', 'woman')

Out[42]: 0.4872672

In [43]: model.wv['doctor']
         model.wv.similarity('doctor', 'man')

Out[43]: 0.47921813

In [57]: model.wv.similarity('father','mother')

Out[57]: 0.8464284

In [58]: model.wv.similarity('father','son')

Out[58]: 0.8195087

In [64]: model.wv.similarity('guy','dude')

Out[64]: 0.677415
```

In this particular, a substantial set of cohesive text used (text8) and word2vec word embedding model was trained using gensim. Looking at the figure above, a couple of examples were observed. First, when it came to gender bias, notice how the model is pretty split between "doctor" and " man and woman", this is good in my opinion because it shows that the model was to trained to not to bias when it comes to occupation. Moving along, something that was interesting was family. When looking a " father and mother" or "father and son", the model learnt the these were similarity. One can believe that this is the case because the context is family or parents. Lastly, "guy and dude" the model learned this to be similar due to the context of a male, which is male. After these examples were analyzed, the model seemed to pretty accurate and not gender bias. Although, improvements can be made to strengthened that.