Homework 4: Bias in Language Models

Name: Sai Sahana Bhargavi Byrapu

1.Introduction: to explore the implicit bias that can be encoded in natural language word embeddings using the tools provided for testing associations with pairs of words and to show that by training on natural data sets, these models also incorporate cultural biases**.**

2.Approach:

Corpora : GloVe based word embedding models trained on large data sets of examples of human language like Wikipedia (dimensions of 50d,100d,200d,300d) and Twitter(25d,50d,100d,200d).

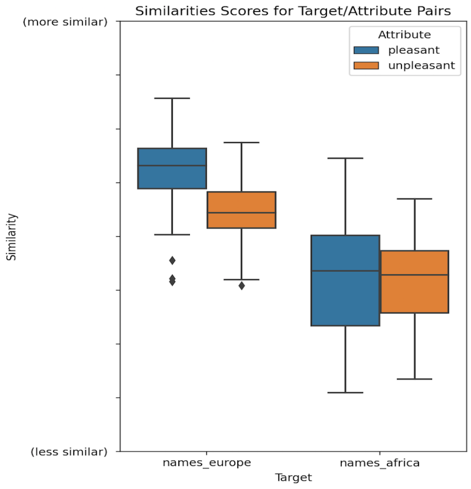
WEAT (Word Embedding Association Test)algorithm:

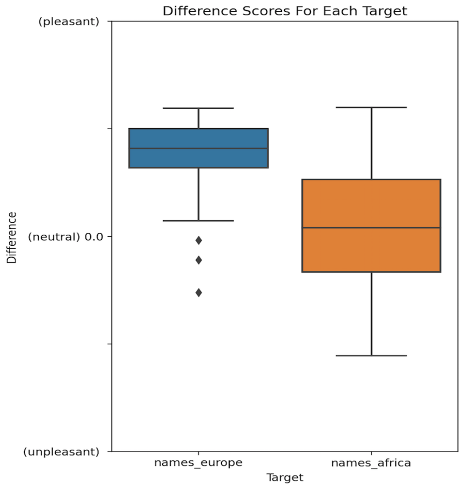
For every run, a pair of targets, and a pair of attribute words are given.

For example, in the below command**,** glove_wikipedia_50d is the numpy file of wikipedia dataset of 50d(dimension) , names_europe and names_africa are list of target words, pleasant and unpleasant are list of attribute words.

```
./weatTest.py glove_wikipedia_50d  names_europe names_africa pleasant unpleasant
```

The algorithm computes the average similarity between target1 and both the attributes and similarly between target2 and both the attributes using cosine similarity. Next, calculate the average difference in similarities divided by standard deviation which gives the final metric termed as 'effect size'. further, it generates box plots for pairwise similarity scores and a box plot for target bias in similarities.like mentioned below.





Part 1:Effect of underlying corpus

After running the pairings for the wikipedia and twitter datasets,the quantitative results are as shown in the table below.

Effect sizes for the underlying corpus		wikipedia				twitter			
Target words	**Attribute words**	50d	100d	200d	300d	25d	50d	100d	200d

European names / African names	pleasant/unpleasant	1.14	1.14	1.22	1.21	1.16	1.11	1.18	1.17
Flowers/Insects	pleasant/unpleasant	1.08	1.18	1.15	1.41	1.25	1.23	1.12	1.10
European names / African names	positive words / negative words	1.22	1.4	1.44	1.44	1.35	1.25	1.24	1.29
Male names/ Female names	career/family	1.76	1.77	1.80	1.75	1.32	1.26	1.28	1.21
Science/Arts	gender_m/gender_f	1.56	1.10	1.34	1.36	0.25	-0.68	-0.05	0.44

As per the WEAT algorithm instructions, the calculated effect size i.e the score is between +2.0 and -2.0. Positive scores indicate that target1 is more associated with attribute1 than target2.Or, equivalently, target2 is more associated with attribute2 than target1.Negative scores have the opposite relationship.Scores close to 0 indicate little to no effect.

case 1,2,3 :Based upon this, it can be observed that there a noticeable change in each case for both wikipedia and twitter datasets of all dimensions where scores are ranging from 1.12,1.2,1.3 for top three cases, thereby stating that there exists bias towards european names compared to african names, flowers compared to Insects while classifying the pleasant and unpleasant positive words are attributed to europeans and negative words are associated with africans. **case 4:**In the next case there is change among the wikipedia and twitter datasets where the score is ~1.8 in former and ~1.26 in later, but there is a bias in concluding that males are more associated with career while females are with family for a given pair of male names and female names.
case 5: Similarly, from the scores of ~1.4 in wikipedia science is more attributed to males when compared to females, deriving that arts is most preferred with females whereas twitter analyses gave mixed response for the varied dimensionalities.

Part 2:New set of WEAT pairings
In order to check for additional biases ,created a new set of WEAT pairings and thus generated a list of words for the target pair and used the existing attribute pairs (unpleasant/pleasant, family/career, male/female) and own one(temporary/permanent). The results are recorded as mentioned below after running the created pairs with the respective attribute words.

Effect sizes for the		**wikipedia**				**twitter**			
Target words	**Attribute words**	**50d**	**100d**	**200d**	**300d**	**25d**	**50d**	**100d**	**200d**
religion: hindu/muslim	pleasant/unpleasant	0.78	0.10	1.02	1.44	-1.29	-1.39	-1.32	-1.14
food: junk/healthy	pleasant/unpleasant	0.85	0.85	0.55	0.65	0.39	0.53	0.64	0.64

nationality: american/indian	career/family	-1.25	-0.29	-0.37	-0.16	-1.25	-1.38	-1.01	-0.84
toys: gadgets /non-gadgets	gender_m/gender_f	1.30	1.13	1.07	1.19	1.25	1.40	1.65	1.49
disease:mental/ physical	temporary/ permanent	-1.07	-1.16	-0.98	-0.92	-0.85	-0.85	-0.77	-0.82

case 1: created a list of word for religions -hindu and muslim in order to check the bias for pleasant/unpleasant, it can be noticed that the values are positive for all dimensions in wikipedia dataset, there exists bias toward hindus compared to muslims where as twitter dataset imbibes bias toward muslims as score is negative.
case2: for the generated list of words for foods-junk and healthy, it can be seen that the values are positive and are close to 0.8 in wikipedia and 0.6 in twitter, this means there is bias towards junk food stating it be pleasant and the healthy foods are unpleasant.
case 3: for the list of words of nationality-american/indian- scores are negative for all dimensions for both wikipedia and twitter, this means there is bias in stating that americans are more attributed to family and indians are associated to careers.
case 4,5:positive scores indicate that gadgets are more attributed to males compared ti females and negative scores contribute that mental diseases are permanent while physical diseases are temporary

List of words for the target and attribute pairs.

hindus	muslims	gadgets	non-gadgets
iyengers	khan	robots	dolls
aryan	kaif	drones	barbies
brahmin	mohammad	lego	kitchen
kshatriya	sheikh	cars	disney
vaishyas	malik	remote-controlled	jewellery
chopra	pathan	airplane	beads
gupta	qureshi	rockets	bracelet
			painting
			Storybook

junk	healthy	mental	physical
coke	lettuce	Depression	Allergies
burger	salad	Anxiety	Cold
pizza	eggs	Obsessive-compulsive	Flu
chips	oats	Bipolar	Conjunctivitis
fries	quinoa	Schizophrenia	Diarrhea
chocolate	juice	trauma	Headaches
brownie	spinach	stress	Mononucleosis
cake	protein		Stomachaches
mcchicken	sandwich		

mutton
cheesesteak
bacon

```
temporary    permanent           temporary    permanent
 ------------------------------------------------------------------------------
 tentatively        always            provisionally          forever
transitorily         aye              briefly                 forevermore
momentarily     eternally                                     indelibly
perpetually       ever                                        perpetually
perennially      everlastingly                               evermore
```

Conclusion: This homework introduced me to word embeddings, mappings from words (or phrases) to vectors of number and how the computer learns the underlying meaning of the human understandings of fairly similar words through assigning similar vector values using a word embedding model implemented by an algorithm like Glove and word2vec and the how the biases can be encoded in the natural language word embeddings.