Andrew Gerlach
HW 1
1/27/2020

1)
P(Comedy | fast) = P(fast | Comedy)*P(Comedy) = $(\frac{1+1}{7+7}) * (\frac{2}{5}) \; = \; 0.06$
P(Action | fast) = P(fast | Action) * P(Action) = $(\frac{2+1}{10+7}) * (\frac{3}{5}) \; = \; 0.11$
Prediction for fast is Action

P(Comedy | couple) = P(couple | Comedy)*P(Comedy) = $(\frac{2+1}{7+7}) * (\frac{2}{5}) \; = \; 0.09$
P(Action | couple) = P(couple | Action) * P(Action) = $(\frac{0+1}{10+7}) * (\frac{3}{5}) \; = \; 0.04$
Prediction for couple is Comedy

P(Comedy | shoot) = P(shoot | Comedy)*P(Comedy) = $(\frac{0+1}{7+7}) * (\frac{2}{5}) \; = \; 0.03$
P(Action | shoot) = P(shoot | Action) * P(Action) = $(\frac{3+1}{10+7}) * (\frac{3}{5}) \; = \; 0.14$
Prediction for shoot is Action

P(Comedy | fly) = P(fly | Comedy)*P(Comedy) = $(\frac{1+1}{7+7}) * (\frac{2}{5}) \; = \; 0.06$
P(Action | fly) = P(fly | Action) * P(Action) = $(\frac{1+1}{10+7}) * (\frac{3}{5}) \; = \; 0.07$
Prediction for fly is Action

2)
Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. This paper showed how debiasing techniques could be expanded from a binary classifier to a multi-class classifier. The researchers came up with a way to identify the bias subspace of an embedding (ex. {man, woman} for a gender subspace) then they first compute the vector differences of the word embeddings of words in each set from the set's mean, and then taking the most significant components of these vectors. To remove those components of the embedded vectors they used two methods: hard-debiasing and soft-debiasing. In hard-biasing each component of the bias embedding is computed and removed from words that should be bias-neutral (doctor, nurse, teacher, etc). Soft-debiasing creates a projection of the embedding matrix that preserves the relationship between the biased and de-biased embeddings with the projection onto the bias subspace of neutral words being minimized. Overall, the team saw an increased cosine distance in the de-biased words making the probability of a biased word association smaller. The evaluation was done on several downstream tasks.

Gender Bias in Contextualized Word Embeddings. This team's goal was to "quantify, analyze, and mitigate gender bias exhibited in ELMo's contextualized word vectors". For quantification and analyzation they found that ELMo has far more occurences of male data than female data which is systematically encoded in the geometry of the word embeddings which them propagates the gender information unequally. The mitigation technique they used involved data augmentation which replaced gender-revealing data with data that indicates the opposite gender and adding it, not replacing it, to the original data. The other technique used was Neutralization. Instead of adding to the training data by swapping gender words, the gender-swapped version of the test instances were created. The average of the gender-swapped and original representations were used. They found that both techniques improved the bias with augmentation doing slightly better than neutralization. They evaluated by measuring how predictable gender is from ELMo representations of occupation words that co-occur with gender revealing pronouns.

Probabilistic Bias Mitigation in Word Embeddings. The de-biasing in this experiment used a loss function that penalizes the discrepancy between the conditional probabilities of a target word (one that should not be affected by the attribute) conditioned on two words describing the attribute (man and woman in the case of gender). That is, for every target word we seek to minimize: P(target|a) - P(target|b) where *a* and *b* are a pair of words in a pair like {man, woman}. The other method was a "Neighborhood Effect". This technique looked at the k/2 nearest neighbors from male-associated bias words and the k/2 female-associated biased words. The loss function minimizes P(target|m) - P(target|b) where *m* is a male-biased word and *f* is a female-biased word sorted by L1 distance. This is summed over the k/2 nearest neighbors. Their assumption was that if the word is unbiased then the probabilities should be close. Their results showed a reduction in bias after employing these techniques.

3) Code submitted separately.

4)

1) The striped bats were hanging on their feet and ate best fishes.

The stripe bat be hang on their foot and eat best fish.

2) At some point, I realized that I started to use the bike more often, not only to get to work but also to catch up with friends and to head out for coffee on weekends.

At some point I realize that I start to use the bike more often not only to get to work but also to catch up with friend and to head out for coffee on weekend.