CSCE 771: Computer Processing of Natural Languages

Student Name:

Objective: The objective of the Quiz is to learn bias issues with the usage of large language models on NLP tasks with hands-on experience.

Tasks:

1. Read paper -

[10 + 10 + 10 = 30]

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, Adam T. Kalai, Neurips 2016, Link:

Now answer the following:

a) According to you, why does this behavior of word embeddings happen when running analogies?

Because analogies explicitly represent the relationships between each word or each phrase. Word embedding exactly not only serves as a dictionary of sorts but also embeds the similar semantic meanings to represent the word meanings.

b) What approach is proposed in the paper to mitigate it?

The authors describe their method, a hard-debiasing algorithm, and they conclude from empirical results that the algorithm significantly reduces gender bias in word embeddings trained on Google News articles.

The main procedure for the experiment was as follows:

- Use a word embedding to create analogies that solve the following task: 'she' is to *x* as 'he' is to *y*
- Run the hard-debiasing algorithm to remove gender bias in the word embedding
- Use the modified word embedding to solve the same analogy task, then analyze the results

c) Do you think the approaches actually mitigate? Any problems you anticipate in practice?

The results of the algorithm were demonstrated by comparing crowd-worker evaluation of analogies formed before and after debiasing. The proportion of analogies that were judged by a majority of the workers as showing gender stereotypes was reduced from 19% to 6%.

2. Create and run notebook

[10 + 10 + 20 + 20 = 60]

Create your own copy python notebook and execute from the tutorial python notebook

at: https://github.com/PLN-FaMAF/Bias-in-wordembeddings/blob/main/main_tutorial_bias_word_embedding.ipynb

Activities to do are:

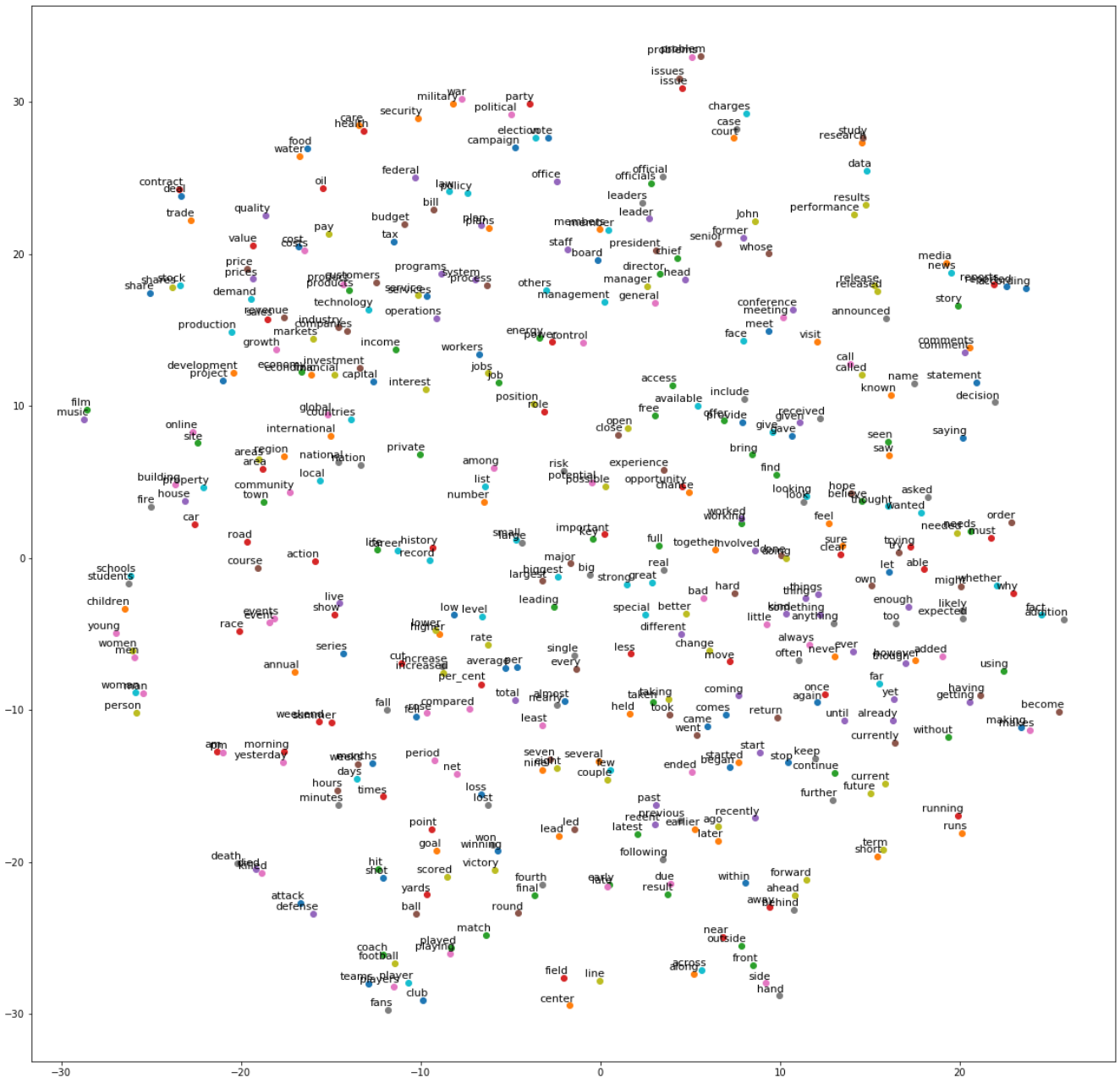
a. load word embedding

b. do visualization

c. run analogies: examples and your own (at least 3)

d. run one mitigation method

Link to your completed notebook is at: <Please fill >



```
In [61]: w2v_small.most_similar(positive=['textbook', 'smaller'], negative=['small'], topn=2)
```

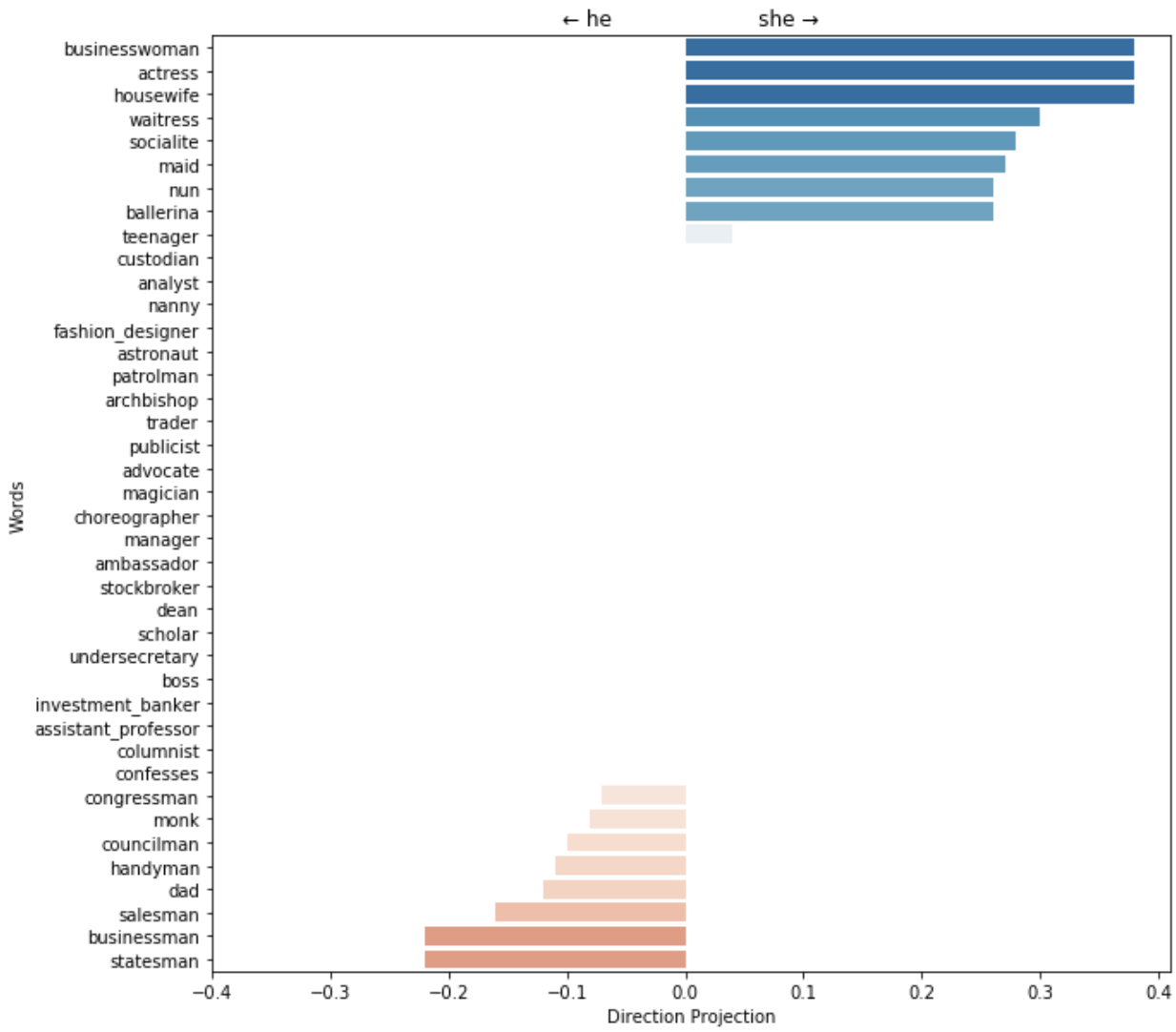
```
Out[61]: [('textbooks', 0.5776058435440063), ('books', 0.38853728771209717)]
```

```
In [62]: w2v_small.most_similar(positive=['computer', 'beautiful'], negative=['ugly'], topn=2)
```

```
Out[62]: [('computers', 0.5134012699127197), ('laptop_computer', 0.47052162885665894)]
```

```
In [63]: w2v_small.most_similar(positive=['garden', 'king'], negative=['queen'], topn=2)
```

```
Out[63]: [('gardens', 0.6007475852966309), ('gardening', 0.48624253273010254)]
```



3. Apply to your project

[3 + 4 + 3 = 10]

Now consider your course project.

a. Does gender bias in word embedding is relevant to your project? How?

Yes. In my project, VQA models have been shown to exploit language bias by learning the statistical correlations between questions and answers without looking into the image content: e.g., questions about the color of a banana are answered with yellow, even if the banana in the image is green.

b. If yes, what strategy will you use to improve to the situation?

We have found that some of the VQA datasets contain harmful samples that exhibit gender or racial stereotypes. Such samples are often found when the associated question is unanswerable from the image content. The ideal solution is to remove harmful samples by conducting a manual filtering.

c. If no, what ethical issue(s) do you anticipate for your project and the strategy one may use?