# CSCE 771: Computer Processing of Natural Languages Prof. Biplav Srivastava, Fall 2022

## Quiz 4 / Instructions

- This quiz is a mix of paper reading and running working code to understand the concepts underlying the paper
- Code has to be submitted in a directory of your GitHub called "Quiz4" with sub-dir for code, data and doc. Code will have your source code, data will have any input or output generated, and doc will have a .pdf of this file (called Quiz4-CSCE771-answers.pdf) along with any answers
- Complete quiz by 1:00 pm on Thursday, Dec 8, 2022. Hand over printout in class / in- person or send pdf as an email to biplav.s@sc.edu confirming completing the quiz and attaching your Quiz4-CSCE771-answers.pdf.
- Total points = 30 + 60 + 10 = 100
- Obtained

= Student

Name:

Guangyi Chen

---

**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:
   https://papers.nips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5- Abstract.html

   Now answer the following:
   *a*)According to you, why does this behavior of word embeddings happen when running analogies?

For example :
man is to king as woman is to? we need women+ (king - man)

(King - man) is a Vector , means a offset , when women add it we can find the answer words

With these vectors, we can perform the kind of arithmetic operations

Then we can start **exploring** our embedding to detect possible harmful bias

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

1. Project each **neutral profession word** on the direction
2. Calculate the absolute value of each projection
3. Average it all

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

not quite

An interesting property of this method is that it gives us a quantitative measure of the bias in the embedding, but keep in mind that this number only takes into consideration the bias present in the *assumed neutral* words with respect to the *defined gender direction*, which we agreed was a simplification of the actual phenomenon.

1 Words that are not in the list.

2 This list did not comprehensive enough.

We intentionally do not reference the resulting embeddings as "debiased" or free from all gender bias and prefer the term "mitigating bias" rather than "debiasing," to guard against the misconception that the resulting embeddings are entirely "safe" and need not be critically evaluated for bias in downstream tasks.
As an approach to mitigating bias, we will remove the gender projection from the words in the gender-neutral word list and then normalize.

This method *needs* a list of gender-neutral words to neutralize, or a list of words that are inherently gendered to not neutralize.

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-word-embeddings/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: < https://github.com/simonachen11/771-quiz4/blob/main/quiz4%20code.ipynb>

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?

Not quite, in this part it only has two genders, which means not A but B. In my project, we need to detect each location which is department, police office highway or something else.

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

I will consider the way to Mitigating Bias.

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

Some places start with the person’ s name, but it was actually a department name.

so, entity extraction identify these as a person because the name contains a person’ s name in it.

In this case, these needs to be selected and relabel as department name.