NLP CS6320 Homework 2

Due February 29, 23:59
30 points

Word2vec - the good (5 points)
The famous property of word2vec vectors are ability to manipulate “meaning” as vector:
king - man + woman = queen (title - gender)
Paris - France + Italy = Rome (country - capital)

Please come up with two more types of such manipulation. For each type, provide at least three working examples. Use word2vec package for that.

Word2vec - the bad (5 points)
Please come up with a relationship that is not captured by word2vec. Give at least three specific examples where it is not working.

Word2vec - the ugly (5 points)
ML can propagate societal biases. In context of the word embedding, it’s about distance between words: "nurse," "teacher," or "secretary" more closely with "woman" and "engineer," "scientist," or "pilot" more closely with "man". Bias can be based on gender, age, ethnicity or culture.

Please come up with your own specific example of bias in word2vec.

Spacy (15 points**)**
Go through spacy 101: https://spacy.io/usage/spacy-101
Come up with your own sentence and provide the output (screenshot) of the following steps:

- Tokenization
- Part of speech tagging
- Named entities
- Dependency parsing

Bonus points (you might need it for the project):

1. Do your own Named Entity extraction using rules: https://spacy.io/usage/rule-based-matching

2. Use visualizers: https://spacy.io/usage/visualizers

You can embed this visualizations into your UI for the project, if relevant