1. True/False (4 pts, recommended 20 minutes)

For each of the statements below, indicate if it is true or false. **If it is false, explain why it is false and provide an example. You may provide an explanation if it is true in case you are wrong and would like to receive partial credit.**

A. The regex **\bThor\b** will result in a higher false positive rate than **Thor** when searching for references to the movie character Thor.

*False. It will reduce the FPR. An FPR here would be something like the would **Thorough**. These are not truly references to Thor but will be categorized as such. Using the word boundaries will reduce these occurrences.*

B. In a Hidden Markov Model's **emission matrix**, assuming rows are **observed states** and columns are **hidden states**, the sum of each row should equal 1.

False. Each row is one word, and each value in a row is the likelihood of a hidden state emitting that word. The columns should sum to 1.

C. A document that has original text **cat litter** and another document that has the text **litter cat** will have identical vectors when using word count vectorization, TF-IDF vectorization, and bag-of-words word2vec vectorization.

True. All of these are BOW (bag of words) models that do not incorporate sequence. Note – word2vec DOES take into account context, but when you combine word embeddings to form a document vector, that does NOT taking into account the order of words.

D. Two documents:

   a. **cat cat dog dog love love**
   b. **cat dog love**

   Would show a **cosine distance** > 0.

False – they will show a cosine distance of 0, since they would show a cosine similarity of 1.

E. There are 3 capture groups in the following regex:
(?:Mr\.|Miss)\s(\w)\s(?P<last_name>\w)

False. There are only two capture groups – the first group is a non-capture group.

F. **UTF-8** and **ASCII** use the same **Unicode codepoint** for the character “a”.

True. UTF-8 is backwards compatible with ASCII, so the first 127 characters are the same. The character a is represented via the code point 97.

G. If a model’s F1 score is 1, it is guaranteed to have 0 false positives and 0 false negatives.

True. If F1 score is 1, then both precision and recall must be 1. If both precision and recall are 1, that means there can be no false positives or false negatives.

H. If you have a word2vec neural network, with **V** total unique words in your entire vocabulary, and are trying to train word embeddings of size **M** dimensions, the output of the **softmax layer** of word2vec is of shape **M x 1**.

False – the output of the softmax layer is the final computation before you compare against the ground truth, which is a vector of size V x 1 (one hot encoded, with only the context word position in the vector = 1, everything else 0). Therefore, the softmax layer should also be V x 1.