Part 1

1. In 1-3 sentences, describe what goes wrong with our topic modeling analysis if we do not normalize and pre-process the text data.

If we don’t normalize and pre-process data, common words like conjunctions, articles will be given disproportionate weightage because of their frequency, thereby providing no meaningful inference. The purpose of LDA is to discover collocated words and group them in a single topic.

1. Stemming/Lemmatizing: If we don’t take the root word of terms, it can be difficult to derive patterns which might be represented by several different forms of the same word
2. Stop words: Not removing stop words can add a lot of noise in our model as these terms have low semantic meaning and don’t contribute much to topic discovery.
3. In 1-3 sentences, describe the importance of using the term frequency inverse document frequency (tfidf) transformation.

TFIDF extracts the most important words by normalizing the frequency in documents by comparing with the entire corpus as a whole. As a result, common words (eg: the , and ,og) that appear in all documents are assigned low weightage whereas unique or low frequency words (eg: neuroscience) are emphasized.

1. In 1-3 sentences, describe what the area under the precision-recall curve tells us about the performance of the classifier.

Precision = TP / (TP + FP).

Recall = TP / (TP + FN).

The area under the precision-recall curve, gives us a measure of performance of the classifier.

An ideal classifier will have 100% precision and 100% recall, and the area under the curve would be 1.

A random classifier will set the baseline, at the initial ratio of TP: TN (i.e, a random classifier will predict P or N with a 50% chance when there are equal number of samples, and thus has 50% precision).

Thus, the AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example,

1. In 1-3 sentences, describe the effect of lines 37-39, where the number of words per topic is being increased.  What are the tradeoffs between too few and too many words?

With few words per topic, the meaning of the words would be hard to discover with a few terms, as pattern discovery is more difficult.

With too many words per topic, we may be capturing lot of noise and irrelevant terms not necessarily sharing a strong relation with the topic data as other terms.

Part 2

B) 3-5 sentences discussing the difference in performance you see. Why is one performing better than the other?  (The results may not be what you expect --- consider why)

The AUC for TFIDF is 0.96 v/s 0.89 for Doc2vec. There are several reasons why Doc2vec would be performing worse than TFIDF

1. Doc2Vec being an unsupervised algorithm requires a large training set to perform with high accuracy.
2. Unsupervised algorithms like Doc2Vec requires specific parametric tuning for high performance. We have not optimized our model for these parameters.
3. Doc2Vec compresses the words into a smaller vector, losing some information along the way in order to find a good representation of a document. This loss in information might cause a drop in the accuracy compared to TF-IDF which is able to generalize well for smaller corpus by retaining more information about the documents.