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.

Normally, we normalize the data to create a linear and robust relationship. Usually, when relationship between two variables is non-linear, normalization helps make it linear.

In this context, normalization helps identify words to stem from the same word and reduce the size of the vocabulary. For example, lying, liar, lies will become lie.

Preprocessing the data helps in increasing the apparent content of topic summaries, removing the redundant words and relevant words that convey the same message or topic.

1. 1-3 sentences, describe the importance of using the term frequency inverse document frequency (tfidf) transformation.

TFIDF tells us how important a word is to a document in the document corpus. It will give a less weightage to the words that occur very frequently in all the documents. This will help us identify the most important words that contribute towards the semantic meaning of the document

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

A precision recall curve shows us the tradeoff between precision and recall for different thresholds.

Higher area in the curve shows high precision and high recall.

The Smaller values on the x-axis of the plot indicate lower false positives and higher number of true negatives.

The Larger values on the y-axis of the plot indicate higher true positives and lower false negatives.

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?

When there are too few words there is no absolute clarity in a document and the topic is not clear.

An increase in the number of words in the topic can help us identify more subtopics but an increase beyond a certain limit can lead to overfitting and may also make the topic too broad to understand.

Part 2:

1. 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 TFIDF classifier had higher accuracy in predicting the job titles and classifying the topics through text classification. Surprisingly, doc2vec did not give a better accuracy, one of the possible reasons can be due to size of the train data, which is small.