Problem 1: (10 pts) Describe why the BOW feature representation limits our ability to model human language. What aspect of language, and specifically word meaning, does BOW ignore? Do approaches like TF-IDF and PMI resolve this shortcoming in BOW?

After building BOW models, we start facing issues when we come across new sentences.

If the new sentences contain new words, then our vocabulary size would increase and thereby, the length of the vectors would increase too. Additionally, the vectors would also contain many 0s, thereby resulting in a sparse matrix. We are retaining no information on the grammar of the sentences nor on the ordering of the words in the text. TF-IDF gives larger values for less frequent words and is high when both IDF and TF are high.

Problem 2: (10 pts) The word2vec language modeling approach was perhaps the first successful method to learn meaningful word representations. How does word2vec assign/measure similarity between two words?

By calculating the Angle between word vectors using inner product and cosine similarity we can determine whether two word-vectors are parallel. If they are parallel, they are similar.

Problem 3: (10 pts) Why are the inner product and cosine similarity used to measure similarity and not Euclidean distance?

The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance because of the size they could still have a smaller angle between them. Small angle indicates higher similarity.

Problem 5: (20 pts) Recall from Lecture 03 that the principle of maximum likelihood makes two qualifying assumptions for any dataset/model combination:

* + all examples are drawn from the same distribution
  + all examples are independently drawn

Which of these qualifying assumptions do we break when learning the parameters of a language model using MLE?

We assume that misspelling in a word is independent of misspellings in other words. The strict conditional probability demands us to count for every word before the target word. We break the second assumption “all examples are drawn from the same distribution” when learning the parameters of a language model using MLE.