**Problems**

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?***

The Bow ignores the syntactic and structural information from the texts so that this will somewhat limit the process of modeling human language. What’s more, the BOW focuses only on the number of occurrences of the word, and not to consider the order.

1. ***(20 pts) The word2vec language modeling approach was perhaps the first successful method to learn meaningful word representations. Answer these questions:***

***(a) How does word2vec assign/measure similarity between two words?***

By calculating the Angle between word vectors using inner product and cosine similarity: if two word-vectors are parallel, then they are similar.

***(b) When $N$ is large, what computational bottle neck arises in word2vec that requires us to change the algorithm?***

When the dimensions are too large, the amount of computation increases dramatically, and this affects the speed of computation.

***Note: we did not tweak the algorithm in the in-class demo, but did mention it during the lecture/demo?***

1. ***(20 pts) Why are the inner product and cosine similarity used to measure similarity and not euclidean distance?***

If word frequency or word vector are used as features, the Euclidean distance of a pair of texts in feature space is usually very large when the length of similarity is very different but the content is similar. If cosine similarity is used, the Angle between them may be small, so the similarity is high.

1. ***(50 pts) Write a Python program to compute the trigram ($n=3$) probabilities of the following Dr. Suess corpus:***

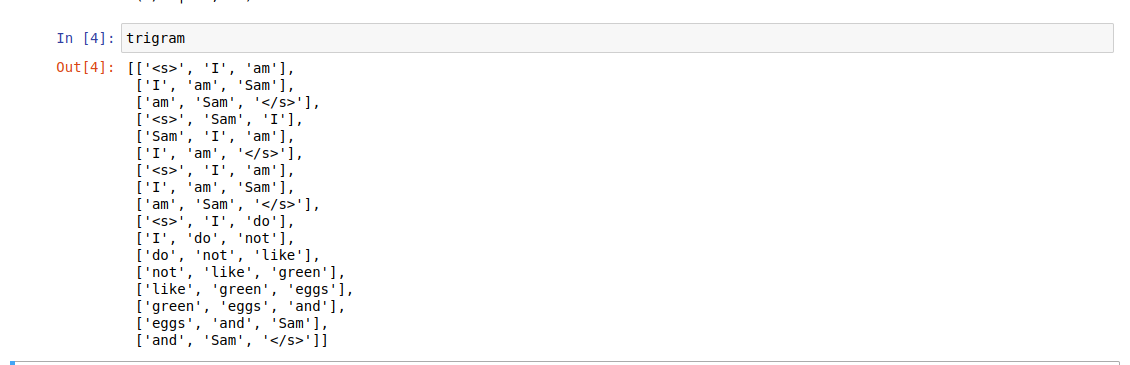
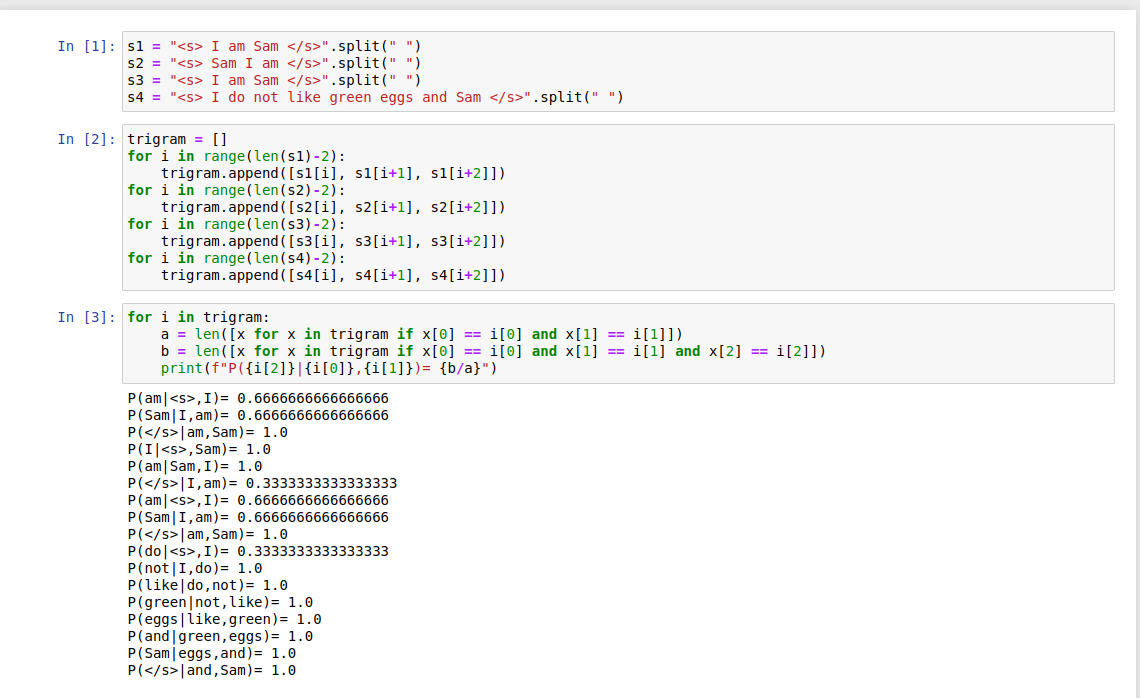
***Hint: See Jurafsky & Martin Chp. 3 for bigram estimation from a similar corpus***

***<s> I am Sam </s>***

***<s> Sam I am </s>***

***<s> I am Sam </s>***

***<s> I do not like green eggs and Sam </s>***

****

1. ***(10 pts Extra Credit) 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 drawn independently***

***Which of these qualifying assumptions does word2vec break (many other LMs do too, as it turns)?***

Word2vec will break the assumption of “ all examples are drawn independently”

***Hint: We make the Markov assumption in language modeling out of necessity; this doesn't mean that it reflects reality!***