ANLY-580-01 : Gabriella Zakrocki
ASSIGNMENT 2

1. 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 feature representation limits our ability to model human language for one main reason: independence. The BOW, which uses a Naïve Bayes approach, assumes the features are independent. This can have a huge issue because in human language we have rely on conditional independence. For example, if I say the word "Abraham" there's a higher probability I will also say the word "Lincoln" in order. Bag of words doesn't consider the order of the word or the effect of the word; the word "it" and the word "hate" have the same level of importance/meaning in classification.

2. 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?
   (b) When N is large, what computational bottle neck arises in word2vec that requires us to change the algorithm?
   Note: we did not tweak the algorithm in the in-class demo, but did mention it during the lecture/demo?

a. Word2vec uses short dense vectors to store words (aka: embedding). The package uses static embeddings which means it learns one fixed embedding for each word displayed in that document's vocabulary. The similarity is captured from the context. Words with similar context end up with similar vectors. The similarity can be calculated using cosine similarity on word vectors. Basically, the word2vec package uses binary classification instead of neural networks to calculate predictions of how often a word shows up near another word. The method uses logistic regression to train a classifier and uses those as weights for the word embeddings.

b. When N is large there becomes an issue with the larger parameter space. This being that the model is going to run extremely slow and that you need a huge amount of training data to tune that many weights to avoid overfitting. The best way to handle this would be to utilize subsampling or negative sampling.

3. Why are the inner product and cosine similarity used to measure similarity and not Euclidean distance?

Cosine similarity is calculated using only the dot product and magnitude of each vector and is therefore affected only by the terms the two vectors have in common, whereas Euclidean has a term for every dimension which is non-zero in either vector. Cosine thus has some meaningful semantics for ranking similar documents, based on mutual term frequency, whereas Euclidean does not. The cosine similarity is advantageous because even if two similar documents are far apart by the Euclidean distance because of the size (like, the word 'cricket' appeared 50 times in one document and 10 times in another) they could still have a smaller angle between them. Smaller the angle, higher the similarity. Basically, angle measure achieves capturing semantic similarity better than distance measure.

4. 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

   Posted in python file titled "assignment2-question4-answer.ipynb"

5. (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)?
   Hint: We make the Markov assumption in language modeling out of necessity; this doesn't mean that it reflects reality!

Word2vec breaks the assumption that examples are drawn independently. This is because with word2vec we have a large unsupervised corpus and for each word in the corpus, we try to predict it by its given context (Continuous BOW), or trying to predict the context given a specific word (skip-gram). This means that they are dependent on each other.