Assignment 2 - Towards Sentiment Analysis - using POS and ngrams to analyze text

You can choose to do this assignment in groups of two, or you can do it independently.

We will continue on exploring a portion of the women's e-commerce clothing reviews now that we have some more sophisticated text analysis techniques.  The full data set can be found here:

[https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews (Links to an external site.)](https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews)

We will consider two files:

[input1.txt](https://canvas.du.edu/courses/122202/files/8064238?wrap=1)[Preview the document](https://canvas.du.edu/courses/122202/files/8064238?wrap=1)

[input2.txt](https://canvas.du.edu/courses/122202/files/8064239?wrap=1)[Preview the document](https://canvas.du.edu/courses/122202/files/8064239?wrap=1)

The first file contains a number of reviews that corresponded to the highest rated reviews.  The second contains a number of reviews corresponding to the lowest ratings.  The files are text files with one review per line.  The review may be (likely is) multiple sentences long.

Your task is to explore these two files and find differences and similarities.  If you did not know which set of reviews was which, could you tell from your analysis??

 I suggest you consider the following (you are free to add in more also) to compare one file to the other:

* Frequent words (in one file versus the other)
* Frequent bigrams/trigrams (in one file versus the other)
* High level POS usage differences (in one file versus the other)
* Most common adjectives, verbs, adverbs, and nouns used (in one file versus the other)
* Differences (in one file versus the other) in bigrams/trigrams that include verbs, adjectives, nouns, adverbs
* The effect of lemmatizing your words before calculating frequencies
* Anything else you can think of that might reveal differences

A few suggestions:

* Do NOT remove stop words before POS tagging.  If you want to remove stop words, first tag, then remove the (word, tag) pairs where the word is a stopword.  And if so, make sure you do it consistently across both sets.
* Given the two files are of very different in size you will want to convert to percentages rather than simply counts.
* After tagging, feel free to map the POS tagset into the universal tagset if you want, it does simplify things considerably.  On the other hand, looking at present tense versus past tense verbs may reveal something also.  That tense information is lost if you map to the universal tagset.
* I recommend using FreqDist( ) to do your counting for you, it is a nice easy to use data structure!  You will likely have many FreqDist( ) objects (for example, one for verbs, one for nouns, etc...)

AGAIN - unlike in many classes where the goal is just to get the code to work, here I am asking you to consider what you are using the code for.  What results make sense and what do not.  What can you learn about the review text by running experiments.

**To turn in:**

a) One or two .py files that illustrate what you have done.  It does not need to cover everything, but should show mostly the techniques you have used.  Comments are your friend.....

b) Your analysis writeup, a pdf file, that includes a description of the experiments you ran, the results from your experiments in prose and in tabular/graph form, and what you deduce from the results.