HW #2 Sentiment Analysis – 30 points (scaled)

Unsupervised Sentiment Analysis

Requirements

- Create the Ipython notebook (named as "yourlastname_firstinitial_
    - HW2_SentimentAnalysis.ipynb") and complete the **tasks** below. **Each cell must be properly numbered and formatted using Markdown**.
- Your Jupyter Notebook needs to be properly formatted using Markdown and comments.
- **Do not print all**. This creates too long of a document. Please use .head on dataframe printouts and limit printing to a few sample lines. For example, avoid using the following code:

```
for i in data:
    print i
```

Instead, you should use the following code (or something similar to print a specific output should it be needed):

```
for i in data[:5]:
    print i
```

- All code must be shown in the Jupiter notebook and you should add a **comment** before or after your code. Use Markdown properly and comments (e.g. #). Use the dataset (moviereviews_HW2.csv) (This dataset comes from IMDB and is all unlabeled movie reviews. Unsupervised sentiment analysis usually uses a lexicon-based approach (positive and negative word lists).

Sections of Homework

1. Use Bing Liu Lexicon for your analysis.
2. Use either Pattern or Textblob for your analysis. You choose the package. If you are using a Mac and neither of these works, put a comment in your homework to that effect.
3. Use **vaderSentiment**.
4. Report the results of the sentiment analyzes for each package in the appropriate section. Then add a few comments where you compare their results. Compare the following aspects of your analysis:
    a. The number of positive reviews
    b. The number of negative reviews

    c. The number of neural reviews
    d. Which method seems to be working better for this movie dataset? Explain why?
5. Use the results of the analysis based on **vaderSentiment**,to conduct word frequency analysis for positive reviews and negative reviews separately. Compare the results using word frequency and word clouds. Here is a method you can use:
    a. First, separate positive and negative reviews
    b. Second, make sure to remove typical stopwords and additional words such as movie, movies, film, films, see, and look. We already know this corpus is about movies so we don't need to see these words in Word Cloud.
    c. Then, perform word frequency analysis and create word cloud for each dataset. For example, what are 50 popular words in positive movie reviews? What are 50 popular words in negative movie reviews?
    d. You may wish to run the analysis again without removing the stopwords. Remember, that Vader attempts to use clues to find reversals in word use. Did this make any difference? This is an optional part of the assignment.

How to Format Your Jupyter Notebook:

- Start with K-State Honor Code "**On my honor, as a student, I have neither given nor received unauthorized aid on this academic work**."
- Must be professional and neat

Submission

- Complete Ipython notebook in **HTML** version (yourlastname_firstinitial_ HW_SentimentAnalysis.html) attached to the CANVAS assignment.