CSCI 1100 — Humanities Computer Science 1

Homework 6

Parsing Files & Sets

Reading: Algorithms of Oppression, pages 1-24, by Safiya Noble

---

OVERVIEW

This homework is worth 100 points total toward your overall homework grade and is due Thursday, **April 4th**, 2019 at 11:59:59 pm. There is only one part to this homework, so the files you will submit for the homework will be:

```
hw6Part1.py
README.txt
```

In this assignment, we'll perform low-level sentiment analysis of articles about current events.

This will use two text files that contain lists of positive and negative words, as determined by Bing Liu, Minqing Hu, and Junsheng Cheng in their research that can be found at www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

The goal of this assignment is to parse an article text file and compare the words in the article with those in the positive and negative word files.

This assignment should be implemented using sets. Your sets should compare the two files with the article text, but the article text can be stored using any of the previously covered material. You will then perform a basic analysis of this data.

Information on the article must include:

1. Headline
2. Author
3. Source
4. Length of article (words)

Analysis of the article must include the following:

1. All of the positive/negative word matches in the article
2. Number of occurrences of each positive/negative word in article
3. Number of occurrences of positive/negative words in article
4. Percentage of words in article that are unique
5. Percentage of all words in article that are positive/negative
6. Percentage of unique words in article that are positive/negative

7. If there positive/negative word matches in the headline

We have provided three lengthy text files for you to test your code on, but we encourage you to go beyond these and test your code with articles you find.

Here are some notes on stripping extraneous characters:

To strip the extraneous characters in the articles, you can use the 're' Python module. Therefore, you must import re at the beginning of the program, and then to use the following code to strip the extraneous characters:

```
#eliminate any extraneous characters from a single word using re import
            regex = re.compile('[^a-zA-Z]')
            word = regex.sub('', word)
#word will now be a stripped word, that can be appended to a list or
other structure that is
#keeping track of all the words in the article
```

If you have any further questions on the re module, try to research it further or ask a mentor. You can also use any other Python modules or functions that you want.

README: In this homework, we will introduce you to some critical deconstruction of algorithmic techniques, in terms of identifying their strengths, weaknesses, and potential impact (or lack thereof) on different communities.

After reading the introductory 24 pages of Safiya Noble's "Algorithms of Oppression," reflect upon the coding portion of the assignment and the sentiment analysis model developed by Liu et. al.. Then, in a README File, address the following questions in two or three paragraphs each. (NOTE: your grading for each question will be determined by your appropriate completion of the question, not on how the grading team determines the "correctness" of your response.)

1: Why do you think Liu et. al. choose to use product reviews as the dataset for their sentiment analysis testing? How does that kind of dataset differ from the datasets that Safiya Noble addresses in her chapter?

2: In what ways could this kind of sentiment analysis be used by companies like Google to address race and gender-based violence in their systems?

3: What are kinds of bias that this sentiment analysis would fail to pick up? In your response, consider some of the overlapping kinds of "texts" that are a part of digital objects, such as readable text, metatags, images, and links.

Other notes:

We will not be providing solutions to this homework, just examples of how we would like it to be formatted. This is done to mimic what it would be like if you actually did this program on your own, and we can discuss this idea further in lab/office hours.

For this homework, you will be graded on:

- Implementing the homework using sets
- Correctly reading in and parsing provided files
- Calculating the seven statistics stated above
- Readability and formatting your output
- Style (commenting, code structure, variable names)
- Appropriately completing the README questions (i.e., making an honest effort to answer each question)