CASS
PROJECT ORCAS

7/25/2019

REQUIREMENTS

1. The Sentiment of a Comment

The output sentiment should be a continuous variable in the range from -1 to 1.

Where values below 0 represent a negative sentiment. Values above 0 represent a positive sentiment.

2. Emotions

The emotions of each word in a comment will be tallied up and included in the output.

3. Edge Cases

The project should account for edge cases in which there are no words in the lexicon.

4. Repeatable Words

The project should count the number of unique words in a comment, therefore a single word repeated multiple times doesn’t influence the overall sentiment.

5. Reading in the Data

The project given a filename of a excel spreadsheet, will extract the data from the comment column.

6. Writing out the Data

The project should output the results of the sentiment and other data related to the emotions to a new file.

7. Spell Check

The project should run the comments through a spell check, and make the spell changes in order to get a better estimate to the correct sentiment.

APPROACH

1. Reading in the Data

The initial approach to reading in the lexicon, is given the lexicon filename as a command line argument, the file is parsed and creates a “dictionary” (hash table) that maps the word to the sentiment values.

The initial approach to reading in the data, is given the data filename as a second command line argument, the file is then read into the program and the comments are parsed into a list.

2. Writing out the Data

The initial approach to writing out the result data, is that given a third command line argument, the filename of the wanted result file, the data will be written to the new file. The results will include an overall sentiment on the scale from -1 to 1 as well as information on emotions used in the comment.

3. Pre-processing the Comment

The initial approach to pre-processing the comments, involves the comments being stripped of their punctuation, then then shifted to all lowercase letters, and finally split by the white space between the words. The words are then sent through a spell check to correct misspelled or unknown words. The words are then represented as a “dictionary” (hash table) that maps the word to the number of times it appeared in the comment.

4. Determining Sentiment

The initial approach to determining sentiment, involves the process of going through each unique word of a comment, and using that word as a key to look up its sentiment and emotion values in the given lexicon. Using these values for all the words in the comments, tallies are totaled, and a sentiment is calculated in the following formula:

(number of unique positive words – number of unique negative words) / (number of unique words).

Logically this works because if all the words are negative words the resulting sentiment would be -1.

Vice versa, if all the unique words are positive words the resulting sentiment would be 1.

5. Output format

The initial approach to the output format, involves displaying the comment, the overall sentiment, and number of emotions that appeared in the comment.