Hausarbeit:

Ich versteh' nur Bahnhof

or

I loved the film "Plan 9 from Outer Space" !!!

February 3, 2015

Your final task is to write a Python program. You have the option of whether you want the program to do:

1. **Language Identification**: guess whether an input is either in German or English
2. **Sentiment Analysis**: guess if a movie reviewer liked a movie or didn't like it

Both programs will have similar structure. They will open and read-in an input file, convert every line to lowercase, and count the frequency of certain words. For language identification, count the frequency of: "the", "and", & "of", as well as "die", "der", & "und". For sentiment analysis, count the frequency of: "good", "great", & "like", as well as "bad", "terrible", & "hate".

After you've read-in the text file and tabulated all the counts of these words, compare the total count of the English words with the total count of the German words (for the language identification task), or compare the total count of the positive words with the total count of negative words (for the sentiment analysis task).

Then print out which of the two choices had the most word occurences. So if you are working on the language identification task, at the end of the file the program should print either "English" or "German". Thus you are guessing whether the input file is written in English or German. If you are working on the sentiment analysis task, at the end of the file the program should print either "Positive" or "Negative" Thus you are guessing whether the movie review is positive (the reviewer liked the movie), or negative (the reviewer did not like the movie).

From the command-line shell, you should ultimately be able to type:

```
python3 hausarbeit.py input.txt
```

And the output should simply be either "German" or "English" (for the language identification task), or either "Positive" or "Negative" (for the sentiment analysis task).

You can download sample input files from:
http://languagemodel.org/classes/uds/shell_and_python_basics/hausarbeit/sample_input

You can (but are not required to) work with **one** other person in the class on this (no more than one). **Be sure to read and follow all steps**, and include lots of useful comments in the script.[1] Email Jon the final python script. In the email affirm that you (and possibly one other person in the class, whom you should name) are the sole author(s) of the program.

Also state in the email how you would improve the program to get more accurate results. You don't need to implement these proposed improvements.

Example Command-line Usage for Language Identification

```
$ python3  hausarbeit.py  language_input1.txt
German
$ python3  hausarbeit.py  language_input2.txt
English
$ python3  hausarbeit.py  language_input3.txt
English
```

Example Command-line Usage for Sentiment Analysis

```
$ python3  hausarbeit.py  movie_input1.txt
Positive
$ python3  hausarbeit.py  movie_input2.txt
Positive
$ python3  hausarbeit.py  movie_input3.txt
Negative
```

Command-line Arguments in Python

In order for your python program to work with the file name that you specify (as above), you should have the following line at the beginning of your python program:

`import sys`

Then you can use the first command-line argument as `sys.argv[1]`. For example:

```
import sys
myfile = open(sys.argv[1])
for line in myfile:
    print(line)
myfile.close()
```

[1] Start a line with the `#` symbol, followed by your useful comments.