

Hausarbeit:

Ich versteh' nur Bahnhof

January 13, 2016

Your final task is to write a Python program, and then write about it. Each part is discussed below. For the program (only the program), you can (but are not required to) work with **one** other person in the class on this (no more than one). There should be two email attachments: your term paper (*your_name_hausarbeit.pdf*) and your program (*your_name_langid.py*). Include the program as an appendix to the paper, **and** as a separate email attachment.

If you have any questions, feel free to email me.

1 Program

Read and follow all steps, and include lots of useful comments in the program.¹ The python program should perform language identification – guess whether an input text is either in German or English

Your program should open and read-in an input file, convert every line to lower-case, and count the frequency of certain words. For language identification, count the frequency of: “the”, “and”, & “of”, as well as “die”, “der”, & “und”.

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.

Then print out which of the two choices had the most word occurrences. So 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.

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

```
python3 name_langid.py input.txt
```

And the output should simply be either “German” or “English”.

You can download sample input files from:

http://languagemodel.org/classes/uds/shell_and_python_basics/hausarbeit/sample_input

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

1.1 Example Command-line Usage for Language Identification

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

1.2 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()
```

2 Paper

First, explain at a very high level what your program does and how it works. Explain it to an audience who does not know how to program.

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

Next, explain how it can be helpful as a translator to know how to process and modify large amounts of text. List several situations where it could help the job of a translator.

Next, do you think the job of a translator will ever be fully replaced by machine translation? Why or why not. Explain, with reasons and with examples.

Affirm that you (and possibly one other person in the class, whom you should name) are the sole author(s) of the program. You should also affirm that you are the only author of the paper.

The paper should be normal length and normal margins & fonts.