Lab 2: Lists

Due date: Monday February 8 2020, 3:05pm

Assignment

1. The main task in this lab is to separate text into lists and do do some statistics on these lists. The data for this lab needs to be downloaded from

   `https://www.gutenberg.org/ebooks/search/?sort_order=downloads`

   Pick a few books you like and **report** (you will need to download the textfile (plain text UTF-8)

   - **The average sentence length.** [use the function in the cheatsheet to read the book in, do not worry about some special cases like titles, sections, numbers etc] To calculate the average sentence length the easiest way is to read the book as a long string, and then split the string on the '.' into a list and then use a for loop over all sentences to split on ' ' and then count the length of that list (`len(list)`) and sum these length up, once all sentences are summed up, divide by the number sentences and print the average length.
   - **The average word length.** The loop from above can be extended that when you split the sentence into words, split each word into characters (using `list(string)` and the also sum these up and report at the end the average word lengths.
   - **How many 1-letter words, 2-letter words, ... 20-letter words are present.** Create an array c with 21 zero entries, every time you encounter a word get the length of the word which can be used as an index into c and you can then augment $c[\text{index}] += 1$. At the end print the c which now contains all counts for the words.
   - **What are the 100 most common words [use a dictionary];** Create a dictionary, fill in the words as keys and the original value for each word is 1; then loop over all words and test wether the word is already a key or not and if not create a new dictionary entry, if yes add 1 to the value.
   - **Compare the word usage of two books** of the same author and two books of different authors.

2. Upload the python script and the output as a zip to Canvas before the deadline. Your output should contain the book titles and the statistics.

Cheat sheet

contains fragments of code that you will need to finish the Lab 2. the real code for the read function can be found in Canvas in the codes examples.

```
#!/usr/bin/env python
# (c) yourname
#---------------------------------------------------------
#
#
# reads the book into a string
def readbook(filename, clean=True):
   ’’’
   Reads a textfile into a single string, ignoring line breaks.
   The function has a single argument: the infile name
   and it returns a single string containing the whole file,
   if the variable clean is True then replace all upper case
   letters with lowercase and also remove ",;:-’
   ’’’
   with open(filename, ’r’) as myfile:
      data=myfile.read().replace(’\n’, ’’).replace(’\r’,’’).replace(’\t’,’ ’)
   if clean:
      data = data.replace(’;’,’ ’).replace(’,’,’ ’).replace(’:’,’ ’)
      data = data.replace(’-’,’ ’).replace(’"’,’’).replace("’","").lower()
   return data

#
# reads all sentences from a text into a list
def get_sentence(text):
   ’’’
   reads all sentences from the text (using a ’.’ as the end of a sentence
   ’’’
   pass  #replace pass with the code that splits the text into a list

#
# count words in sentence
def count_words(listofsentences):
   ’’’
   loops over all setences and splits each setence, counts words per setence
   and returns the average number of words per sentence
   ’’’
   pass  #replace pass with your code

#
# count characters in word
def count_chars_in_word(listofwords):
   ’’’
   loops over all words in a sentence and counts characters in a word
   sums them up and returns the average number of words and also the
   number of words of length 1,2,3,4..,20
   ’’’
   pass



if __name__ == ’__main__’:
    data = readbook(’testbook.txt’)
    # more stuff to follow
```