

Introduction to Web Science

Assignment 6

Prof. Dr. Steffen Staab

staab@uni-koblenz.de

René Pickhardt

rpickhardt@uni-koblenz.de

Korok Sengupta

koroksengupta@uni-koblenz.de

Institute of Web Science and Technologies

Department of Computer Science

University of Koblenz-Landau

Submission until: December 6, 2016, 10:00 a.m.

Tutorial on: December 9, 2016, 12:00 p.m.

Please look at the lessons 1) **Simple descriptive text models** & 2) **Advanced descriptive text models**

For all the assignment questions that require you to write code, make sure to include the code in the answer sheet, along with a separate python file. Where screen shots are required, please add them in the answers directly and not as separate files.

Team Name: mike

Group members: Shohel Ahamad, Anish Girijashivaraj, Slobodan Kocevski

1 Digging deeper into Norms (10 points)

You have been introduced to the concept of a norm and have seen that the uniform norm $\|\cdot\|_\infty$ fulfills all three axioms of a norm which are:

1. Positiv definite
2. Homogeneous
3. Triangle inequality

Recall that for a function $f : M \rightarrow \mathbb{R}$ with M being a finite set¹ we have defined the L_1 -norm of f as:

$$\|f\|_1 := \sum_{x \in M} |f(x)| \quad (1)$$

In this exercise you should

1. calculate $\|f - g\|_1$ and $\|f - g\|_\infty$ for the functions f and g that are defined as
 - $f(0) = 2, f(1) = -4, f(2) = 8, f(3) = -4$ and
 - $g(0) = 5, g(1) = 1, g(2) = 7, g(3) = -3$
2. proof that all three axioms for norms hold for the L_1 -norm.

1.1 Hints:

1. The proofs work in a very similar fashion to those from the uniform norm that was depicted in the videos.
2. You can expect that the proofs for each property also will be "three-liners".
3. Both parts of this exercise are meant to practice proper and clean mathematical notation as this is very helpfull when reading and understanding research papers. Discuss in your study group not only the logics of the calculation and the proof (before submission) but try to emphasize on the question whether your submission is able to communicate exactly what you are doing.

Answer:

¹You could for example think of the function measuring the frequency of a word depening on its rank.

Ans 01: Given $f(0) = 2$, $f(1) = -4$, $f(2) = 8$, $f(3) = -4$
and $g(0) = 5$, $g(1) = 1$, $g(2) = 7$, $g(3) = -3$

$$\begin{aligned} * \|f-g\|_1 &= \sum_{i=0}^4 |f_i - g_i| \\ &= (|2-5| + |-4-1| + |8-7| + |-4+3|) \\ &= 3 + 5 + 1 + 1 = 10 \text{ Ans} \end{aligned}$$

$$\begin{aligned} * \|f-g\|_\infty &= \max(|f_1 - g_1|, |f_2 - g_2|, \dots, |f_n - g_n|) \\ \text{Here } n &= 4. \\ \therefore &= \max(|2-5|, |-4-1|, |8-7|, |-4+3|) \\ &= \max(3, 5, 1, 1) \\ &= 5 \text{ Ans} \end{aligned}$$

Positive Definition:

$$\|f\|_{\infty} = 0 \Leftrightarrow \sup_{n \in M} \|f(n)\|_R = 0$$

$$\Rightarrow \|f(n)\|_R = 0 \forall n$$

$$\Rightarrow f(n) = 0 \forall$$

$$\therefore f(n) = 0 \forall$$

$$f = 0.$$

again

$$\|g\|_{\infty} = 0 \Leftrightarrow \sup_{n \in M} \|g(n)\|_R = 0$$

$$\Rightarrow \|g(n)\|_R = 0 \forall n$$

$$\Rightarrow g(n) = 0 \forall$$

$$\therefore g = 0.$$

As a result for all those combinations the results will be 0.

$$\therefore \|f - g\|_{\infty} = 0 \Leftrightarrow \sup_{n \in M} \|f(n) - g(n)\|_R = 0$$

$$\Rightarrow \sup_{n \in M} \|f(n)\|_R - \sup_{n \in M} \|g(n)\|_R = 0$$

$$\Rightarrow \|f(n)\|_R - \|g(n)\|_R = 0 \forall n$$

$$\Rightarrow f(n) - g(n) = 0 \forall$$

$$\Rightarrow f - g = 0 \underline{\underline{\text{sum}}}$$

Homogeneous: The condition will be.

$$\|a(f-g)\|_1 = a\|f-g\|_1 \left[\begin{array}{l} \text{where "a" is} \\ \text{a variable} \end{array} \right]$$

Let, $a = 2$, then,

$$\begin{aligned} & \|2(f-g)\|_1 \\ &= \left[|(2*2) - (5*2)| + |(-4)*2 - (1*2)| + |(8*2) - (7*2)| \right. \\ & \quad \left. + |(-4*2) + (9*2)| \right] \\ &= 6 + 10 + 2 + 2 = 20. \end{aligned}$$

On the other hand we proved that $\|f-g\|_1 = 10$

$$\therefore 2\|f-g\|_1 = 10*2 = 20 \text{ (proved).}$$

Triangle inequality: The condition will be.

$$\|f+g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$$

$$\|f+g\|_\infty \leq \sup_{n \in \mathbb{N}} \|f(n) + g(n)\|_{\mathbb{R}}$$

$$15 \leq \sum_{i=0}^4 f_i + \sum_{i=0}^4 g_i$$

$$15 \leq 2 + 10$$

$$15 \leq 12 \text{ (proved)}$$

2 Coming up with a research hypothesis (12 points)

You can find all the text of the articles from Simple English Wikipedia at <http://141.26.208.82/simple-20160801-1-article-per-line.zip> each line contains one single article.

In this task we want you to be creative and do some research on this data set. The ultimate goal for this exercise is to practice the way of coming up with a research hypothesis and testable predictions.

In order to do this please **shortly**² answer the following questions:

1. What are some observations about the data set that you can make? State at least three observations.
2. Which of these observations make you curious and awaken your interest? Ask a question about why this pattern could occur.
3. Formulate up to three potential research hypothesis.
4. Take the most promising hypothesis and develop testable predictions.
5. Explain how you would like to use the data set to test the prediction by means of descriptive statistics. Also explain how you would expect your outcome.

(If you realize that the last two steps would not lead anywhere repeat with one of your other research hypothesis.)

2.1 Hints:

- The first question could already include some diagrams (from the lecture or ones that you did yourselves).
- In step 3 explain how each of your hypothesis is falsifiable.
- In the fifth step you could state something like: "We expect to see two diagrams. The first one has ... on the x-axis and ... on the y-axis. The image should look like a ... The second diagram ...". You could even draw a sketch of the diagram and explain how this would support or reject your testable hypothesis.

Answer:

1. Observations:
 - a) **There are different spellings of the same word in different article.Simple English Wikipedia uses a native words from American and British English.**

²Depending on the question shortly could mean one or two sentences or up to a thousand characters. We don't want to give a harsh limit because we trust in you to be reasonable.

- b) The article provide information about scientific topics
- c) The articles are separated from each other by a new line.

"There are different spellings of the same word in different article.Simple English Wikipedia uses a native words from American and British English.." This observation make us curious.

2. Question

- a) Is it true that the spelling of the same word is different in different articles?
- b) What makes one language popular?
- c) Why the content on the web is written in American English spelling?
- d) How to measure popularity of a language?

3. Hypothesis.

- a) Simple English Wikipedia has more then 100000 articles
- b) SEW has more than 50% of its articles for IT science.
- c) **Simple English Wikipedia uses more words from American English rather than British English**

For the first hypothesis, It will be falsifiable if we get less then 1000000 articles after counting the appearances of the sign for new line.

For second hypothesis it will be falsifiable if there is no information that is related with the keywords in the IT sector.

For third hypothesis we can say that it is falsifiable if the same word that is used in different articles is written the same.

4. Testable predictions.

We have to find out how are spelled the same words in British English and how they are spelled in American English. What is the difference in the spelling between this two languages.

-Example of counting the words.

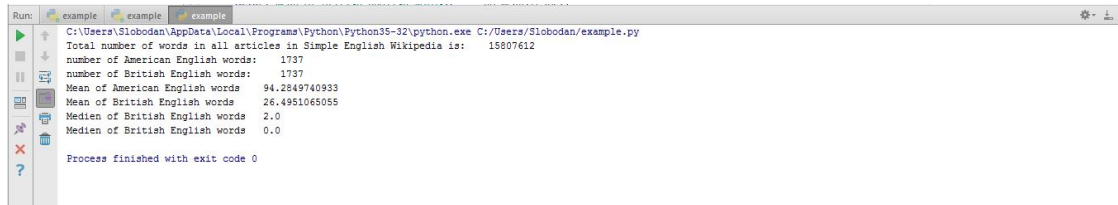
- 5. First, We are going to get a list of all the words that have the same meaning but are spelled differently on American English and British English.After that we compare each word from the list with each word from each article. We will save the number of matching words in all the articles.

We expect to see one diagram which is a Cumulative Frequency Diagram(CDF) which has % of American English and British English words in x-axis and the frequency in y-axis. If this diagram shows that the number of American English words is greater than the number of British English words, our hypothesis will be

proved.

```
1: import numpy as np
2:
3: f1=open("SEWcontent.txt", "r", encoding="utf-8")
4: m = f1.read()
5: splittedContent = m.split(' ')
6:
7: print("Total number of words in all articles in Simple English Wikipedia is: \t ")
8: AE = open("us wordss.txt", "r", encoding="utf-8")
9: aew = AE.read()
10: splittedAE = aew.split('\n')
11: print("Total number of American English words: \t " + str(len(splittedAE)))
12:
13: BE = open("be words.txt", "r", encoding="utf-8")
14: bew = BE.read()
15: splittedBE = bew.split('\n')
16: print("Total number of British English words: \t " + str(len(splittedBE)))
17:
18: def Countwords(splA):
19:     lst=list()
20:     for j in range(0, len(splA)):
21:         if(j%400==0):
22:             print("Word number:" + str(j) + " / " + "" + str(len(splA)) + "\t ")
23:             x = (str(m)).count(splA[j])
24:             lst.append(x)
25:     return lst
26:
27:
28: resAE = Countwords(splittedAE)
29: resBE = Countwords(splittedBE)
30:
31: print("American words found in SEW:" + str(resAE))
32: print("British words found in SEW:" + str(resBE))
33:
34: from matplotlib import pyplot as plt
35:
36: print("Mean of American English words\t ", np.mean(resAE))
37: print("Mean of British English words\t ", np.mean(resBE))
38:
39: print("Medien of British English words\t ", np.median(resAE))
40: print("Medien of British English words\t ", np.median(resBE))
41:
42: plt.hist(resAE, bins = np.arange(0, 1300, 5), normed=True, color='yellow', label="A")
43: plt.hist(resBE, bins = np.arange(0, 1300, 5), normed=True, color='gray', label="B")
44: plt.title("Comparing of the usage of words from American and British English")
45: plt.xlabel("Number of words")
46: plt.ylabel("Frequency")
47: plt.legend( shadow=True, fancybox=True)
```


48: `plt.show()`



The screenshot shows a Python IDE console window with the following output:

```
Run: example example example
C:\Users\Slobodan\AppData\Local\Programs\Python\Python35-32\python.exe C:/Users/Slobodan/example.py
Total number of words in all articles in Simple English Wikipedia is: 15807612
number of American English words: 1737
number of British English words: 1737
Mean of American English words 94.2849740933
Mean of British English words 26.4951065055
Median of British English words 2.0
Median of British English words 0.0
Process finished with exit code 0
```

3 Statistical Validity (8 points)

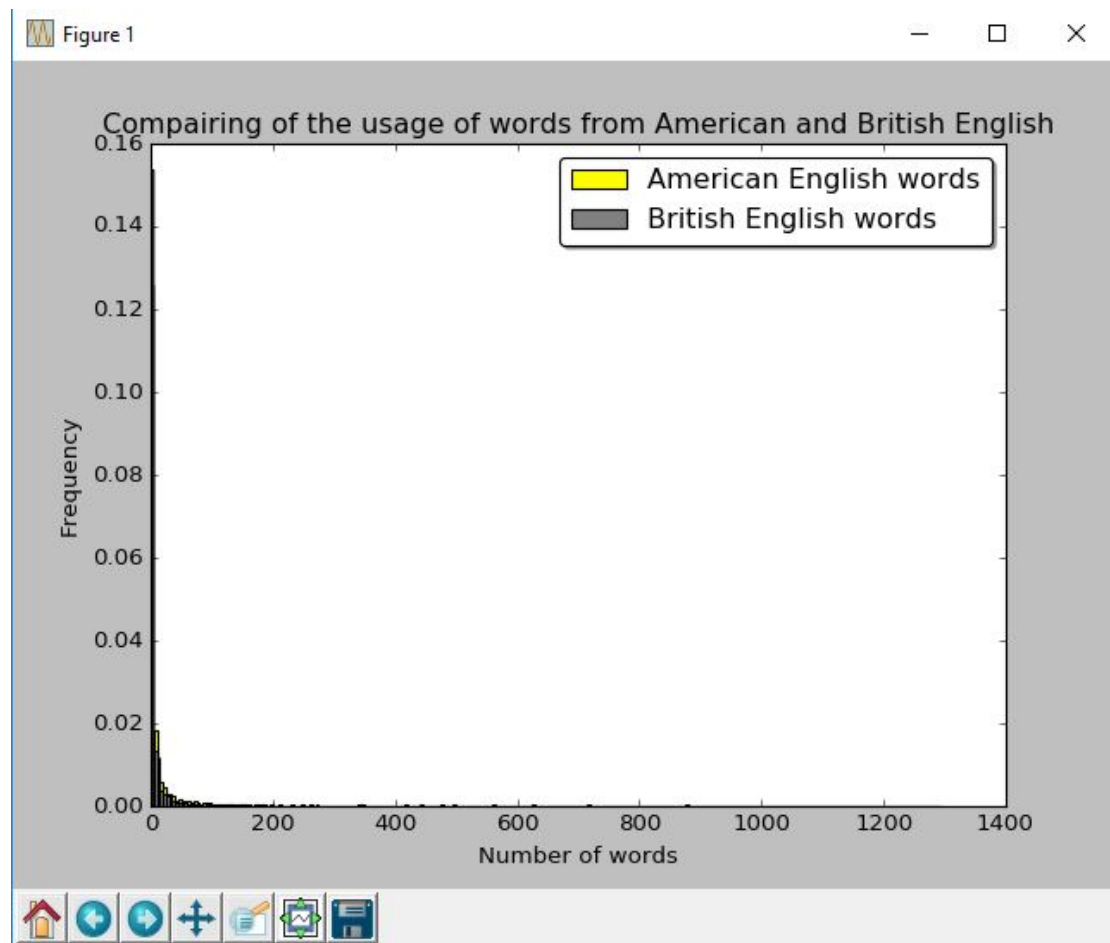
In the above question, you were asked to formulate your hypothesis. In this one, you should follow your own defined roadmap from task 2 validate (or reject) your hypothesis.

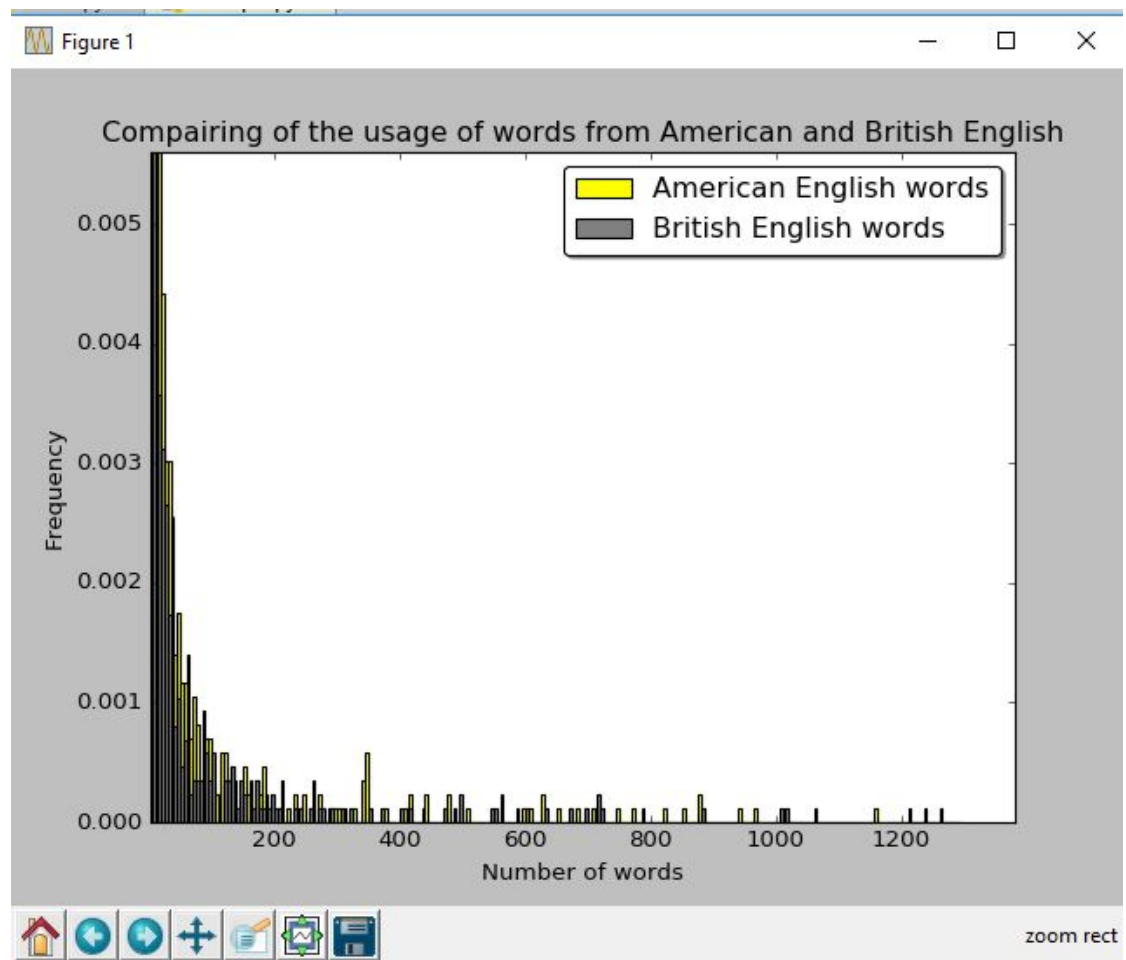
3.1 Hints:

- In case feel uncomfortable to test one of the predictions from task 2 you can "steal" one of the many hypothesis (and with them implicitly associated testable predictions) or diagrams depicted from the lecture and reproduce it. However in that case you cannot expect to get the total amount of points for task 3.

Answer: In this process we use 1737 native British and 1737 native American English words that have the same meaning but are spelled differently. After calculation we found that the mean of the American English word in the given articles is 94.2849740933. Which means every article have 94.29% of its words written on American English, while the mean for the British English words is 26.4951065055. After seeing the results, it is obvious that Simple English wikipedia uses more Words from American English language than British English language.

Answer:





Important Notes

Submission

- Solutions have to be checked into the github repository. Use the directory name `groupname/assignment6/` in your group's repository.
- The name of the group and the names of all participating students must be listed on each submission.
- Solution format: all solutions as *one* PDF document. Programming code has to be submitted as Python code to the github repository. Upload *all* `.py` files of your program! Use **UTF-8** as the file encoding. *Other encodings will not be taken into account!*
- Check that your code compiles without errors.
- Make sure your code is formatted to be easy to read.
 - Make sure you code has consistent **indentation**.
 - Make sure you comment and document your code adequately in English.
 - Choose consistent and intuitive names for your identifiers.
- Do *not* use any accents, spaces or special characters in your filenames.

Acknowledgment

This latex template was created by Lukas Schmelzeisen for the tutorials of "Web Information Retrieval".

LA_TE_X

Currently the code can only be build using **LuaLaTeX**, so make sure you have that installed. If on Overleaf, there's an error, go to settings and change the **L**A_TE_Xengine to **LuaLaTeX**.