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: Echo

1 Digging deeper into Norms (10 points)

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

1. Positiv definite
2. Homogeneous
3. Triangle inequality

Recall that for a function $f : M \longrightarrow \mathbb{R}$ with M being a finite set[1] we have defined the L_1-norm of f as:

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

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
   - $g(0) = 5, f(1) = 1, g(2) = 7, g(3) = -3$
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.

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

Answer:

We know that,

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

So for,

$$||f - g||1 = ||f - g||1 := \sum_{x \in M} |f(x) - g(x)|$$

$$= |2-5| + |-4-1| + |8-7| + | -4+3|$$

$$= 3+5+1+1$$

$$= 10$$

And,

$$||f - g||\infty = \sum \max \{ |f - g|: i....n\}$$

$$= \max \{ |2-5| , |-4-1| , |8-7| , | -4+3|\}$$

$$= \max \{ 3, 5, 1, 1 \}$$

$$= 5$$

Positive Definite:

To prove the axioms positive definite, it has to be proven that,

If, $||f||\infty = 0$ then $f = 0$

For L-1 norm it will be if $||f||_1 = 0$ then $f = 0$

Here we have to prove that,

if, $||f - g||_1 = 0$ then $f - g = 0$

Normally we take the summation of the function values.

Here, if $||f - g||_1 = 0$, it we can assume that if the summation of all the values which are non - negative is 0 then the values for the function are also 0.

So,

$$f = 0$$

(Proved)

Homogeneous :

$||\alpha(f - g)||1 = \alpha ||f - g||1$

Let the value be α=5, then

L.H.S (Left hand side)

$$||5f - 5g||1 = |2*5 - 5*5| + |-4*5 - 1*5| + |8*5 - 7*5| + | -4*5 + 3*5|$$

$$= |10 - 25| + | -20 - 5| + |40 - 35| + | -20 + 15|$$

$$= 15 + 25 + 5 + 5 = 50$$

R.H.S (Right hand side)

$$\alpha ||f - g||1 = 5 ||f - g||1$$

$$= 5*10 = 50$$

(Proved)

Triangular Inequality:

$||f + g||_1 \leq ||f||_1 + ||g||_1$

L.H.S $||f + g||_1$

$= \sum |f(x) + g(x)|$

$= |2+5| + |-4+1| + |8+7| + |-4-3|$

$= 7+ 3+ 15 + 7$

$= 32$

R.H.S $= ||f||_1 + ||g||_1$

$= \sum |f(x)| + \sum |g(x)|$

$= (2+ 4+ 8+ 4) + (5+1+ 7+ 3)$

$= 18 + 16$

$= 34$

Here, $||f + g||_1 < ||f||_1 + ||g||_1$

(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 obervations about the data set that you can make? State at least three obervations.
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 potentiel research hypothesis.
4. Take the most promesing 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:
After observing the first 10-20 articles in the dataset:

[2]Depending 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.

- In this dataset, computer science terms occurances are so less.
- In this dataset, each article has a specific topic
- This dataset contains more articles with positive affect than negative
- This dataset does not contain any non-english words

2.Explain One observation:
In this dataset, computer science terms occur so less.
To find the average percentage of computer science terms occurance in all articles in the dataset is interesting because then we know how frequent it is to have computer science terms distributed in different kinds of article topics.

3. Three Hypothesises FOR "In this dataset, computer science terms occurances are so less"

- Computer science words occurances are less than 30% in all articles in the dataset
  Falsible: Computer science words occurances are more than 30% in all articles of the dataset
- Computer science words occurances are less than 30% in each article in terms of all words in each article
  Falsible: Computer science words occurances are more than 30% in each article in terms of all words in each article
- Computer science related articles are less than 20% in the dataset than other topics.
  Falsible: Computer science related articles are more than 20% in the dataset

4.Testable prediction FOR "1. Computer science words occurances are less than 30% in all articles in the dataset"
(a) The average percentage of the computer science terms (the list given in computer.txt) in the dataset will be less than (30%) The percentage that we are predicting of computer science terms in the dataset is less than 30%. In our case we have chosen to check the distribution of computer science terms in the dataset, and that is justified by counting the number of computer science terms occurances in each article over the number of all

words in the article. Then sum all the percentages.

5.Plan:

- We will go through each article, find the average amount of computer science terms (from the list in computer.txt) in each article.

- Collect in a dictionary the average amount of computer science words in each article, then sum up all percentages to find the average percentage of the occurances of computer science terms in all the dataset

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 imlicitly 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.

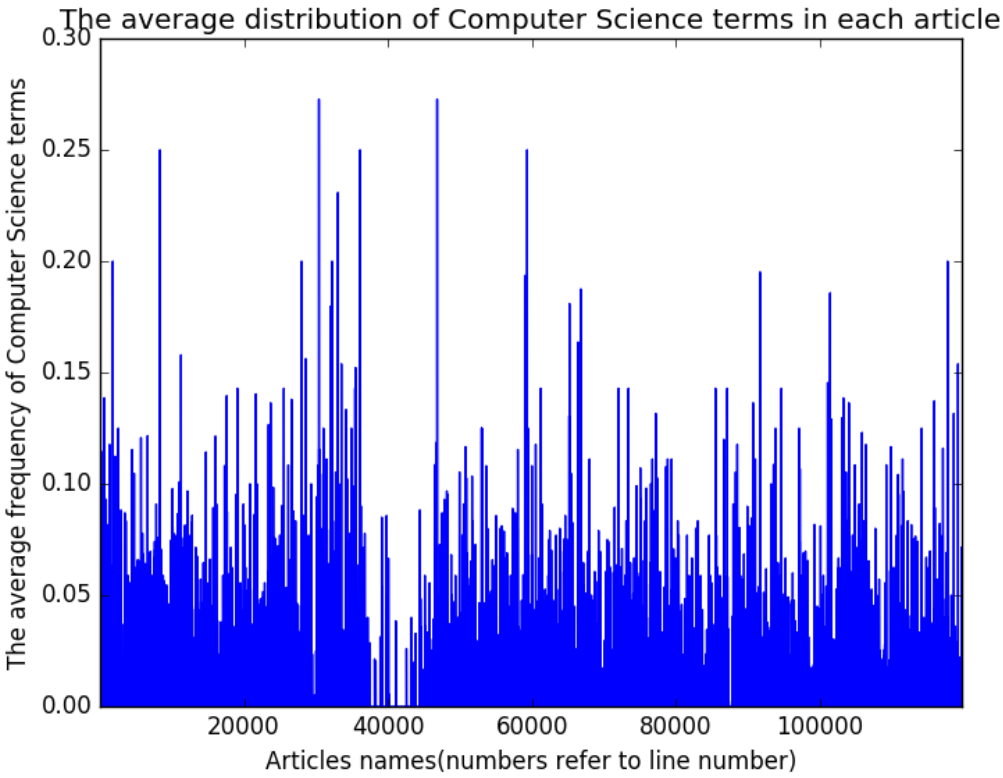
Answer: Echo_assignment6_3.py

```
import time
import pandas as pd
import matplotlib.pyplot as plt


def readFile(x):
    lines = []
    for line in open(x):
        lines.append(line)
    return lines

computer = readFile('computer.txt')
# using a sample of 500 articles only
articles = readFile('articles')

computerwords = []
for computerwrds in computer:
    computerwords.append(computerwrds.split('\n')[0])

start_time = time.time()
dictionry = {}
wordscomp = {}
num = 1
for article in articles:
    matchedwords = []
    allwords = len(article.split())
    for word in article.split():
        l = filter(lambda c: c.isalpha(), word)
        if l in computerwords:
            matchedwords.append(l)
    dictionry[num] = len(matchedwords)
    try:
        wordscomp[num] = float(len(matchedwords))/allwords
```

```
34:     except:
35:         pass
36:     num += 1
37: print("--- %s seconds ---" % (time.time() - start_time))
38:
39: start_time = time.time()
40: fd = pd.Series(wordscomp)
41: print 'The average amount of computer science terms in all articles in the data se
42: fd.plot()
43: plt.title('The average distribution of Computer Science terms in each article')
44: plt.ylabel('The average frequency of Computer Science terms')
45: plt.xlabel('Articles names(numbers refer to line number)')
46: plt.show()
47: print("--- %s seconds ---" % (time.time() - start_time))
```

assignment6.py has some statistics to validate that the hypothesis of (Computer science words occurances are less than 30% in all articles in the dataset) since the average amount of computer science terms in all articles in the data set was : 0.0014686982286

```
C:\Python27\python.exe C:/Users/Hanadi/Desktop/Uni/WS1617/Intro/Echo/Echo/assignment6/task2/assignment6.py
--- 125.657999992 seconds ---
The average amount of computer science terms in all articles in the data set:  0.0014686982286
--- 51.7120001316 seconds ---

Process finished with exit code 0
```


The average distribution of Computer Science terms in each article
The average frequency of Computer Science terms
0.30
0.25
0.20
0.15
0.10
0.05
0.00
20000
40000
60000
80000
100000
Articles names(numbers refer to line number)

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".

LaTeX

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 LaTeXengine to `LuaLaTeX`.