Introduction to Web Science

Assignment 6

Prof. Dr. Steffen Staab   René Pickhardt

staab@uni-koblenz.de   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.

Delta Group Members

Oana Dumitrasc
odumitrasc@uni-koblenz.de

Alisa Becker
alisabecker@uni-koblenz.de

Omar Aly
oaly@uni-koblenz.de

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.

Answer 1:
Calculate $||f - g||_1$

$$||f - g||_1 = |2 - 5| + |-4 - 1| + |8 - 7| + |-3 - 4| \tag{2}$$

$$||f - g||_1 = |-3| + |-5| + |1| + |-7| \tag{3}$$

$$||f - g||_1 = 3 + 5 + 1 + 7 = 16 \tag{4}$$

Calculate $||f - g||_\infty$

$$||f - g||_\infty = max\{|f_1 - g_1|, ..., |f_n - g_n|\} \tag{5}$$

$$||f - g||_\infty = max\{|2 - 5| + |-4 - 1| + |8 - 7| + |-3 - 4|\} \tag{6}$$

---

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

$$||f - g||_\infty = max\{|-3| + |-5| + |1| + |-7|\} = 7 \tag{7}$$

Answer 2:   Too prove for positive definite:

$$||f||_1 = 0 -> f = 0 \tag{8}$$

Proof:

$$||f||_1 = 0 <-> \sum |f(x)| = 0 \tag{9}$$
$$-> |f(x)| = 0 \forall f(x) \tag{10}$$
$$-> f(x) = 0 \forall x -> f = 0 \tag{11}$$

Too prove for homogeneous:

$$\alpha||f||_1 = ||\alpha f||_1 \tag{12}$$

Proof:

$$\alpha||f||_1 = ||\alpha f||_1 := \alpha \sum |f(x)| = \sum |\alpha f(x)| \tag{13}$$
$$-> \alpha \sum |f(x)| = \sum \alpha |f(x)| \tag{14}$$
$$-> \alpha \sum |f(x)| = \alpha \sum |f(x)| \tag{15}$$

,acc. to factoring rules Too prove for Triangle inequality:

$$||f + g||_1 \leq ||f||_1 + ||g||_1 \tag{16}$$

Proof:

$$||f + g||_1 \leq ||f||_1 + ||g||_1 := \sum |f(x) + g(x)| \leq \sum |f(x)| + \sum |g(x)| \tag{17}$$
$$-> if f(x), g(x) \geq 0 : \tag{18}$$
$$\sum f(x) + g(x) \leq \sum f(x) + \sum g(x) \tag{19}$$
$$-> \sum f(x) + g(x) \leq (f(x_0) + ... + f(x_n)) + (g(x_0) + ... + g(x_n)) \tag{20}$$
$$\sum f(x) + g(x) \leq (f(x_0) + g(x_0)) + ... + (f(x_n) + g(x_n)) -> \tag{21}$$
$$\sum f(x) + g(x) = \sum f(x) + g(x) \tag{22}$$
$$-> if f(x), g(x) \leq 0 : \tag{23}$$
$$\sum |f(x) + g(x)| = f(x) + g(x) \, and \sum |f(x)| + \sum |g(x)| = \sum f(x) + \sum g(x) -> \tag{24}$$

Proof acc. to if f(x),g(x) 0

$$-> if f(x) < 0, g(x) \geq 0:\tag{25}$$

$$\sum |f(x)| + \sum |g(x)| = \sum a|(x)| + \sum |g(x)|, if a(x) \geq 0 and f(x) = -a(x) -> \tag{26}$$

it's enough to proof

$$\sum |f(x) + g(x)| \leq \sum |a(x) + g(x)|, if a(x) \geq 0 and f(x) = -a(x)\tag{27}$$

which is given because f(x) < a(x)

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.

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.

<u>Answer</u>

1. In simple English wikipedia we observed that:

   - the numbers in each article(all articles) are not that much

   - the word summer is used more than other seasons

---

[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.

- the word female is used more than the word male the tables are considered words

- the words with 4 or more characters contain less vowels than consonants

2. The observation about frequency of numbers (sequence of digits) per article made us curios. This pattern may occur because many of the numbers in the articles are still wrote using letters. As a question we can ask ourselves if Simple English Wikipedia will be easier to read if more digit numbers would be used.

3. We have two hypotheses:

   - There is no article which has more than 30% of its words being numbers (we define numbers as being words formed using only digits, digits separated by one of the characters ”,” or ”.” and digits containing the character ”e” which stands for exponential) and (we define a word as being everything contained between two spaces)

   - 50% of the articles have no numbers (we define numbers as being words formed using only digits, digits separated by one of the characters ”,” or ”.” or containing the character ”e” which stands for exponential) and (we define a word as being everything contained between two spaces)

4. If our hypotheses are correct we should expect to have maximum percentage of ”number words” per article to be 30% and to have half of the articles with no number words at all.

5. On the data set provided we will use a regex to count how many of the words in an article are numbers. Using that we will calculate a % per article and we will be able to observe if our hypothesis is correct or not.

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.
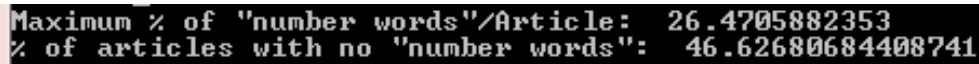
3.1 Hints:

- 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 So to answer this question we create a python script

```
 1: import pandas as pd
 2: import re
 3:
 4: articles = []
 5: numOfWordsPerArticle = []
 6: wordsThatAreNumbersPerArticle = []
 7: percentageOfNumberWordsInArticle = []
 8:
 9: #reading lines from simple article file
10: #and doing our opperations on each line(article)
11: with open('simple-20160801-1-article-per-line', 'rb') as fp:
12:     for line in fp:
13:         articles.append(line.decode("utf-8"))
14:         xWords = len(str(line).split(" "))
15:         numOfWordsPerArticle.append(xWords)
16:         xNumWords = len(re.findall(r"\s[+-]?\d+(?:\.\d+)?(?:[eE][+-]?\d+)?\s", str
17:         wordsThatAreNumbersPerArticle.append(xNumWords)
18:         percentage = ((xNumWords/xWords)*100) if xWords != 0 else 0
19:         percentageOfNumberWordsInArticle.append(percentage)
20:
21: #creating dictionary
22: data = {"Articles": articles,
23:         "Words/Article": numOfWordsPerArticle,
24:         "Words that are Numbers/Article": wordsThatAreNumbersPerArticle,
25:         "% of \"Number Words\"/Article": percentageOfNumberWordsInArticle}
26:
27: #creating dataframe
28: dataFrame = pd.DataFrame(data, columns=['Articles',
29:                                         'Words/Article',
30:                                         'Words that are Numbers/Article',
31:                                         '% of \"Number Words\"/Article'])
32:
33: #sorting data frame according to values in % column
34: dataFrame = dataFrame.sort_values('% of \"Number Words\"/Article')
```

```
35: #writing to xlsx file
36: writer = pd.ExcelWriter("testAgain.xlsx", engine='xlsxwriter')
37: dataFrame.to_excel(writer, sheet_name='Sheet1', index=False)
38: writer.save()
39:
40: #Showing the results of our calculations
41: print("Maximum % of \"number words\"/Article: ", max(dataFrame['% of \"Number Word
42: articlesWithNoNumWords = len(dataFrame[dataFrame['% of \"Number Words\"/Article']
43: allArticles = len(dataFrame['% of \"Number Words\"/Article'])
44: print("% of articles with no \"number words\": ", (articlesWithNoNumWords/allArti
```

And we got those results as shown in figure 1

Figure 1: Python Script Results

As the results show after implementing our test prediction, the first hypothesis was validated and the second one was falsified. Of course our hypotheses can be rephrased and extended.

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`.