

# Introduction to Web Science

## Assignment 6

Prof. Dr. Steffen Staab

[staab@uni-koblenz.de](mailto:staab@uni-koblenz.de)

René Pickhardt

[rpickhardt@uni-koblenz.de](mailto:rpickhardt@uni-koblenz.de)

Korok Sengupta

[koroksengupta@uni-koblenz.de](mailto: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: hotel

Andrea Mildes - [mildes@uni-koblenz.de](mailto:mildes@uni-koblenz.de)

Sebastian Blei - [sblei@uni-koblenz.de](mailto:sblei@uni-koblenz.de)

Johannes Kirchner - [jkirchner@uni-koblenz.de](mailto:jkirchner@uni-koblenz.de)

Abdul Afghan - [abdul.afghan@outlook.de](mailto:abdul.afghan@outlook.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  $\|\cdot\|_\infty$  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 set<sup>1</sup> 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.

---

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

**Answer:**

$$f : M \rightarrow \mathbb{R}$$

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

$$\|f\|_\infty := \sup_{x \in M} \|f(x)\|_{\mathbb{R}} = \sup\{|f(x)| : x \in M\}$$

$$\begin{aligned} f(0) &= 2, f(1) = -4, f(2) = 8, f(3) = -4 \\ g(0) &= 5, g(1) = 1, g(2) = 7, g(3) = -3 \end{aligned}$$

$$\|f - g\|_1 = \sum_{x \in M} |f(x) - g(x)|$$

$$\begin{aligned} &= |2 - 5| + |-4 - 1| + |8 - 7| + |-4 - (-3)| \\ &= 3 + 5 + 1 + 1 \\ &= 10 \end{aligned}$$

$$\|f - g\|_\infty = \sup_{x \in M} \|f(x) - g(x)\|_{\mathbb{R}} = \sup\{|f(x) - g(x)| : x \in M\}$$

$$\begin{aligned} |f(0) - g(0)| &= |2 - 5| = 3 \\ |f(1) - g(1)| &= |-4 - 1| = 5 \\ |f(2) - g(2)| &= |8 - 7| = 1 \\ |f(3) - g(3)| &= |-4 - (-3)| = 1 \end{aligned}$$

$$\rightarrow \sup_{x \in M} \|f(x) - g(x)\|_{\mathbb{R}} = 5$$

### Positive definite

Assumption:  $\|f\|_1 = 0 \Rightarrow f = 0$

Proof:

$$\|f\|_1 = 0 \Leftrightarrow \sum_{x \in M} |f(x)| = 0$$

$$\Rightarrow \sum_{x \in M} |f(x)| = 0$$

$$\Rightarrow |f(0)| + |f(1)| + \dots + |f(x)| = 0$$

$$\Rightarrow f(x) = 0 \quad \forall x$$

$$\Rightarrow f = 0$$

### Homogeneous

Assumption:  $\|\alpha f\|_1 = |\alpha| \|f\|_1, \alpha \in \mathbb{R}$

Proof:

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

$$\Rightarrow \sum_{x \in M} |\alpha| |f(x)| = |\alpha| \cdot |f(0)| + |\alpha| \cdot |f(1)| + \dots + |\alpha| \cdot |f(x)|$$

$$\Rightarrow |\alpha| \cdot (|f(0)| + |f(1)| + \dots + |f(x)|)$$

$$\Rightarrow |\alpha| \cdot \sum_{x \in M} |f(x)|$$

$$\Rightarrow |\alpha| \|f\|_1$$

Since  $\alpha$  has to be positive for our usage, we can write  $\alpha$  instead of  $|\alpha|$  and thus gain:

$$\alpha \|f\|_1$$

**Triangle inequality**

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

Proof:

$$\sum_{x \in M} |f(x) + g(x)| \leq \left( \sum_{x \in M} |f(x)| \right) + |g(x)| \leq \sum_{x \in M} |f(x)| + \sum_{x \in M} |g(x)| = \|f\|_1 + \|g\|_1$$

## 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**<sup>2</sup> 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.

- Not every article contains quotation marks
- Not every article contains parenthesis

---

<sup>2</sup>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.

- More than one article contains numbers
  - Articles contain vowels and consonants
  - The length of the articles differ
  - Many articles start with an article
2. The most interesting would be the start of an article, as there the user decides whether to continue reading or not.
- 3.
- In the whole article text of the Simple English Wikipedia are more occurrences of vowels than consonants.  
This hypothesis is false if there are more consonants than vowels in the article texts of the Simple English Wikipedia.
  - In the whole article text of the Simple English Wikipedia are more occurrences of letters than numbers.  
This hypothesis is false if there are more occurrences of numbers than letters in the article texts of the Simple English Wikipedia.
  - In the Simple English Wikipedia more than 30% of article texts start with a definite or indefinite article.  
This hypothesis is false if 30% or less of the opening words in the article texts of the Simple English Wikipedia are articles.
4. We choose hypothesis three: "In the Simple English Wikipedia more than 30% of article texts start with a definite or indefinite article.". The hypothesis is already testable.
5. We will extract every opening word of each article. Then we will compare it to a list of articles<sup>3</sup> not regarding upper and lower case. If the opening word matches, we will increase a counter. Finally, we will divide the counter by the total number of articles. Therefore we have the percentage of articles starting with an article. We expect our outcome to have a deviation of +/- 5%.

---

<sup>3</sup>a, an, the

### 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:

Out of 119,753 article texts 24,304 started with a definite or indefinite article. This corresponds to a percentage of 20.3%. Therefore, the hypothesis "In the Simple English Wikipedia more than 30% of article texts start with a definite or indefinite article." must be rejected.

#### Code:

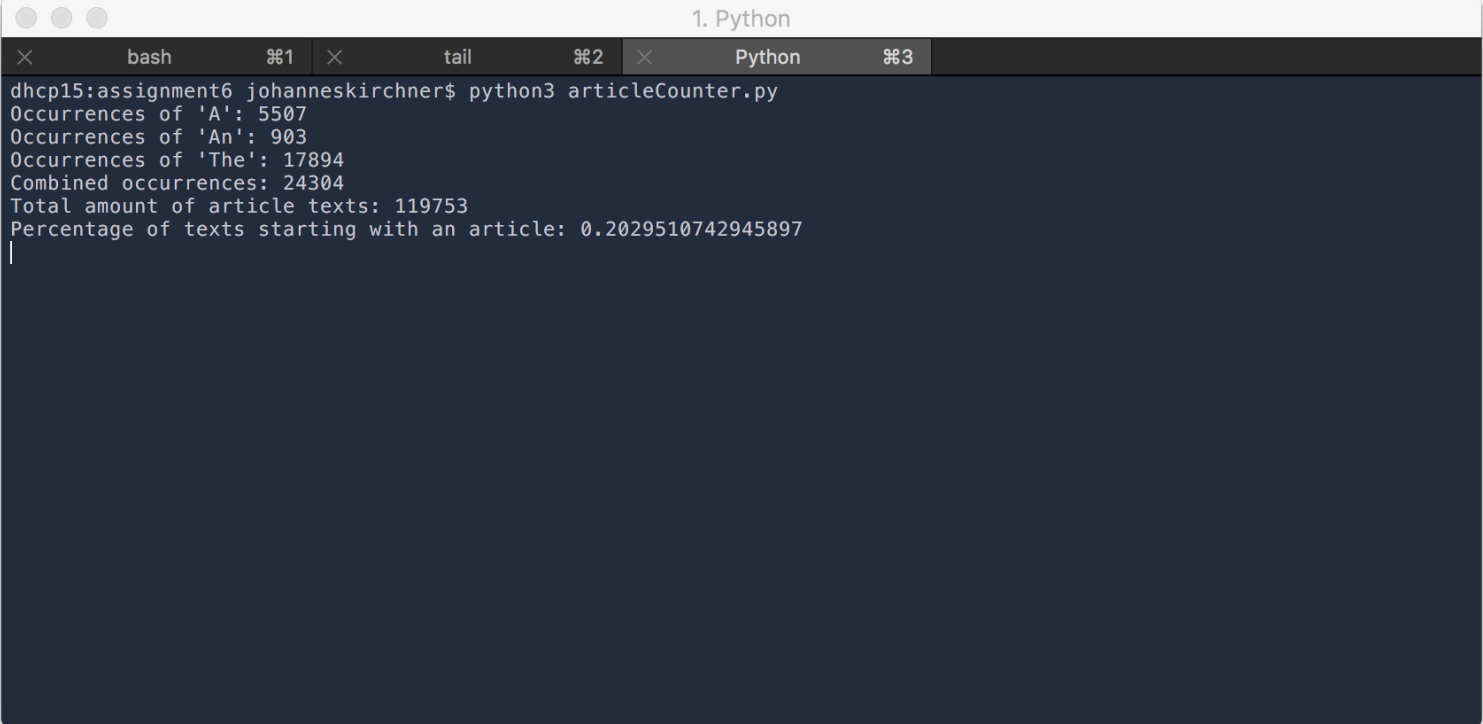
```
1: # assignment 6 task 3
2: # Andrea Mildes - mildes@uni-koblenz.de
3: # Sebastian Blei - sblei@uni-koblenz.de
4: # Johannes Kirchner - jkirchner@uni-koblenz.de
5: # Abdul Afghan - abdul.afghan@outlook.de
6:
7: import numpy as np
8: import matplotlib.pyplot as plt
9:
10:
11: # Draw a graph showing the amount of articles starting with a definite
12: # or indefinite article.
13: def draw_graph(a, an, the, count, total):
14:     objects = ('An', 'A', 'The', 'Combined')
15:     y_pos = np.arange(len(objects))
16:     occurrences = (an, a, the, total)
17:
18:     plt.bar(y_pos, occurrences, align='center', alpha=0.5)
19:     plt.xticks(y_pos, objects)
20:     plt.ylabel('Occurrences')
21:     plt.title('Article texts of the Simple English Wikipedia starting '
22:             'with articles')
23:     textstr = 'Occurrences of \'an\' = %.0f\n' \
24:             'Occurrences of \'a\' = %.0f\n' \
25:             'Occurrences of \'the\' = %.0f\n' \
26:             'Combined occurrences = %.0f\n' \
27:             'scanned article texts = %.0f\n' \
```



```
28:         'Percentage of texts starting \n' \
29:         'with an acrticle = %.4f' % \
30:         (an, a, the, total, count, ((total/count)*100))
31:     props = dict(facecolor='white', alpha=0.5)
32:     plt.text(-0.4, 24000, textstr, fontsize=14, verticalalignment='top', bbox=prop)
33:
34:     plt.show()
35:
36:
37: # Read the given file and store each line (article text) in a list
38: def readLines(file):
39:     with open(file) as f:
40:         wiki_articles = f.readlines()
41:     return wiki_articles
42:
43:
44: # Iterate through each article text and check if the first word is an article
45: def countArticles(wiki_articles):
46:     a = 0
47:     an = 0
48:     the = 0
49:     count = 0
50:
51:     for article in wiki_articles:
52:         first_word = article.split(' ', 1)[0]
53:         first_word = first_word.strip()
54:         first_word = first_word.lower()
55:
56:         count += 1
57:         if first_word == "a":
58:             a += 1
59:         elif first_word == "an":
60:             an += 1
61:         elif first_word == "the":
62:             the += 1
63:
64:     total = a + an + the
65:     return a, an, the, count, total
66:
67:
68: def main():
69:     file = "simple-20160801-1-article-per-line"
70:     wiki_articles = readLines(file)
71:     a, an, the, count, total = countArticles(wiki_articles)
72:     print("Occurrences of \'A\': " + str(a))
73:     print("Occurrences of \'An\': " + str(an))
74:     print("Occurrences of \'The\': " + str(the))
75:     print("Combined occurrences: " + str(total))
76:     print("Total amount of article texts: " + str(count))
```

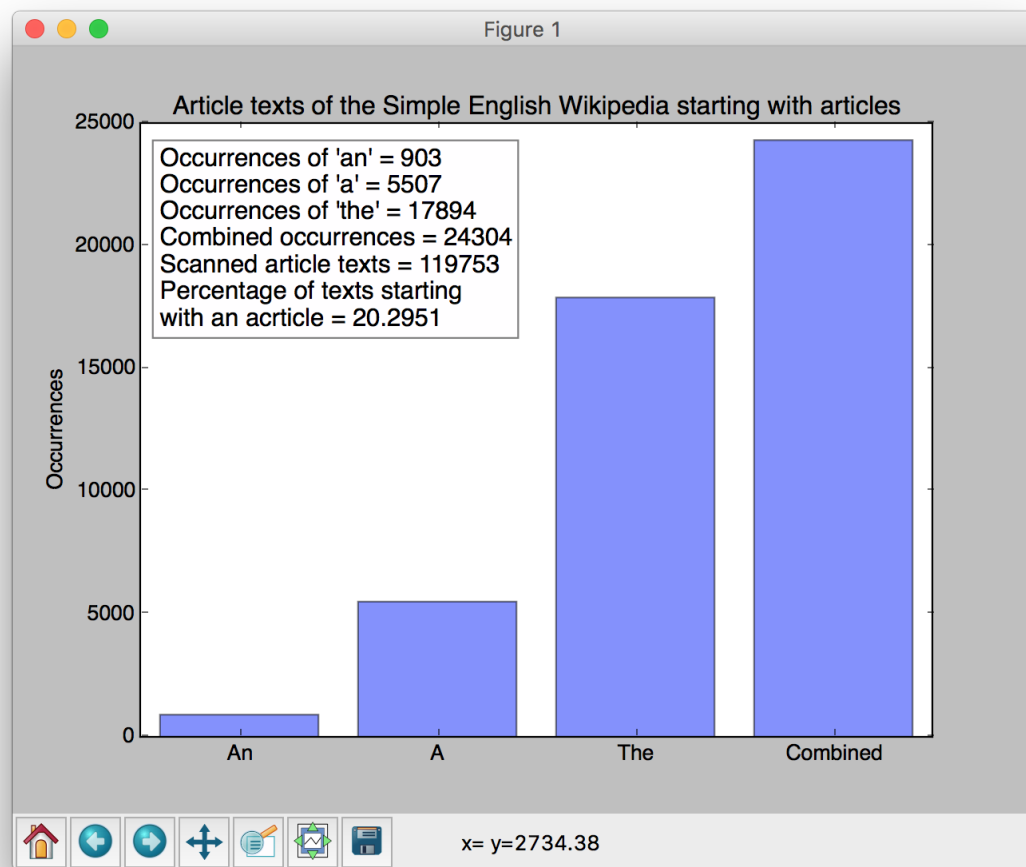
```
77:     print("Percentage of texts starting with an article: " + str(total / count))
78:     draw_graph(a, an, the, count, total)
79:
80:
81: if __name__ == '__main__':
82:     main()
```

---



The screenshot shows a terminal window with a dark blue background. The title bar at the top reads "1. Python". Below the title bar, there are three tabs: "bash" (active), "tail", and "Python". The terminal content shows the execution of a Python script named "articleCounter.py". The output of the script is as follows:

```
dhcp15:assignment6 johanneskirchner$ python3 articleCounter.py
Occurrences of 'A': 5507
Occurrences of 'An': 903
Occurrences of 'The': 17894
Combined occurrences: 24304
Total amount of article texts: 119753
Percentage of texts starting with an article: 0.2029510742945897
|
```



## 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  $\LaTeX$ engine to LuaLaTeX.