

# Introduction to Web Science

## Assignment 10

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)

Olga Zagovora

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

Institute of Web Science and Technologies  
Department of Computer Science  
University of Koblenz-Landau

Submission until: January 25, 2016, 10:00 a.m.

Tutorial on: January 27, 2016, 12:00 p.m.

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: XXXX

## 1 Modeling Twitter data (10 points)

In the meme paper<sup>1</sup> by Weng et al., in Figure 2<sup>2</sup> you find a plot, comparing the system entropy with the average user entropy. Your task is to reproduce the plot and corresponding calculations.

1. We provide you with the file 'onlyhashtag.data', containing a collection of hashtags from tweets. Use this data to reproduce the plot from the paper. Once you have the values for average user entropy and system entropy calculated per day create a scatter plot to display the values.
2. Interpret the scatter plot and compare it with the authors interpretation from the graph showed in the paper. Will the interpretations be compatible to each other or will they contradict each other? Do not write more than 5 sentences.

### 1.1 Hints

1. Use formulas from the lecture to calculate the entropy for one user and the system entropy.
2. Do not forget to give proper names of plot axes.

---

<sup>1</sup><http://www.nature.com/articles/srep00335>

<sup>2</sup>Slide 27, Lecture Meme spreading on the Web

## 2 Measuring inequality (10 points)

We provide you with a sample implementation of the Chinese Restaurant Process<sup>3</sup>.

Assume there is a restaurant with an infinite number of tables. When a new customer enters a restaurant he chooses an occupied table or the next empty table with some probabilities.

According to the process first customer always sits at the first table. Probability of the next customer to sit down at an occupied table  $i$  equals ratio of guests sitting at the table  $(c_i/n)$ , where  $n$  is the number of guests in the restaurant and  $c_i$  is the number of guests sitting at table  $i$ .

Probability of customer to choose an empty table equals :  $1 - \sum_{i=1}^S p_i$ , where  $S$  is the number of occupied tables and  $p_i = c_i/n$ .

Provided script simulates the process and returns number of people sitting at each table. We will study restaurants for 1000 customers. Now you should modify the code and evaluate how unequal were the customers' choices of tables.

Calculate the Gini- coefficient measuring the inequality between the tables, until the coefficient stabilizes. Do five different runs and plot your results in a similar way that plots in the lecture slides are done, cf. Slide 32 and Slide 33.

---

<sup>3</sup>File "chinese\_restaurant.py"; Additional information can be found here: [https://en.wikipedia.org/wiki/Chinese\\_restaurant\\_process](https://en.wikipedia.org/wiki/Chinese_restaurant_process)

### 3 Herding (10 points)

Let us consider the altitude of Koblenz to be 74 m above sea level. You are asked to figure out the height of the Ehrenbreitstein Fortress and the Fernmeldeturm Koblenz without googling.

The exercise is split in two parts:

#### Part 1 : The Secret

In *complete secrecy*, each member of the team will write down their estimated height of the Ehrenbreitstein Fortress without any form of discussion. Please keep in mind that you need to have reasons for your assumption. Once you are done, then openly discuss in the group and present you values in a tabulated format with the reasons each one assumed to arrive at that value.

#### Part II : The Discussion

Discuss amongst yourself with valid reasoning what could be the height of the Fernmeldeturm Koblenz. Only after discussing, each member of the group is asked to arrive at a value and present this value in a tabulated format as was done in Part I.

Calculate the Mean, Standard Deviation and Variance of your noted results for both the cases and explain briefly what you infer from it.

**Note:** This exercise is for you to understand the concepts of herding and not to get the perfect height by googling information. There is in fact no point associated with the height but with the complete reasoning that you provide for your answers.

## Important Notes

### Submission

- Solutions have to be checked into the github repository. Use the directory name `groupname/assignment10/` 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".

### **L**A<sub>T</sub>E<sub>X</sub>

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<sub>T</sub>E<sub>X</sub>engine to **LuaLaTeX**.