Exercises 4
17.03.2014

Rules: The document contains a set of 5 exercises/subpoints, each of them worth 2 points. You need to provide for each of the first two exercises one pdf file per subpoint with the solution to the problem, and for the third one a Python script and a chart. All files must be put in a ZIP archive named FirstName_LastName_ex*n*.zip, where n is the number of the exercise session (see ex_set_1.pdf). The files must be named exercise_m(_a/b).py/pdf, where m is the number of the exercise and a/b the subpoint, if it's the case. The ZIP archive must be uploaded on ILIAS until the specified deadline.
For each exercise session you will get either one point or zero. One point is given if 4 out of 5 exercises/subpoints in this session are correctly solved.
Note: You can also use the tables at this link to find the limit values for the T-Test (consider the significance level of the first row, for two-tailed tests) and Chi-square test.

Exercise 1. You have a text of 1 million words (n = 1000000). The word *Agatha* is found 30 times in the text, the word *Christie* is found 117 times in the text, while the bigram *Agatha Christie* has 20 occurrences. Using a **T-Test**, and assuming a significance level of 1% and infinite degrees of freedom (dof = inf):

a) decide if the words *Agatha* and *Christie* appear independently in the text

b) what if the number of occurrences for *Agatha* is 200, for *Christie* is 300, and for the entire bigram is 6?

Exercise 2. Association between terms *I* and *you* and texts written by women. Compute the **Chi-square** values for the two cases below and decide if there is a correlation between the term and woman authors. Assume the degrees of freedom equal to 1.

a)

Observed	*Woman*	not *Woman*	
you	350	650	1000
not *you*	19650	39350	59000
	20000	40000	60000

b)

Observed	*Woman*	not *Woman*	
I	450	550	1000
not *I*	19550	39450	59000
	20000	40000	60000

Exercise 3. Use the input file input_ex3.txt. Write a Python script that parses the file and counts the number of word tokens, word types and the frequency for the first 100 most frequent words. Demonstrate the **Zipf's Law** on these 100 most frequent terms. Plot a log-log chart that proves this.