Exercises 10
12.05.2014

Rules: The document contains a set of 3 exercises: exercise 1 and 2 worth 4 points each, exercise 3 worth 2 points. You need to provide a *separate* PDF file for each of the three exercises. All files must be compressed 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 ZIP file must be uploaded on ILIAS until the specified deadline.
Attention! You also have a bonus exercise worth 5 points. This is not mandatory, but the points can be added to a former set, where you lost points and you want to raise your grade.
Good luck!

Exercise 1. You have the profile of 3 authors and test text Q below. Decide if Q is likely to be written by any of the authors, using the Z-score and Burrows' Delta.

a)

Profile	**A1**	**A2**	**A3**	**Q**
tf	??	??	??	??
term1	20	21	3	25
term2	36	100	23	40
term3	90	100	12	45
term4	75	3	67	10
term5	12	57	34	20

b)

Profile	**A1**	**A2**	**A3**	**Q**
tf	??	??	??	??
term1	2	36	100	3
term2	78	45	5	2
term3	10	30	6	5
term4	11	32	23	4
term5	5	29	0	2

Exercise 2. Using the same input tables as in Exercise 1, decide if Q is likely to be written by any of the authors based on the Kullback-Leibler Divergence (use Laplace smoothing with lambda=1).

Exercise 3. The following text is encrypted using a simple substitution method. The plaintext is part of an English text encoded in upper case characters without punctuation marks. Using the distribution of the characters in English texts below, recover the plaintext and explain in detail the steps.

ODQSOCL OW GIU BOEE QRROHOCS QV GIUR KIA QF Q DQCQSLR WIR
ICL IW CQFQF EIYQE YIDJUVLR FGFVLDF GIU SLV OCVI GIUR
IWWOYL IC VXQV DICPQG DIRCOCS VI WOCP VXL JXICLF ROCSOCS
LHLRG YQEELR OF Q POFVRQUSXV YICWUFLP CQFQ BIRMLR QCP
LHLRG YQEELR QFFURLF GIU VXQV XOF IR XLR WOEL IR
QYYIUCVOCS RLYIRP IR RLFLQRYX JRIKLYV LHLRG ICL IW BXOYX
OF DOFFOCS WRID VXL YIDJUVLR FGFVLD OF QAFIEUVLEG HOVQE

Distribution of characters in English texts:

A 8.167	H 6.094	O 7.507	V 0.978
B 1.492	I 6.996	P 1.929	W 2.360
C 2.782	J 0.153	Q 0.095	X 0.150
D 4.253	K 0.772	R 5.987	Y 1.974
E 12.702	L 4.025	S 6.327	Z 0.074
F 2.228	M 2.406	T 9.056	
G 2.015	N 6.749	U 2.758	

Hint: Start by computing the distribution of characters in your given cipher text. Then, think about bigrams, trigrams and their frequency in English.

Bonus. Write a Python script that parses the following corpus in input_2.txt file:

- Corpus 1 written by Author 1: texts 2, 3 and 4
- Corpus 2 written by Author 2: texts 15, 16 and 17

The script should decide based on the Kullback-Leibler Divergence if a given text is written by Author 1 or Author 2, by following the steps:

a) compute the frequency of the most frequent 100 english terms (you can find these terms in the frequent100.txt file) in the given corpus, for each term and each sub-corpus

b) compute the size in each sub-corpus

c) divide by the size

d) use Laplace with lambda=1 for smoothing (add-one smoothing technique)

e) compute KLD divergence between each author profile and the given test text.

Based on the script's results, is text 37 written by any of the 2 authors? What about text 8?