Assignment 2

Remarks:

1. Please note:
   - You can work on the exercises in groups with up to three students.
   - Groups can be changed for each assignment sheet.
   - Please comment which students have worked on the current assignment sheet.
   - Please **submit only once** per group.
   - Submit your exercises on the course homepage on Ilias. You find it at: `https://ilias3.uni-stuttgart.de/` → Magazin → Ingenieurwissenschaften → Maschinelle Sprachverarbeitung / Computational Linguistics → Lehrveranstaltungen WS 12/13 → Information Retrieval und Text Mining (WS 12/13)
   - Please submit your work until **November, 07 23:59** on Ilias.
2. If you need any **help**, send an email to `max.kisselew@ims.uni-stuttgart.de` or drop in at my office: 0.012 (Pfaffenwaldring 5b).

Exercise 1 (IIR 13) [2 P.]

Based on the data below, estimate a multinomial Naive Bayes classifier (the type of NB classifier we introduced in class) and apply the classifier to the test document. Calculate the probability that the classifier assigns the test document to $c = \textit{Japan}$ or $\overline{c}$.

	docID	words in document	in c = Japan?
training set	1	Kyoto Tokyo Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Beijing	no
test set	5	Taiwan Taiwan Kyoto	?

Exercise 2 (IIR 3) [1 P.]

If you wanted to search for `s*ng` in a permuterm wildcard index, what key(s) would one do the lookup on?

Exercise 3 (IIR 3) [3 P.]

Compute the Levenshtein matrix for the distance between the strings “obama” (input) and “romney” (output). Use this format (as introduced in class):

		c		a		t		c		a		t	
	0	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**	**6**	**6**
c	**1**	**0**	2	2	3	3	4	**3**	5	5	6	6	7
	1	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**
a	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5	5	6
	2	3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**
t	**3**	3	**2**	2	**1**	**0**	2	2	3	3	4	**3**	5
	3	4	**2**	3	**1**	2	**0**	**1**	**1**	**2**	**2**	**3**	**3**

After you have calculated the distance between the two strings: Trace the editing operations for one possible editing path as demonstrated in class:

cost	operation	input	output
1	insert	*	c
1	insert	*	a
1	insert	*	t
0	(copy)	c	c
0	(copy)	a	a
0	(copy)	t	t

Exercise 4 (IIR 13) [2 P.]

Rank the documents in collection $\{d_1, d_2\}$ for query q using the language model approach to IR introduced in class with Jelinek-Mercer smoothing. Use the mixture coefficient $\lambda = 0.4$.

- d_1: The European Union Act 2011 prevents additional powers being passed to Brussels without a referendum
- d_2: EU will not ban Chanel 5 perfumes over allergy findings
- Query q: European Union

Exercise 5 (IIR 12/13) [3 P.]

Consider the following frequencies for the class *coffee* for four terms in the first 100,000 documents of Reuters-RCV1:

term	N_{00}	N_{01}	N_{10}	N_{11}
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,524	119	1118	34
roasted	99,824	143	23	10

a) Which two terms will be selected in frequency-based feature selection and why?

b) Compute the MI values and order the terms according to MI. Which two terms will be selected in MI-based feature selection?

Due date: Tuesday, November 07, 2012, 23:59