Assignment 1

Remarks:

1. To pass the course you will need to ...
   - achieve at least 50 % of the points you can get over all assignment sheets
   - vote 50 % of all exercises in the course
   - pass the **final exam** at the end of the course (your final exam grade will be your course grade)
2. 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 **October, 23 23:59** on Ilias.
3. 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 1) [1 P.]

Shown below is a portion of a positional index in the format:
term: doc1: ⟨position1, position2, . . . ⟩; doc2: ⟨position1, position2, . . . ⟩; etc.

ANGELS: 2: ⟨36,174,252,651⟩; 4: ⟨12,22,102,432⟩; 7: ⟨17⟩;
FOOLS: 2: ⟨1,17,74,222⟩; 4: ⟨8,78,108,458⟩; 7: ⟨3,13,23,193⟩;
FEAR: 2: ⟨87,704,722,901⟩; 4: ⟨13,43,113,433⟩; 7: ⟨18,328,528⟩;
IN: 2: ⟨3,37,76,444,851⟩; 4: ⟨10,20,110,470,500⟩; 7: ⟨5,15,25,195⟩;
RUSH: 2: ⟨2,66,194,321,702⟩; 4: ⟨9,69,149,429,569⟩; 7: ⟨4,14,404⟩;
TO: 2: ⟨47,86,234,999⟩; 4: ⟨14,24,774,944⟩; 7: ⟨199,319,599,709⟩;
TREAD: 2: ⟨57,94,333⟩; 4: ⟨15,35,155⟩; 7: ⟨20,320⟩;
WHERE: 2: ⟨67,124,393,1001⟩; 4: ⟨11,41,101,421,431⟩; 7: ⟨16,36,736⟩;

Which document(s) (if any) match each of the following queries at which positions, where each expression within quotes is a phrase query? (i) “fools rush in” (ii) “fools rush in” AND “angels fear to tread”.

Exercise 2 (IIR 1) [3 P.]

Write a (Python) program that implements a simple version of Boolean retrieval. The program should be able to interpret simple queries consisting of one term like ’stuttgart’ or conjunctive queries with two terms connected by ’AND’ like ’stuttgart AND university’. You do not have to treat OR or NOT in this exercise.

Supporting files for this assignment can be found in the zip archive 'assignment_1_supp.zip' on the course homepage. The archive contains text files which you can test your program on. In addition, a simple tokenizer is included which you may import into your code to tokenize the text files. Use the function `'tokenize(...)'` for this purpose.

Implement the following functions/methods:

(a) `make_inverted_index(folder_with_text_files)`: opens text files in the specified folder, normalizes them using the tokenizer and then adds all the tokens found in the documents to the inverted index which is returned. A good way to store the index could be a dictionary.

(b) `process_simple_query(query)`: returns and displays the postings list (and the corresponding document names) for a one-word query.

(c) `intersect(p1, p2)`: Implements the algorithm presented in class that intersects two postings lists and returns the intersection.

(d) `process_conjunctive_query(query_tokens)`: Calls the intersection algorithm and displays the intersection (with the corresponding document names) of two posting lists for a query connected by an 'AND'.

Notes:

- To get a user input in the console you can use the `raw_input(...)` function:

```
>>> query = raw_input("Please enter your query: ")
Please enter your query: blahblah
>>> query
'blahblah'
```

- You can access all file names in a particular folder by means of the `listdir(...)` command from the `os` module. E. g.:

```
import os

for f in os.listdir("files"):
    ...
```

- If you are not familiar with Python you may also use other programming languages like Java or C/C++.

Exercise 3 (IIR 2) [1 P.]

The following pairs of words are stemmed to the same form by the German version of the Porter stemmer[1]. Which pairs, would you argue, should not be conflated? Give your reasoning.

(a) geräumige/geräumigen (→ geraum)

(b) musik/musiker (→ musik)

(c) neuer/neues (→neu)

(d) persönlich/persönlichkeit (→ person)

[1] http://snowball.tartarus.org/algorithms/german/stemmer.html

(e) schlaf/schlafen (→ schlaf)

(f) unternehmer/unternehmung (→ unternehm)

(g) wetten/wetter (→ wett)

Exercise 4 (IIR 1) [1 P.]

For a conjunctive query, is processing postings list in order of size guaranteed to be optimal? Explain why it is, or give an example where it is not.

<u>Due date:</u> Tuesday, October 23, 2012, 23:59