The Term Vocabulary (Chapter 2)

Definition 1 (Recall)
Recall describes how many of the relevant documents are retrieved.

$$recall = R = \frac{\#relevant\ retrieved}{\#relevant}$$

Definition 2 (Precision)
Precision describes how many of the retrieved documents are relevant.

$$precision = P = \frac{\#relevant\ retrieved}{\#retrieved}$$

Definition 3 (Porter stemmer)
The entire Porter's algorithm is too complex to present here. It consists of 5 phases of word reductions, applied sequentially. The first phase uses the following rule group. Importantly, only the rule that applies to the longest suffix is used.

Rule			***Example***		
SSES	$\rightarrow$	*SS*	*caresses*	$\rightarrow$	*caress*
IES	$\rightarrow$	*I*	*ponies*	$\rightarrow$	*poni*
SS	$\rightarrow$	*SS*	*caress*	$\rightarrow$	*caress*
S	$\rightarrow$		*cats*	$\rightarrow$	*cat*

Exercise 2/1

Are the following statements true or false?

1. In a Boolean retrieval system, stemming never lowers precision.
2. In a Boolean retrieval system, stemming never lowers recall.
3. Stemming increases the size of the vocabulary.
4. Stemming should be invoked at indexing time but not while processing a query.

Exercise 2/2

Suggest what normalized form should be used for these words (including the word itself as a possibility):

1. 'Cos
2. Shi'ite
3. cont'd
4. Hawai'i
5. O'Rourke

Exercise 2/3

The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldn't be conflated. Give your reasoning.

1. abandon/abandonment
2. absorbency/absorbent
3. marketing/markets
4. university/universe
5. volume/volumes

Exercise 2/4

For the Porter stemmer group shown in Definition 3:

1. What is the purpose of including an identity rule such as SS $\rightarrow$ SS?
2. Applying just this rule group, what will the following words be stemmed to?

   circus, canaries, boss

3. What rule should be added to correctly stem *pony*?
4. The stemming for *pony* and *ponies* might seem strange. Does it have a deleterious effect on retrieval? Why or why not?

Posting Lists (Chapter 2)

Exercise 2/5

Below is a part of index with positions in the form
doc1: $\langle pos1, pos2, pos3, \ldots\rangle$; doc2: $\langle pos1, pos2, \ldots\rangle$; ...

- angels: $2 : \langle 36, 174, 252, 651\rangle$; $4 : \langle 12, 22, 102, 432\rangle$; $7 : \langle 17\rangle$;
- fools: $2 : \langle 1, 17, 74, 222\rangle$; $4 : \langle 8, 78, 108, 458\rangle$; $7 : \langle 3, 13, 23, 193\rangle$;
- fear: $2 : \langle 87, 704, 722, 901\rangle$; $4 : \langle 13, 43, 113, 433\rangle$; $7 : \langle 18, 328, 528\rangle$;
- in: $2 : \langle 3, 37, 76, 444, 851\rangle$; $4 : \langle 10, 20, 110, 470, 500\rangle$; $7 : \langle 5, 15, 25, 195\rangle$;
- rush: $2 : \langle 2, 66, 194, 321, 702\rangle$; $4 : \langle 9, 69, 149, 429, 569\rangle$; $7 : \langle 4, 14, 404\rangle$;
- to: $2 : \langle 47, 86, 234, 999\rangle$; $4 : \langle 14, 24, 774, 944\rangle$; $7 : \langle 19, 319, 599, 709\rangle$;
- tread: $2 : \langle 57, 94, 333\rangle$; $4 : \langle 15, 35, 155\rangle$; $7 : \langle 20, 320\rangle$;

- where: $2 : \langle 67, 124, 393, 1001 \rangle$; $4 : \langle 11, 41, 101, 421, 431 \rangle$; $7 : \langle 15, 35, 735 \rangle$;

The following terms are phrase queries. Which documents correspond to the following queries and on which positions?

a) *fools rush in*

b) *fools rush in* AND *angels fear to tread.*

The index is incorrect. How?

Exercise 2/6

Below is a part of index with positions in the form
doc1: $\langle pos1, pos2, pos3, \ldots \rangle$; doc2: $\langle pos1, pos2, \ldots \rangle$; ...

- ostrich: 1 : <1,7>; 2 : <4,5>;
- hippo: 1 : <5,8,9>; 2 : <6,9>;
- lion: 1 : <3,6>; 2 : <3,7>;
- giraffe: 1 : <2,4>; 2 : <1,2,8>;

Which documents correspond to the phrase query *lion giraffe hippo* and on which positions? Include intermediate results.

Exercise 2/7

Consider a query composed of two terms. Non-positional postings list of one term is composed of 16 items $P = [4, 6, 10, 12, 14, 16, 18, 20, 22, 32, 47, 81, 120, 215, 300, 500]$ and the second term has the postings list of only a single element $R = [47]$. Find out how many comparisons (and why) are necessary to find out the intersection of the lists that are organized as follows:

a) standard postings lists

b) postings lists with skip pointers of skip frequency $\sqrt{|P|}$

Exercise 2/8

Consider a query composed of two terms. Non-positional postings list with skip pointers of one term is composed of 16 items $P_1 = [4, 6, 10, 12, 14, 16, 18, 20, 22, 32, 47, 81, 120, 215, 300, 500]$ with skip frequency of square root of its length and the second term has the standard postings list $P_2 = [18, 32, 60]$. How many comparisons are necessary to find out the intersection of the lists?

Exercise 2/9

List the comparisons performed to intersect the following sorted non-positional postings lists with skip pointers of frequency 5.

$$P_1 = [2, 10, 12, 16] \quad \text{and} \quad P_2 = [1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]$$

Exercise 2/10

List the comparisons performed to intersect the following sorted non-positional postings lists with skip pointers of frequency 5.

$$P_1 = [4, 5, 6, 7, 8, 9, 10, 13, 14, 15] \quad \text{and} \quad P_2 = [1, 2, 3, 4, 5, 10, 11, 15, 16]$$
