Information Retrieval 2004 – Paper 7 Question 12 (SHT)

a) What role does stemming play in automatic indexing? [4]

[**Stemming is discussed in lecture 3; bookwork**]

Answer: Stemming maps words which are morphologically related onto one string, the stem. This is important in IR, because words that are morphologically related typically mean something similar (eg. ’skied’, ’Ski’ and ’Skiing’). As stemming increases the number of documents that match (by mapping query and document terms to the stem), it can increase the recall, while possibly lowering precision (because some of the now-related terms might not in fact have the same semantics). [5]

b) Briefly describe the principles behind the Porter Stemmer. [5]

[**The Porter Stemmer is discussed in examples and particular rules in supervisions. This is an easy question, bookwork.**]

Answer: The Porter stemmer describes English words with the use of a suffix lexicon, but without a stem lexicon. Rules are expressed as conditions of which letters are removed under which conditions, eg. -ed is only removed if the preceding part of the word contains at least one Vowel-Consonant sequence. Conditions can be about the structure of the syllable, or about certain letters contained in parts of the word, or letters in certain positions in the word. Most conditions are about the word’s measure (number of Vovel-Consonant sequences). Conditions can be combined with Boolean operators. Rules are grouped, and in each group only one rule must match. [3]

[**not all of these conditions must be described; this is rather a list of things they can say about the conditions. It is most important that the realize that a suffix not stem lexicon is used!**]

c) One extreme form of ”stemming” is the mapping of all words with a certain prefix onto one term. What effect will this prefix mapping have if the prefix length chosen is extremlely short, eg. only two letters long? Compare to a situation with more linguistically motivated stemming. [5]

[**This is a more difficult question – prefix mapping with length 6 was used in early IR research, and I mention this briefly in lecture 3. This question checks whether they do understand what stemming does and asks them to think creatively. Two points for the more difficult consequences; one for obvious ones. Below, a few are listed. Also, one to two points for convincing examples of what would go wrong**]

Answer:

- Many words will be mapped onto each other that don’t really mean the same thing
- I expect examples here, eg: ’dodo’ and ’domestic’ or ’it(alian)’ and ’it’. – one point for (a) convincing example(s)
- the Zipfian hypothesis in IR (medium-frequency words hold most meaning) is undermined by this type of ’stemming’.
- TF/IDF weighting will no longer work as a consequence.
- the overall number of terms in a system will be drastically reduced (to 26X26 terms) – this will decrease indexing a lot

- precision will go down dramatically; recall might go up but only slightly as so many irrelevant documents will be picked up

- One needs to seriously apply stop lists to make this less catastropic, otherwise. all words starting with 'th' will be mapped on each 'the'.

d) Consider a query with 4 relevant documents, and a ranked IR system which returns them in the following order ('X' represents a relevant document, '–' represents an irrelevant document).

Rank	**Relevance**	**Prec.**	**Recall**
1	X	P=1.00	R=.25
2	X	P=1.00	R=.5
3	–		
4	–		
5	–		
6	–		
7	–		
8	–		
9	–		
10	X	P=.3	R=.75
11	X	P=.3636	R=1.0

Calculate this system's mean precision at seen documents and 11-point average precision.[6]
[**Standard IR evaluation measures, discussed in detail with examples in lecture 4.**
Answer:
Mean precision at seen documents: $\frac{1.0+1.0+.3+.3636}{4} = .665$
11pt precision:
(boldface means value is interpolated)

Recall	**Prec.**
0	**1.0**
.1	**1.0**
.2	**1.0**
.3	**1.0**
.4	**1.0**
.5	1.0
.6	**0.3**
.7	**0.3**
.8	**0.3636**
.9	**0.3636**
1.0	0.3636
AVG	.698181