Inf1B Data and Analysis
Tutorial 8 (week 10)

Hutchins-Korte and Simpson

6 March 2008

- Please answer all questions on this worksheet in advance of the tutorial, and bring with you all work. Tutorials cannot function properly unless you do the work in advance.
- Data & Analysis tutorial exercises are not assessed, but are a compulsory and important part of the course. If you do not do the exercises then you are unlikely to pass the exam.
- Attendance at tutorials is obligatory; please let your tutor know if you cannot attend.
- *Background Reading:* Lecture notes 10 & 11.

Question 1

You are trying to find information about cheap flights to China for the 2008 Olympics. Using CQP, you try executing the following query on a corpus of advertisements that has been annotated with POS tags:

```
CQP> [word="fly"] []* [word="China"];
```

(a) Describe what this query will return.

(b) After executing the query, you realize it is mostly returning ads for insect control products which were produced in China. Suggest an improvement to the query above which would filter out the insect control ads.

(c) Describe what the query you provided in (b) will return.

Question 2

Next, you search on the terms: **cheap**, **flights**, **olympic**, **games**, **China** using an *information retrieval system*. You find three possible documents. You are given the frequency of each of the terms in each document, as shown below:

Terms	cheap	flights	olympic	games	China
Document 1	10	8	0	2	1
Document 2	0	0	9	9	8
Document 3	2	2	4	4	6
Query	1	1	1	1	1

(a) Compute the *cosine similarity measure*, relative to the query, for all three documents. (N.B. For this question, do this using the 5-value vectors in the table. In reality an IR system would use a higher-dimensional vector space indexed by all words in the document collection.)

(b) Given your answers to (a), rank the documents in order of preference.

(c) Discuss the appropriateness of using the given query and of ranking it using the *cosine* similarity measure for the task of obtaining the desired information.

Question 3

Two different systems for information retrieval, System A and System B, are evaluated in terms of their ability to retrieve relevant documents from a collection. A document collection of 1000 documents is queried. 17 of the documents are relevant to the Olympics query of Question 2. System A retrieves 50 documents, 16 of which are relevant. System B retrieves 10 documents, 8 of which are relevant.

(a) Calculate the precision and recall for the two systems. Which has the higher precision? Which has the higher recall?

(b) In order to choose between the two systems, you decide to evaluate them using an F-score to combine the precision and recall values. Because you dislike looking through non-relevant documents, you decide that precision is three times as important as recall, so you use the F-score F_α with $\alpha = 0.75$. Compute this measure for both systems. Which system is better?