

# Informatics 1 Data & Analysis

## Tutorial 7

Week 10, Semester 2, 2012

- Please attempt all questions on this worksheet in advance of the tutorial, and bring with you all work, including printouts of code and other results. Tutorials cannot function properly unless you do the work in advance.
- Data & Analysis tutorial exercises are not assessed, but they 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: if you are ill or otherwise unable to attend one week then email your tutor, and if possible attend another tutorial group in the same week.

### Introduction

In this tutorial we will work on Information Retrieval (IR). We will focus on two aspects of the information retrieval problem, as discussed in lectures: evaluation of IR systems, and choice of retrieval model.

### 1 Evaluating an Information Retrieval System

Consider the following information retrieval scenario: Suppose that it has been found at the Royal Infirmary of Edinburgh that due to equipment malfunction, the results of blood tests taken on 21/12/2010 are less trustworthy if the patient was diabetic. The hospital would like to contact all diabetic patients who had any kind of blood test on that day, to repeat the test. The hospital uses an information retrieval system to identify these patients. Suppose the collection of patients' medical records contains 10000 documents, 150 of which are relevant to the above query. The system returns 250 documents, 125 of which are relevant to the query.

- (a) Calculate the precision and recall for this system, showing the details of your calculations.
- (b) Based on your results of (a), explain what the two measures mean for this scenario. How well would you say that the hospital's information IR system works?
- (c) According to the precision-recall tradeoff, what will likely happen if an IR system is tuned to aim for 100% recall?
- (d) For the given scenario, which measure do you think is more important, precision or recall? Why? Given your answer, what value would you give to the weighting factor  $\alpha$  when calculating the F-score measure for the hospital's IR system?
- (e) In Informatics 1: Computation and Logic you have discussed the properties of soundness and completeness. Can you relate them to precision and recall of an IR system?

## 2 Information Retrieval Model

You are looking for information on the **Economic Recession in Scotland** in a large document collection. You decide to search using the terms: **economy**, **recession**, **Scotland**, **banks** and **business** using an *information retrieval system* and this recommends three possible documents. You are given the frequency of each of the terms in each document, shown in the table below:

Terms	economy	Scotland	recession	banks	business
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

You have no additional information about the documents; and to actually retrieve any one document will cost money.

- (a) One possible measure for determining which of the 3 documents is the *cosine* similarity measure, which measures the cosine of the angle between the query vector and that of each document. Compute this measure, for all three documents.
- (b) Based on your results of (a), which document is the best match for this query? Why?
- (c) Do you agree with the results of this analysis? What are the strengths and weaknesses of cosine measure?