COM3110: Text Processing

Introduction

Mark Hepple
Department of Computer Science
University of Sheffield
m.hepple@dcs.shef.ac.uk

Course Details

Instructors	Mark Stevenson (marks@dcs.shef.ac.uk) Room G27 @ Regent Court
	Mark Hepple (m.hepple@dcs.shef.ac.uk) Room G28d @ Regent Court
Classes	Lecture 1: Monday, 12:10, SG-LT02 (Mappin)
	Lecture 2: Thursday, 14:10, SG-LT01 (Mappin)
	Lab/Tutorial: Thursday, 10.00, Regent Court
	— to take place as announced (not every week)
Homepage	www.dcs.shef.ac.uk/~marks/campus_only/com3110
Assessment	100% Exam
Office hours	Email requests for appointments

Course Goals

- Develop an understanding of the problems of handling large large volumes of digitally stored text.
- Acquire familiarity with techniques for handling text.
- Develop ability to construct simple systems for applying such techniques.
- Develop an understanding of the basic problems and principles underlying text processing applications.

Prerequisites:

- Interest in language and basic knowledge of English.
- Some mathematical basics, e.g. basic probability theory
- Some programming skills.

Motivation

What is text processing and why study it? Proposed definition:

> *The creation, storage and access of text in digital form by computer*

Reasons for studying text processing now include:

- **The Web**
  - *Access* – more text than ever, available to more people than ever, in more languages than ever
    - widely discussed problem: *information overload*
    - premium on technology that can facilitate *information access*
  - *Creation* – automatic creation/update of web content

Motivation (contd)

. . . reasons for studying text processing (contd):

- **Metadata** – databases are out; text is in
  - *Access* – embedded semantic tags mean programs can crawl text sources and locate specific information
  - *Creation* – automatic creation/update of metadata
- **Convergence with NLP**
  - *NLP* (natural language processing) seeks to build programs that can “understand” texts
  - *Text Processing* – usually seen to have more modest, engineering aims
  - *Convergence* – increasingly they are borrowing ideas and techniques from each other
    - particularly in area of *statistical language processing*

Applications: Text Processing or NLP?

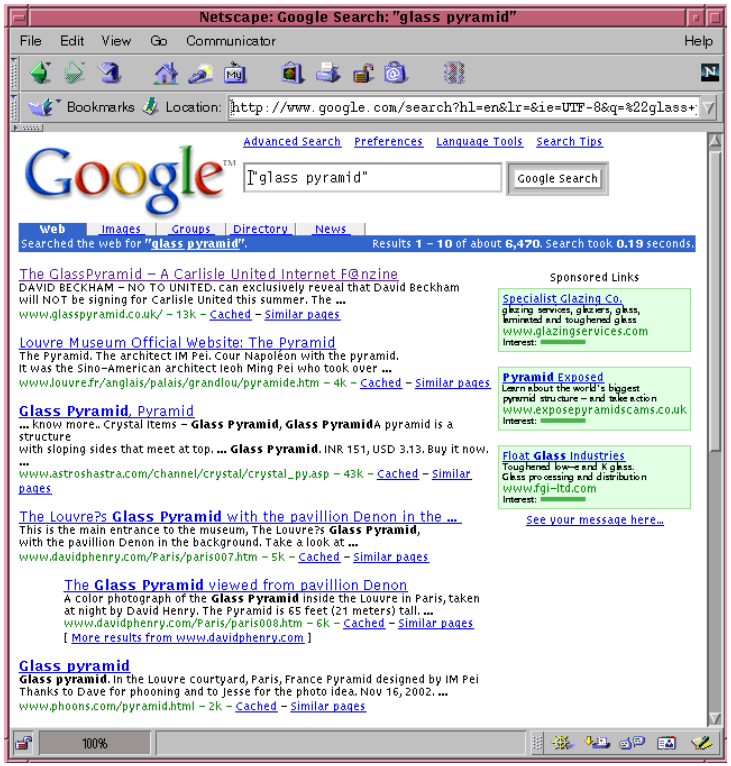
Distinction commonly seen in terms of whether task requires some ‘understanding’ of language, or special linguistic knowledge.

- Information Retrieval
- Text Categorisation
- Automatic Summarisation
- NL Generation
- Machine Translation
- Information Extraction

Applications: IR

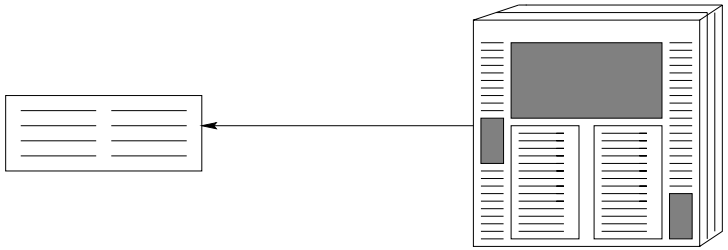
IR is concerned with developing algorithms and models for retrieving relevent documents from text collections.

- task of extracting the required information left to user

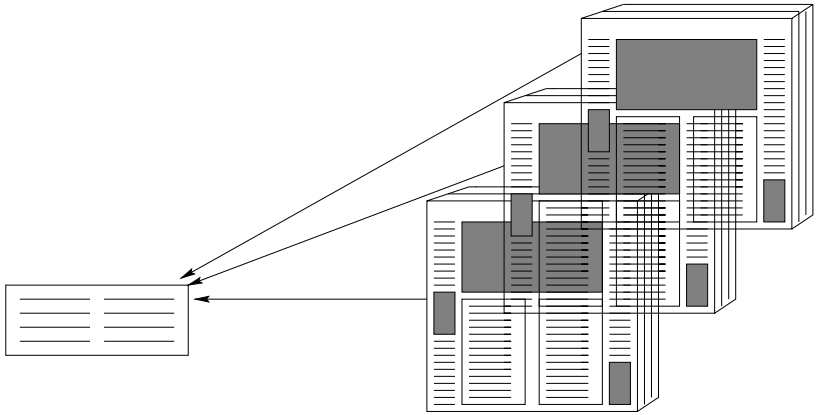


Applications: Summarization

Single Document



Multiple Document



Applications: Summarization (contd)

How does summarization work?

- Extract (e.g.) 25% of the initial document. What to extract?
  - sentences;
  - phrases;
  - words.
- How do you decide which parts to extract?
  - select at random;
  - select beginning of document;
  - select salient units.
- How do you decide whether one method works better than another?

Applications: Machine Translation

- Translate text from one language to another

  e.g. English to French and/or vice versa

- Write a computer program to do the translation.

- **Very difficult problem!**

- Requires immense amount of knowledge about language and the world.

- Learn from corpora that are translations of each other.

Course Content: Major topics

- Programming for text processing
  - *PERL* programming language
  - provides excellent *pattern matching* facilities
  - easily usable complex data structures, suited for TP
- Shallow linguistic analysis for TP applications
  - stemming / morphological analysis
  - part-of-speech tagging
  - word sense disambiguation
- Some major applications
  - Information Retrieval (IR)
  - Machine Translation

Reading

Major sources:

- PERL programming:
  - R. I. Schwartz and T. Phoenix, Learning Perl, 3rd edition, O'Reilly, 2001.
  - L. Wall, T. Christiansen and R. I. Schwartz, Programming Perl, 2nd edition, O'Reilly, 1996.
- Techniques and linguistic background:
  - C.D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.