COM6115: Text Processing

Introduction

Mark Hepple

Department of Computer Science
University of Sheffield

Course Details

Lecturers Mark Hepple (m.hepple@sheffield.ac.uk)
Chenghua Lin (c.lin@sheffield.ac.uk)

- COM6115 module homepage — links to it from:
  - MOLE unit
  - dept module description page
  - my homepage at: staffwww.dcs.shef.ac.uk/people/M.Hepple
  - module homepage is *campus-only accessible*
    - so run VPN for off-campus access
- Consult the homepage for:
  - all key course details
  - lecture materials
  - assignment
  - past exam papers
  - announcements

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):

- **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*

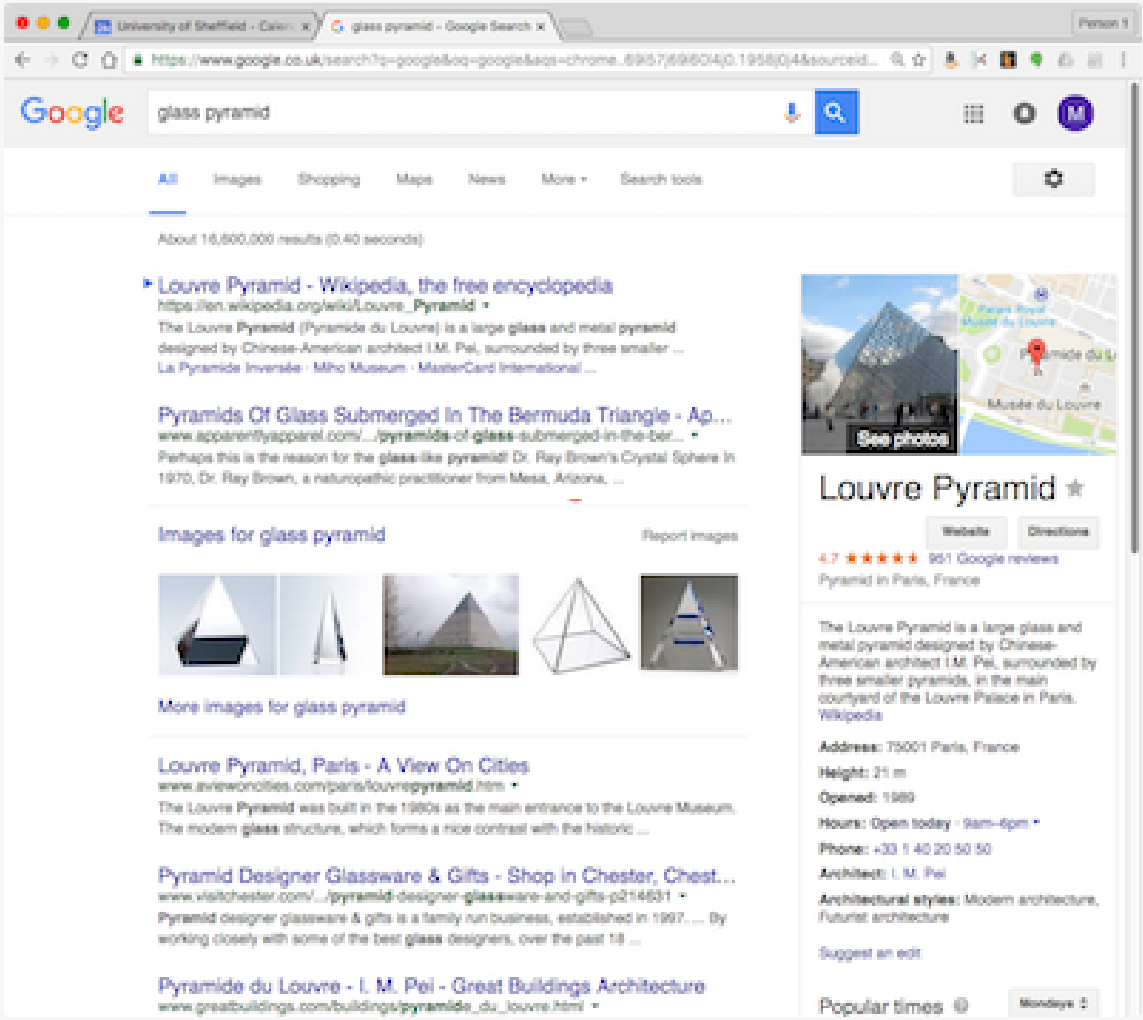
Applications: Text Processing or NLP?

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

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

Applications: IR

Information Retrieval (IR): concerned with developing algorithms and models for retrieving *relevent* documents from text collections.



Applications: IR (contd)

- Text collection = some set of 'documents'
  - originally, few hundred/thousand electronically stored documents, e.g. journal paper abstracts
  - now, billions of pages on the WWW
- Query: user indication of what s/he wants
  - commonly, just 2 or 3 words — good basis for retrieval?
- How decide what docs are relevant?
  - how decide if one method works better than another?
- Much work is still left to the user:
  - task of selecting which of returned docs *are* relevant
  - task of *extracting* the relevant information

Applications: IE

- IR contrasts with *Information Extraction* (IE).
- IE recognises *specific* information in documents, making it available to subsequent automated processes
  - type information to be extracted must be decided *in advance*
- Information recognised can be:
  - extracted and stored in a structured record, e.g. database system
  - stored in document itself as embedded mark-up

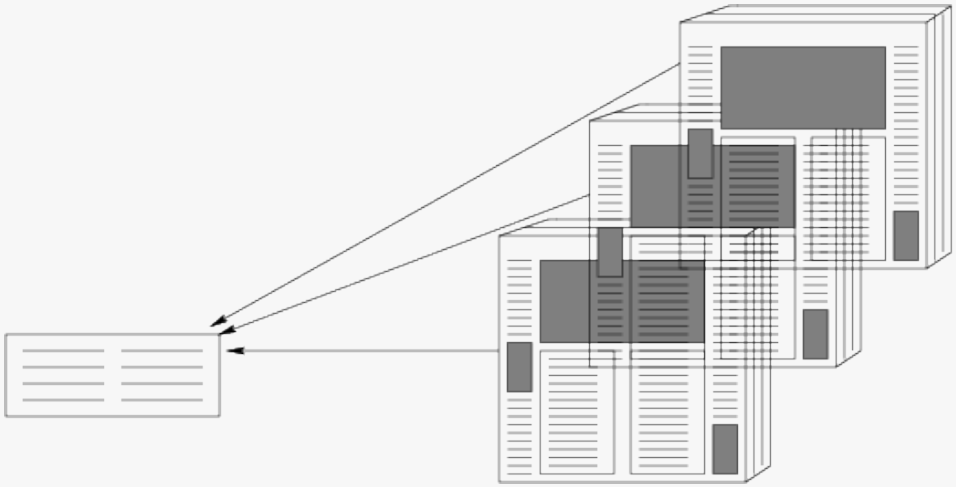
Applications: Text Categorization

- Task: automatically assign texts to different categories
  - e.g. for email — assign to categories: *junk* vs. *non-junk*
  - e.g. for newspaper articles — assign to categories: *sport* vs. *politics* vs. *other*
- Usually achieved by having a set of documents that are representative of each category
  - use statistical/probabilistic methods to decide which set of documents a new document is most like

Applications: Summarization

Single Document

Multiple Document



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 Outline

- Programming for Text Processing (with Python)
- Information Retrieval
- Natural Language Generation
- Information Extraction
- Sentiment Analysis

Reading

Major sources:

- Programming — see module homepage for suggestions
- Information Retrieval
  - Baeza-Yates and Ribeiro-Neto, Modern Information Retrieval. New Yorl: ACM Press, 1999.
  - C. Manning, P. Raghavan and H. Schtze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- General:
  - C.D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
  - D. Jurafsky and J. Martin, Speech and Language Processing, Prentice-Hall, 2007 (2nd edn).