COM3110/4115/6115: Text Processing

Introduction

Mark Stevenson

Department of Computer Science
University of Sheffield

Course Details

Lecturers Mark Stevenson (mark.stevenson@sheffield.ac.uk)
Rob Gaizauskas (r.gaizauskas@sheffield.ac.uk)

- Common module homepage for COM 3110 / 4115 / 6115 at:
  - links to it from MOLE page, from dept module description page, and from my homepage at:

`staffwww.dcs.shef.ac.uk/people/M.Stevenson/campus_only/com3110/`

  - campus-only accessible (so run VPN for off-campus access)
- Consult the homepage for:
  - all key course details
  - lecture materials
  - lab materials
  - assignments
  - past exam papers

Timetable

- Lectures
  - Tuesday 15:00 - 15:50 (Diamond, LT-05)
  - Wednesday 11:00 - 11:50 (Broad Lane Block, LT-07)
- Lab class
  - Thursday 9:00-9:50 (Diamond room 201)
  - Weeks 1 - 4

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)

- **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*

Course Outline

- Programming for Text Processing (with Python)
- Information Retrieval
- Text Compression
- Information Extraction
- Sentiment Analysis

Assessment

Depends on which module code you are sitting the course under:

- COM3110 (10 credit version)
  - Assignment on Information Retrieval (25%)
  - Exam (75%)

- COM4115 or COM6115 (15 credit versions)
  - Assignment on Information Retrieval (25%)
  - Assignment on Sentiment Analysis (25%) Extra assignment!
  - Exam (50%)

- Dates
  - Assignment on Information Retrieval
    - Release week 4, due week 7 (TBC)
  - Assignment on Sentiment Analysis
    - Release week 8, due week 11 (TBC)

Reading

Major sources:

- Programming — see module homepage for suggestions
- Information Retrieval:
  - C. Manning, P. Raghavan and H. Schtze, Introduction to Information Retrieval, Cambridge University Press. 2008.
  - R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval. New York: ACM press, 1999.
- 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).