Lecture 1a: administration

Trevor Cohn
tcohn@unimelb.edu.au

COMP90042, 2015, Semester 1
Course overview

Introduction to text processing

▶ Information Retrieval (Weeks 1 to 6)
▶ Natural language processing (Weeks 7 to 12)

Recommended Textbooks

▶ Manning, Raghavan and Schtze (MRS): Introduction to Information Retrieval (PDFs online)
▶ Jurafsky & Martin (JM): Speech and Language Processing

Assessment

▶ Two assignments: 50% (details next week)
▶ Exam: 50%
▶ Hurdle: mass path both components
Pre-requisites and outcomes

Pre-requisites

- Some Python programming experience
- COMP90049 / COMP30018, “Knowledge Technologies”
- no knowledge of linguistics or statistics is assumed

Expectations

- develop Python skills
- keep up with readings
- classroom participation

Outcomes

- Practical familiarity with range of text analysis technologies
- Understanding of theoretical models underlying these tools
- Competence in reading research literature (including maths!)
Assessment

Assignments and Exam

- Two projects (50% total)
  - Individual work
  - First assignment is programming & analysis
due week 7 (TBC)
  - Second assignment is a research report
due week 12

- Exam (50%)
  - two hour
  - closed book
  - Covers content from lectures, tutorials and prescribed reading
Why is text processing important

The information society and ubiquitous computing

▶ Knowledge locked up in unstructured text and audio on the web
▶ Handheld devices

Multimodal dialogue systems including speech and language technologies:

▶ e.g., Apple Siri, IBM Watson
▶ the race for better IR and NLP is on, with Google, Microsoft, Facebook etc leading the way
After 5 decades of research:

- Language is becoming a key component of HCI and Web technologies
- Some technologies:
  - automatic hyphenation, spelling correction, OCR
  - speech recognition and understanding
  - text retrieval, understanding and summarization
  - machine translation
  - spoken dialogue
- many useful systems already exist
  - Parsing, coreference, entity recognition
    http://nlp.stanford.edu:8080/corenlp/process
  - Machine translation http://translate.google.com
  - Text to speech
    http://www2.research.att.com/~ttsweb/tts/demo.php
Natural Language Processing

- Primary concern: creating programs to perform useful and interesting tasks with human languages
  **Understanding, generation, learning**
- Secondary concern: using computational metaphors leading to deeper understanding of human language
- Foundations in: CS, linguistics, maths, stats, electrical engineering, psychology
- Studying NLP involves:
  - Studying natural languages
  - Formal representations of linguistic information
  - Algorithms for their manipulation
Information Retrieval

- Focus is on information access
  - From unstructured corpora, e.g.,
    - Text documents
    - Web pages
    - Mixtures of raw text and markup (e.g., wikipedia, citation graphs)
  - Based on a query, and returning a set of matching documents
    - e.g., Peter AND Paul AND Mary, Britny Spears, Title:Cats AND NOT Body:Dogs
- Typically shallower methods than NLP
  - But overall there are many similarities and shared techniques
Information Retrieval

- Starts with an *information need*
  - E.g., wanting to know how the various technologies work for internet broadband
- User formulates a *query*
  - internet broadband work
- A system *retrieves* a set of documents that match the query
  - How broadband Internet works - Explain that Stuff
  - Naked DSL - How it Works - WhistleOut
  - Broadband - Wikipedia, the free encyclopedia
- Assess whether the information need is met
  - May reformulate the query, e.g., including DSL or other keywords
- Evaluation in terms of *relevance of results*
  - Precision: what fraction of the documents returned were relevant?
  - Recall: of the relevant documents in the collection, what fraction were returned?
Course outline: information retrieval

- Boolean retrieval methods
- The term-document matrix and term frequency scoring
- TDM as a vector space; cosine distance; document similarity
- Search and information retrieval using cosine distance
- Evaluation of effectiveness
- Probabilistic document similarity
- Language modelling for search and retrieval
- Anchor text, page rank, and other graphical models
Course outline: natural language processing

- Part of speech tagging
- Information extraction & Named entity identification
- Hidden Markov models and Conditional Random Fields
- Text classification
- Syntactic parsing
- Machine translation
What we will not be covering

- Many machine learning models (in any depth)
- Text clustering (LSI, LDA, etc)
- Automatic text summarization
- Dialogue systems
- Sentiment analysis
- Semantics
- Speech recognition
- Text compression
- Most engineering, optimization, and efficiency issues
- ...
Teaching Staff

Lecturer *Trevor Cohn*
- Research into NLP, especially parsing, translation and statistical modelling of text
- Recently joined unimelb, following roles at U. Sheffield & U. Edinburgh

Teaching Assistant *Jeremy Nicholson*
- Highly experienced in NLP and text processing
- Head tutor for several CIS subjects
Recommended Text

IR part follows this text closely:

Other Texts

Selective readings from:

<table>
<thead>
<tr>
<th>Time</th>
<th>Contact Person and Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday 11:00 - 12:00pm</td>
<td>Doug McDonell-309</td>
</tr>
<tr>
<td>Friday 2:15 - 3:15pm</td>
<td>Doug McDonell-309</td>
</tr>
<tr>
<td>Wednesday 5:15 - 6:15pm</td>
<td>Alice Hoy-109 (Comp Lab)</td>
</tr>
<tr>
<td>Thursday 4:15 - 5:15pm</td>
<td>Alice Hoy-236 (Comp Lab)</td>
</tr>
<tr>
<td>Friday 3:15 - 4:15pm</td>
<td>Alice Hoy-211 (Comp Lab)</td>
</tr>
<tr>
<td>Thursday 2:15 - 3:15pm</td>
<td>Doug McDonell 6.25 (contact hour)</td>
</tr>
</tbody>
</table>
Consultation

- Regular consultation, Thursday 2:15pm to 3:15pm; room 6.25
- Email me at tcohn@unimelb.edu.au
- Email Jeremy at nj@unimelb.edu.au
- I will post answers to a FAQ section on the LMS
Workshops

- Weekly workshops in Alice Hoy labs
- Mixture of computer programming and analysis work
- You will need to prepare ahead of time
- Taken by Jeremy Nicholson