Lecture 1a: administration

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 1a

COMP90042: Web Search and Text Analysis

Course outline

- Text-document matrix (Week 1)
- Geometric models (Weeks 2 to 5)
- Probabilistic models (Weeks 6 to 9)
- Beyond the TDM (Weeks 10 to 12)

Course outline: text-document matrix

- Parsing, stemming, stopping, and other transformations
- The “bag-of-words” representation of text
- The text-document matrix
- Term frequency scoring

Course outline: geometric models

- TDM as vector space; cosine distance; document similarity
- Search and information retrieval using cosine distance
- Text clustering
- Text classification using support vector machines
- SVD methods and LSA

Course outline: probabilistic models

- Probabilistic document similarity
- Language modelling for search and retrieval
- Text classification using:
  - Naive Bayes
  - Maxent (logistic regression)
- Topic modelling

Course outline: beyond bag-of-words

- POS tagging, phrase and named entity identification
- Anchor text, page rank, and other graphical models
- Cross-lingual information retrieval
- Perhaps other things ... (e.g. CJK segmentation)

What we will not be covering

- Machine translation
- Sentiment analysis
- Deep parsing or other advanced NLP techniques
- Information extraction
- Automatic text summarization
- Text compression
- Engineering, optimization, and efficiency issues
- ...

About myself

- Industry consultant in text analytics (not an academic, though active as independent researcher)
- Masters by thesis in distribution information retrieval (UMelb)
- PhD in information retrieval evaluation (UMelb)
- Postdoc in e-discovery (University of Maryland)

Pre-requisites and outcomes

Pre-requisites

- Python programming skills
- COMP90049 / COMP30018, "Knowledge Technologies"

Outcomes

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

Contact hours

Tuesday 9:45 - 10:45am	Th:Doug McDonell-309 (Denis Driscoll Theatrette)
Wednesday 2:15 - 3:15pm	Th:Chemical & Biomolecular Engineering-Theatre
Wednesday 4:15 - 5:15pm	Alice Hoy-236 (Comp Lab)
Thursday 2:15 - 3:15pm	Alice Hoy-222 (Comp Lab)

Consultation

- No office on campus!
- Regular consultation, Wednesday 3:15pm to 4:15pm; room to be announced
- Email me at `william@williamwebber.com`
- I will curate a mailing list

Expectations on students

You are CS masters students at Australia's top university. I expect competence accordingly.

- You will be assessed to this standard!
- Code to be neat, (comparatively) bug-free, documented
- Expected to be able to install software, packages, tools
- Expect fluency in written English

I will use all tools available to me to detect plagiarism, and will be *very strict in prosecuting it.*

Student weekly work

- Readings will be provided in advance of lectures (not this week though). Expect c. 2 hours of reading.
- Weekly worksheets need to be finished in full. Workshops are *only to begin* these worksheets.
- All workshop work (and all project work) to be committed to shared subversion repository (details at first workshop).
- We are over-full on workshops. Hopefully students will drop out.
- I am taking both workshops.

Assessment

- 10% on workshop work, as committed to Subversion repository. (Option: you may instead give 20 minute research presentation at workshop, but you need to email me by the end of next week to take this option).
- 40% on projects. There will be two projects:
  - First project due end of Easter non-teaching break. Set work.
  - Second project due at end of semester. Free choice research-y project.
- 50% on final exam. This will be closed book. Everything on the course (in readings, lectures, workshops, projects) is examinable!

Learning resources

- Lecture notes are primary resources.
- No text book as such, but following texts are useful:
  - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. (Available free online)
  - Charu C. Aggarwal and ChengXiang Zhai (ed.), *Mining Text Data*, Springer, 2012. (c. $US 150)
- Citations to other readings will be given as required.
- Wikipedia is a very good place to start!