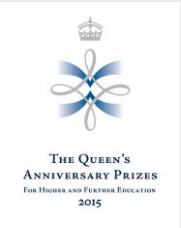
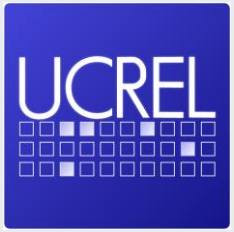
Lancaster University

UCREL NLP Summer School Session 1: Web scraping theory and methods

Dr Stephen Wattam @StephenWattam
Dr Paul Rayson @perayson
Andrew Moore @apmoore94
School of Computing and Communications
Lancaster University





Key sources of information

Lancaster University

- Steve Wattam's Phd Thesis:
  - https://stephenwattam.com/cv/papers/thesis.pdf
- ACL SIGWAC and annual workshops:
  - https://www.sigwac.org.uk/
- Kilgarriff and Grefenstette (2003) Introduction to the Special Issue on the Web as Corpus. Computational linguistics.
  - http://dx.doi.org/10.1162/089120103322711569
- WaCky corpus collection
  - Baroni et al (2009) The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. LRE. http://dx.doi.org/10.1007/s10579-009-9081-4
- Mehler et al (eds) (2011) Genres on the web. [Book]
  - http://www.springer.com/gb/book/9789048191772
- Schäfer & Bildhauer (2013). Web Corpus Construction. [Book]
  - http://sites.morganclaypool.com/wcc/

Why use web data?

- Scale
  - Sufficient examples of study RQ features
- Availability
  - new genres
- Googleology
  - precision/recall, ranking, personalised results, estimated counts
- Accessibility
  - Download for further local processing
- Cost

In the old days (Brown/LOB), you actually had to visit the British Library and/or retype data from books or newspapers!

Browsing vs searching

Lancaster University

- Statistical sampling strategies
- Representativeness
- Browsing
  - web crawlers
- Searching
  - using web search engines and BootCat approach

Ethical considerations

- Eysenbach & Till (2001) Ethical issues in qualitative research on internet communities. BMJ.
  - http://dx.doi.org/10.1136/bmj.323.7321.1103
- Seale C, Charteris-Black J, MacFarlane A, et al. Interviews and internet forums: a comparison of two sources of data for qualitative research. Qual Health Res 2010;20:595–606.
  - http://dx.doi.org/10.1177/1049732309354094

Summary points

Internet communities (such as mailing lists, chat rooms, newsgroups, or discussion boards on websites) are rich sources of qualitative data for health researchers

Qualitative analysis of internet postings may help to systematise and codify needs, values, and preferences of consumers and professionals relevant to health and health care

Internet based research raises several ethical questions, especially pertaining to privacy and informed consent

Researchers and institutional review boards must primarily consider whether research is intrusive and has potential for harm, whether the venue is perceived as "private" or "public" space, how confidentiality can be protected, and whether and how informed consent should be obtained

Legal/copyright considerations

Lancaster University

- Republishing the data is problematic although exceptions exist in some countries for text mining (UK, France, US)
- Ways around this:
  - Host the corpora and restrict concordances to fair use
  - Release URL lists / Tweet IDs
- BYU corpora
  - approach for download is to blank out 5% of words
- COW corpora
  - sentences are shuffled
- But what effect does all this have on NLP results derived from these datasets?
- CommonCrawl solution
  - subset of 10B words from Creative Commons pages: Habernal et al (2016) C4Corpus: Multilingual Web-size corpus with free license. http://www.lrec-conf.org/proceedings/lrec2016/summaries/388.html

Reliability and comparability of data

Lancaster University

- Do you need to filter or clean the data?
  - Many different types of 'noise'
- For lexicography
  - Kilgarriff et al: Good Example finder algorithm.
- Metadata
  - Age, gender, etc information
  - Hoffmann (2007) Processing Internet-derived Text—Creating a Corpus of Usenet Messages http://dx.doi.org/10.1093/llc/fqm002
- Do you need native language examples for national corpora?
- Document attrition
  - S. Wattam, P. Rayson & D. Berridge: "Document Attrition in Web Corpora: an Exploration", Proceedings: Language Resources and Evaluation '12, Istanbul, May 2012 http://www.lrec-conf.org/proceedings/lrec2012/pdf/806_Paper.pdf
- Search engine counts and ranking problems
  - http://www.infolab21.lancs.ac.uk/news_and_events/news/blogspot?article_id=1547

We will return to this in session 2: "Web as corpus creation and cleaning"


how?
discover
questions
asking questions
where?
?
why
challenge
who?
QUESTIONS
clues
ask
what?
who?
discover
when?
Knowing
clues
why?
Knowing
investigation
?
how
ask
investigation