UCREL NLP Summer School Session 2: Web as corpus creation and cleaning

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

Key sources of information

Lancaster University

- In addition to those from session 1 …
- Cleaneval shared task organised by SIGWAC
  - http://cleaneval.sigwac.org.uk/
- Boilerplate removal tools
  - https://github.com/miso-belica/jusText
  - http://corpus.tools/wiki/Justext
  - https://boilerpipe-web.appspot.com/

Picking up from session 1 …

1. Inspect the data that you downloaded in session 1
2. Thinking about requirements of storage, retrieval, analysis and annotation (coming later in the standard NLP pipeline, and the summer school) – what potential problems do you discover?
3. Spend 20 minutes doing this and then we'll compare notes

Time’s up …

- Let’s compare notes

Brief reminder: representativeness and statistical sampling strategies

Lancaster University

- It's easy to grab all the data first (e.g. Reddit corpus) and think about what you want to do with it later! ☺
- Worth reminding yourself of your research questions and how you want to design your experiment/paper/thesis
- What is the data (twitter, blogs, news) that you've downloaded representative of? And what can you claim on the basis of it?

- Miller et al (2015) The road to representativity: a Demos and Ipsos MORI report on sociological research using Twitter. http://www.demos.co.uk/project/the-road-to-representivity/

Potential issue: ‘metadata’ and corpus structure

- Metadata: what do you need to preserve as dimensions or variables for your study
  - Age, gender, location information
  - Dates and titles from HTTP headers
  - Threads in online forums
  - Emails and response threads (preserving quoted text?)
  - Duplication / text reuse

- Hoffmann (2007) Processing Internet-derived Text: Creating a Corpus of Usenet Messages. http://dx.doi.org/10.1093/llc/fqm002

Potential issue: ‘noisy’ data

Lancaster University

- Spelling variation
- Ungrammatical sentences
- Non-native language
- OCR from PDF sources

- What you do with these potential problems depends on your application: national corpus collection, lexicography?

- Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlý (2008). GDEX: Automatically finding good dictionary examples in a corpus. In Proceedings of the 13th EURALEX International Congress. Spain, July 2008, pp. 425–432. https://www.sketchengine.co.uk/user-guide/user-manual/concordance-introduction/gdex/

Further Reading Material

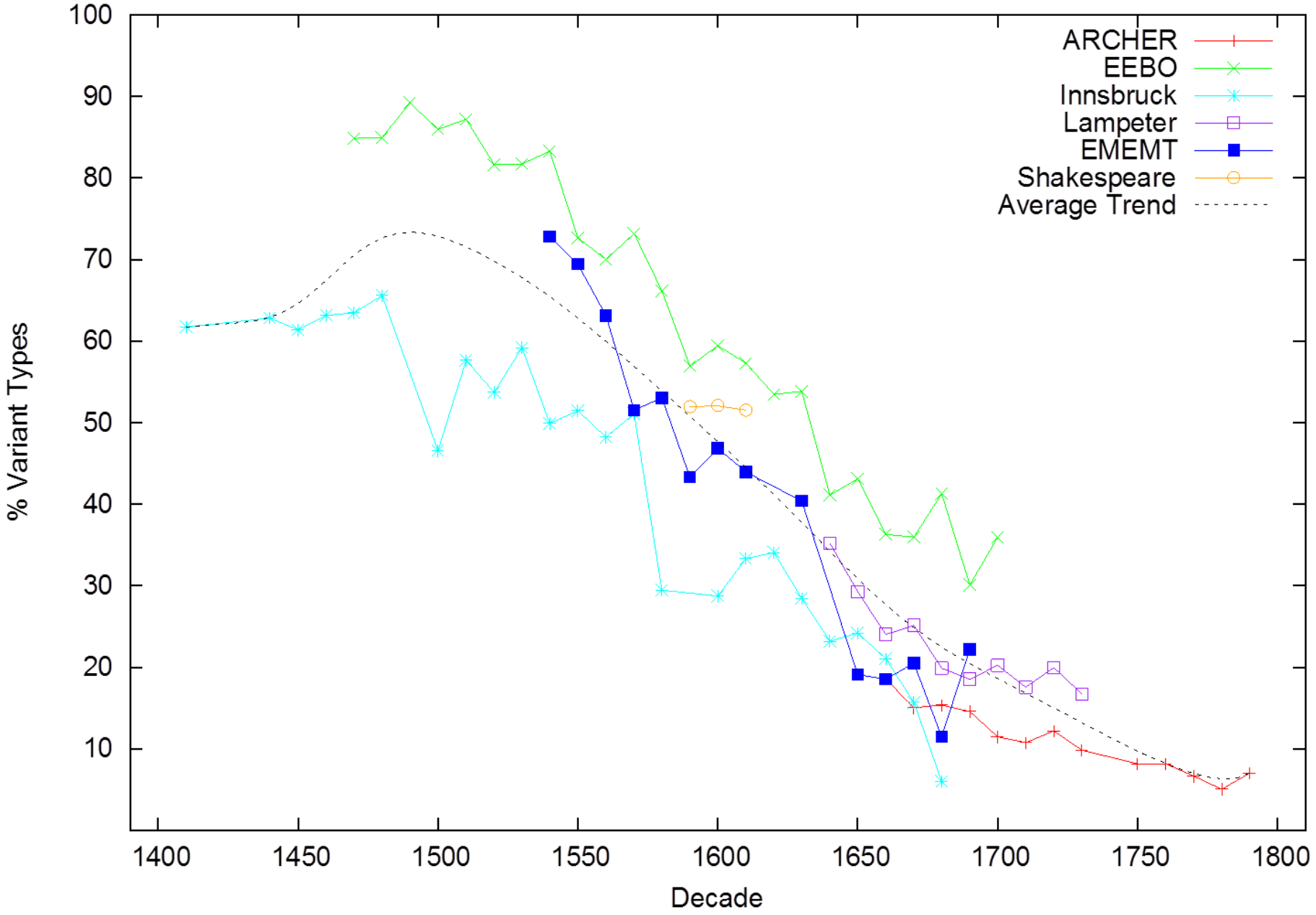
Potential issue: ‘noisy’ data

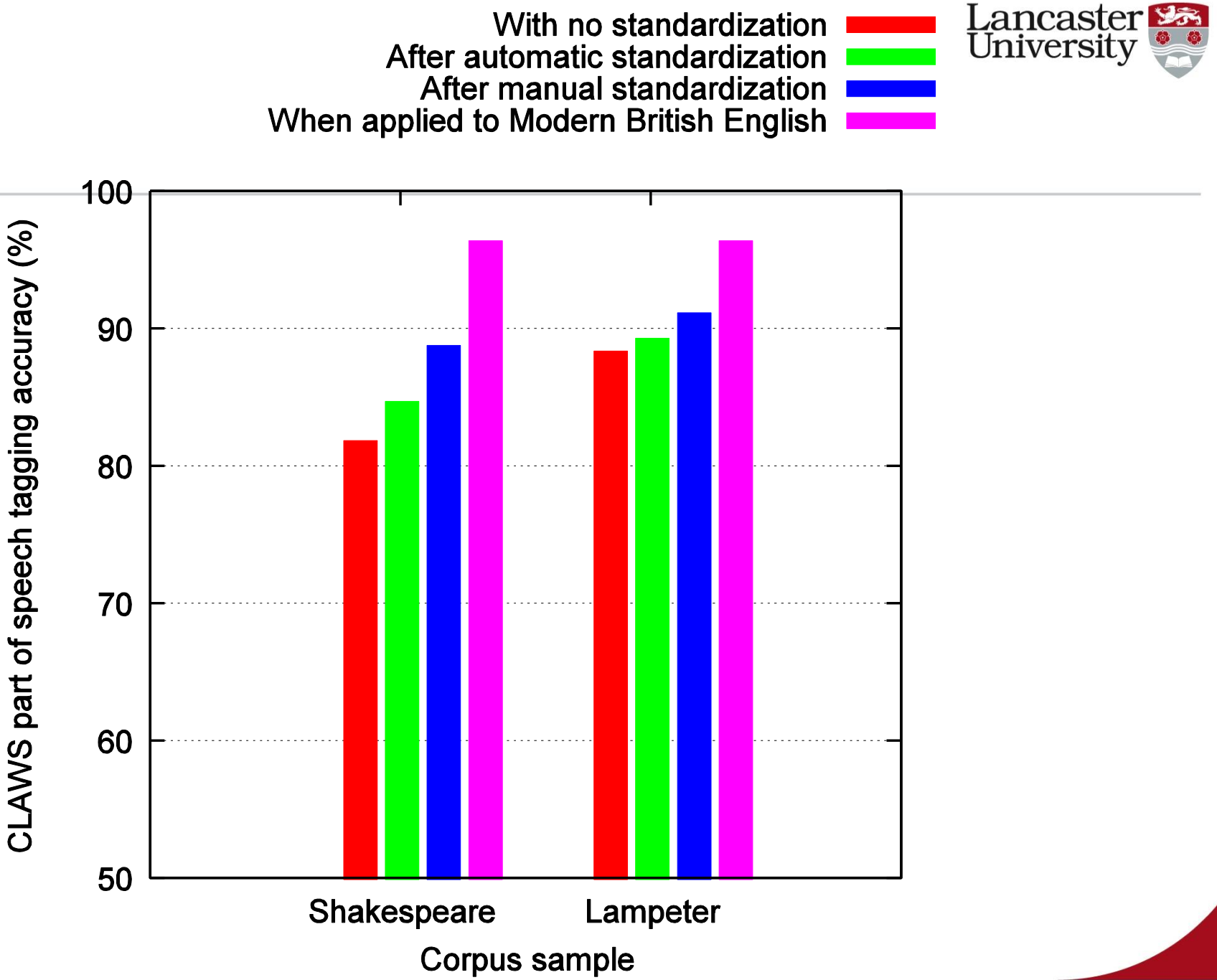
Lancaster University

- Of course, you might be studying spelling variation and need to preserve it for study e.g. historical data, SMS, twitter, OSN

Though I **speake** with the tongues of men & of Angels, and **haue** not charity, I am become as sounding **brasse** or a tinkling cymbal. And though I **haue** the gift of **prophesie**, and **vnderstand** all mysteries and all knowledge: and though I **haue** all faith, so that I could **remooue mountaines**, and **haue** no **charitie**, I am nothing...

(Authorised Version of the Bible, 1611)

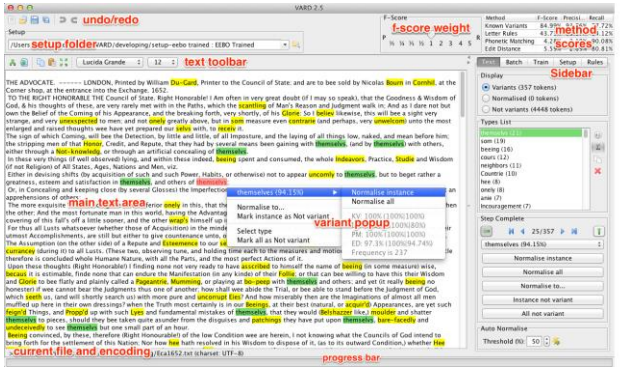

ARCHER
EEBO
Innsbruck
Lampeter
EMEMT
Shakespeare
Average Trend
% Variant Types
Decade
100
90
80
70
60
50
40
30
20
10
1400
1450
1500
1550
1600
1650
1700
1750
1800


With no standardization
After automatic standardization
After manual standardization
When applied to Modern British English
CLAWS part of speech tagging accuracy (%)
100
90
80
70
60
50
Shakespeare
Lampeter
Corpus sample

Potential issue: ‘noisy’ data

- Automatic semantic analysis of EmodE corpora
  - Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16. UCREL, Lancaster University, pp. 22 - 31.
- Automatic POS tagging of historical corpora
  - Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In proceedings of Corpus Linguistics 2007, July 27-30, University of Birmingham, UK.
- http://ucrel.lancs.ac.uk/vard/

Further Reading Material



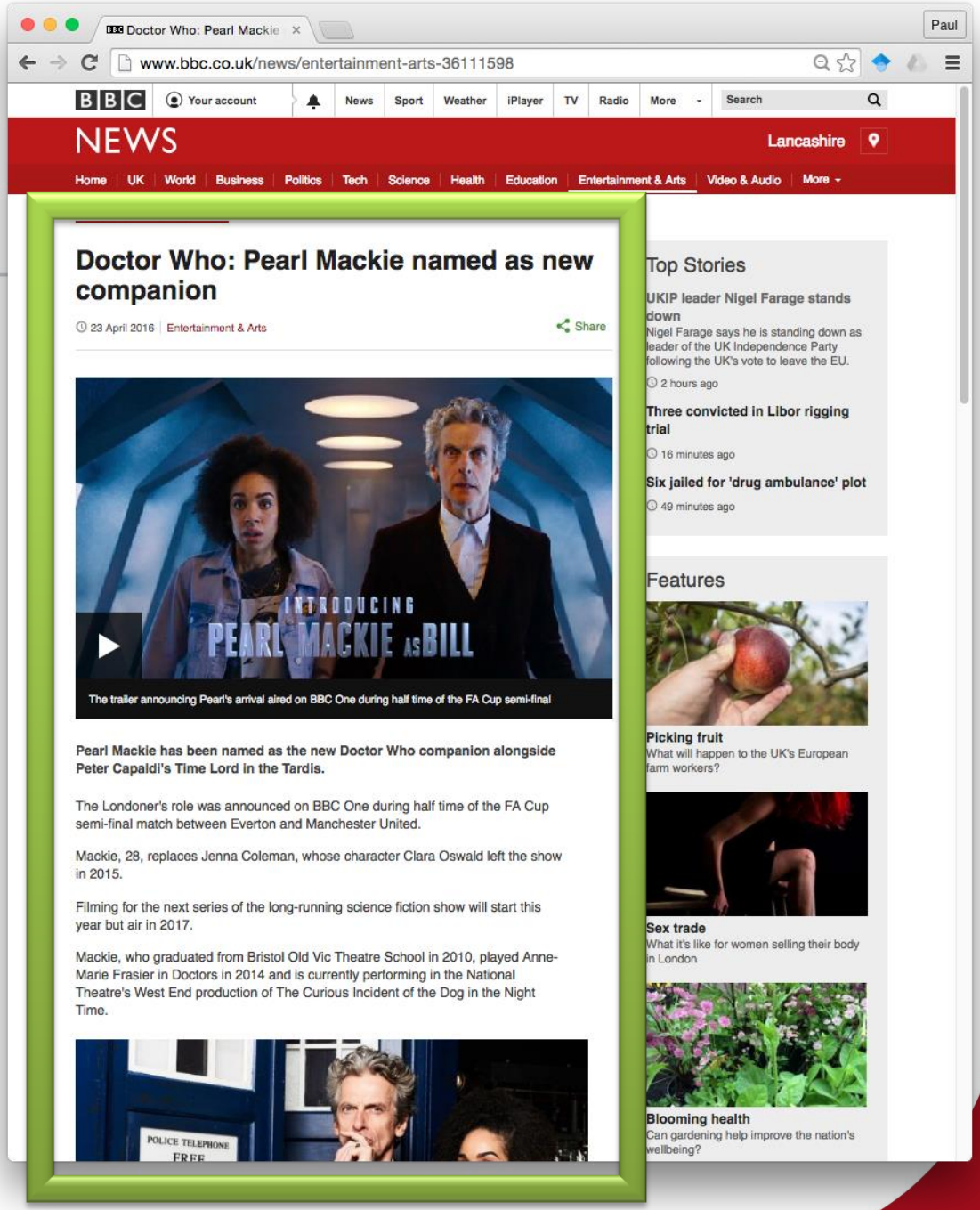
Potential issue: encoding

Lancaster University

- Character set normalisation
  - e.g. Headers say UTF-8 but actual file encoding is different

- Python Unicode docs:
  - https://docs.python.org/3/howto/unicode.html

Potential issue: boilerplate



Corpus creation: recommendations

Lancaster University

- Document all the steps that you've taken
- Release scripts alongside papers wherever possible
- If creating a corpus to release, you'll need to give serious thought to ethics and legal copyright issues.
- If distributing via links or tweet IDs, what about document attrition or deleted tweets?


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

Practical: Aim: turn your scraped data into a corpus

- Focussing on three of the main issues:
  - character set normalisation
  - boilerplate removal
  - preserving & standardising metadata from HTTP headers
- Git repo for S2: https://github.com/UCREL/web-cleaning

Task 1/5: Character set normalisation

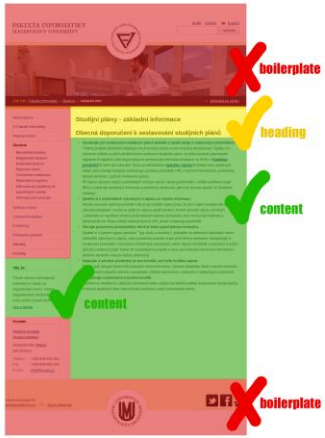
Lancaster University

- Potential mismatch between page header and actual encoding of a crawled page
- See README instructions on how to run ‘encoding’ script
- Investigate files with potential problems indicated by the log file output
- (Hacker option) Other pre-processing steps that you might consider creating and running:
  - Remove hashtags and URLs from the data

Task 2/5: Boilerplate removal

- We'll use jusText developed by Jan Pomikálek (Masaryk University) which is part of the Sketch Engine
  - http://corpus.tools/wiki/Justext

1. Read the algorithm linked from that page
   - http://corpus.tools/wiki/Justext/Algorithm
2. Try out the online demo using a BBC news webpage
   - https://nlp.fi.muni.cz/projects/justext/
   - http://www.bbc.co.uk/news/world-europe-36712550
   - Manually compare the original page and the filtered output side by side
3. Try out the online demo on one of the pages in your scraped corpus
4. Automatically process your whole corpus using 'boiler_removal' script
   - See README for instructions



Task 3/5: Manually examine your metadata in SQLite

Lancaster University

- Already done in session 1 or at the start of this session?
- If not, do so now!

- Open Sqliteman-1.2.2 folder
- Double click sqliteman
- File -> Open
  - web-corpus construction folder
  - output folder
  - Double click on metadata database
- Double click ‘output’ within Tables
- Then examine metadata in full view

Task 4/5: Normalise metadata (hacker option)

Lancaster University

- Missing or malformed data
- Some suggestions as to how …
  - Pull data out, normalise and push it back into SQLite.
  - Output as CSV
  - Or edit live directly in sqliteman

Task 5/5: Final export of metadata.

Lancaster University

- Run ‘export_metadata’ script to export metadata as CSV
- Edit script to export as TSV or JSON if you prefer
- See README for instructions …

You now have completed your corpus creation and cleaning!