OMGZ0rz WTF n00b u d0n7 5|*34k]33t!!1 u r n0t 4 h4><0rz!!!

SCC.413: Applied Data Mining

Dr. Alistair Baron
a.baron@lancaster.ac.uk
B53, Infolab21

Aims of module

- To apply some of the things you have learnt so far on “real-world” data.
- Learn how to collect and process online data to address business needs and research questions.
- Perform various analyses to extract meaningful information and gain insights.
- Understand the current trends of research analysing online activity.
- Learn new practical skills, especially for dealing with online text.

Overview (pipeline)

- Collection online (textual) data:
  - Web scraping
  - Social networks
- Extracting meaningful information:
  - Words, phrases, topics, language style, relationships…

- Topic modelling
- Summarisation
- Information extraction
- Sentiment analysis
- Stylometry

Week 14 (today)

- Introduction to module
  - Plan (subject to change)
  - Labs
  - Assessment
- Collecting online data
  - Web scraping
  - Using APIs

Week 15

- Pre-processing
  - Regular expressions
  - Cleaning
  - Tokenisation

Week 16

- Feature extraction
  - Frequency lists / bag of …
  - Annotations
  - Normalisation
  - (probably) word embeddings.
  - (or) corpus linguistics methods.

Week 17

- Analysis: classification

- Authorship analysis

Week 18

- Topic modelling
- Summarisation

Week 19

- Information extraction (Named Entities)
- Sentiment analysis.

Week 20

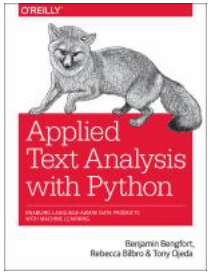
- Group research presentation

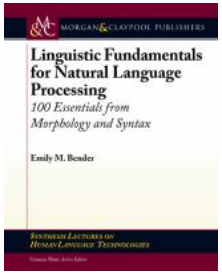
Reading list

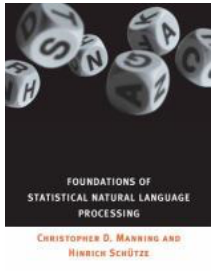
- (will be) available via Moodle.

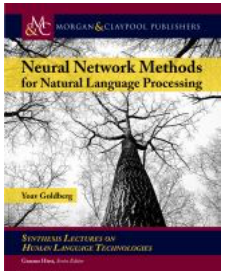
- Key texts available for free online:
  - http://www.nltk.org/book/
  - https://web.stanford.edu/~jurafsky/slp3/

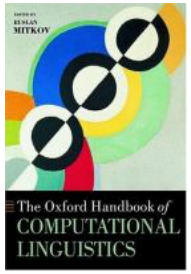
- Others:

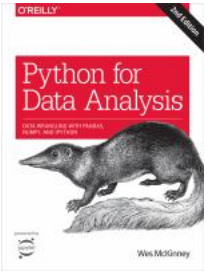






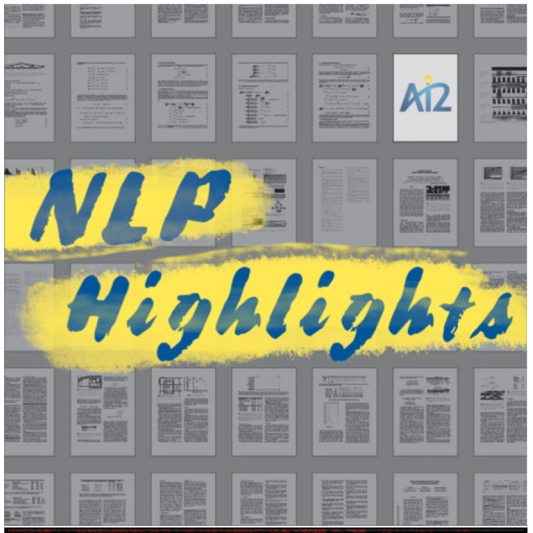






Listening list: podcasts

- Allen AI: NLP Highlights
  - https://allenai.org/podcasts/podcasts-all.html



- Claire Hardaker: en clair
- (forensic linguistics)
  - http://wp.lancs.ac.uk/enclair/

Labs

- Series of practical exercises each week, designed to support lecture material.
- Will prepare you for final assignment.
- Mainly using Python.
- Collaborative.
- TA: Edward Dearden

Assessment

- Weekly lab work (20%)
- Group research presentation (20%)
- Individual assignment (60%)

Assessment

- Weekly lab work
  - 20% of final mark.
  - Must be demonstrated in lab (same week or week after).
  - To show engagement and understanding of lab work. Not marked, but verbal feedback given.
  - Weeks 14-19 lab work: 1, 2, or 3 marks given depending on amount completed. (/18 marks)
  - Week 20: 2 marks given if final assignment work proposed and discussed in lab.

Assessment

- Group research presentation
  - 20% of final mark.
  - Due Friday Week 19.
  - Presented Wednesday Week 20.
  - 10 minute presentation.
  - Groups of 4-5 people.
  - 1 mark per group (unless severe lack of engagement).
  - Set topics, with seed papers.
  - Read papers, discuss, present summary.

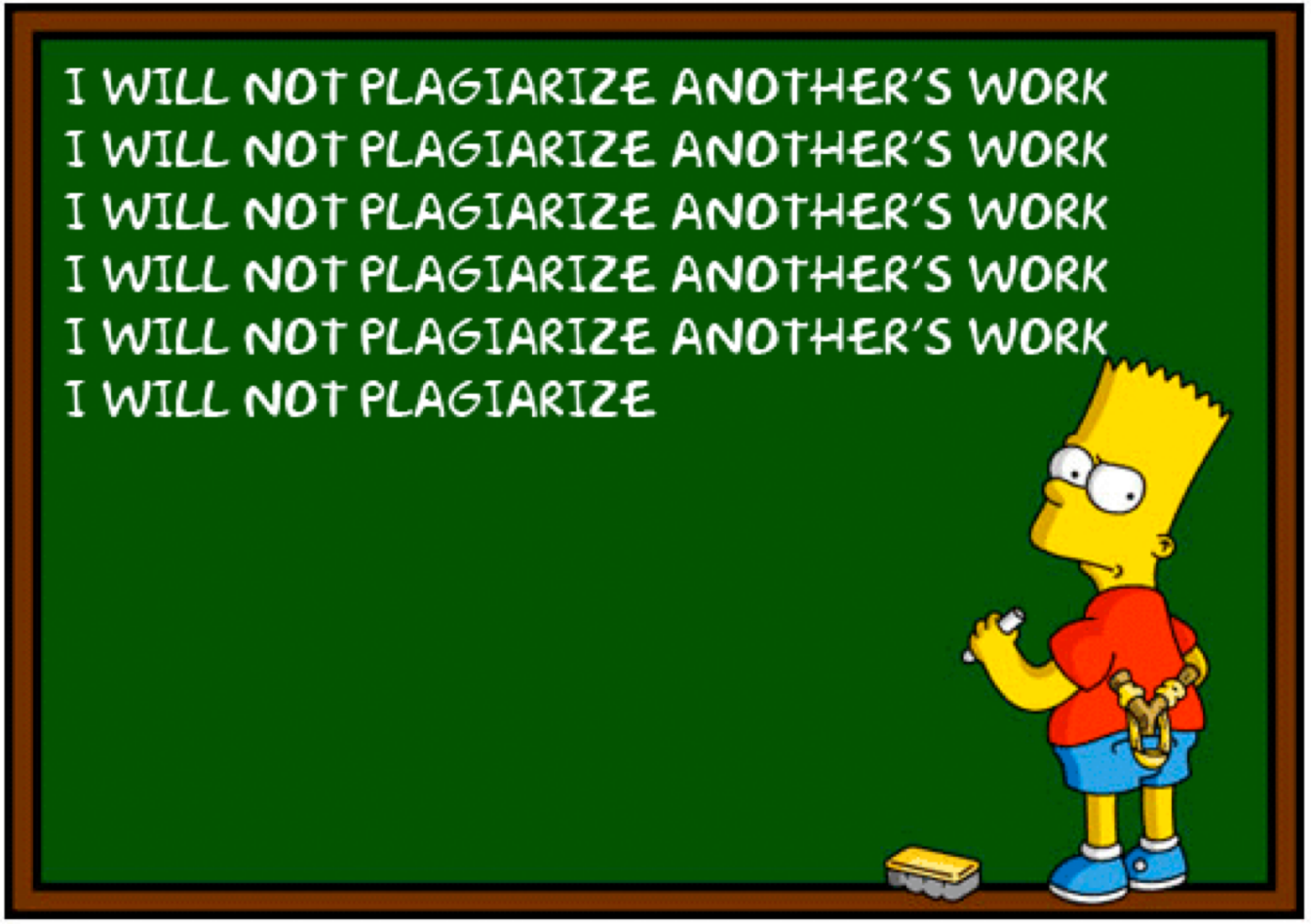
Assessment

- Final assignment
  - 60% of final mark.
  - Due: 3rd May.
  - Individual work.
  - A mini research project investigating a research question.
  - Some datasets will be provided.
  - 4-6 page workshop-style paper.

Feedback

- You can get feedback (and ask questions) on your work in the labs from myself and the TA.
- Verbal feedback will be given whilst checking weekly lab work.
- Written feedback on other assessment will be given within 4 weeks.

Plagiarism


I WILL NOT PLAGIARIZE ANOTHER'S WORK
I WILL NOT PLAGIARIZE ANOTHER'S WORK
I WILL NOT PLAGIARIZE ANOTHER'S WORK
I WILL NOT PLAGIARIZE ANOTHER'S WORK
I WILL NOT PLAGIARIZE ANOTHER'S WORK
I WILL NOT PLAGIARIZE

Plagiarism

- Passing off someone else's work as your own, including:
  - submitting (e.g.) code that someone else wrote
  - paying for someone else to do it for you
  - working on one piece of non-group work together as a group, and submitting it as individual work
  - sharing of code that you then possibly adapt

What I expect from you

- Integrity (no plagiarism, no faking results).
- Effort: **active** learning.
- Come to lectures: Wednesdays 9:15 – 12:00
- Come to labs: Wednesdays 14:00 – 18:00
- Take notes.
- Ask questions, in lectures and in labs.
  - Ask your peers, Ask me, Ask the TA.
- Answer questions. Please help to make lectures interactive – not a one-way flow of information.
- Read around the subject and try things for yourself.
- Plan your time carefully.

What you can expect from me
I will do my best …

- To make lecture notes available on Moodle.
- To give you references to follow up.
- To be present, with the TA, in all labs, available to answer any questions and give feedback.
- To arrange extra support if you've already tried the normal routes (web, peers, TAs).
- To give you prompt feedback on your work.
- To respond to email (ideally, a last resort!)
  - Note: I get far too many emails, and do not check them 24/7, I will reply as soon as I can.

Questions?

Collecting online data

- On with the show…

Sources of (text) data

- Pre-prepared
- APIs
- Data dump
- Web scraping
- + eliciting from people, e.g. MTurk

Sources: pre-prepared

- Often for specific business need, or research purpose.
- Sometimes general purpose.
- Examples:
  - Attached to research paper or project, e.g. http://www.research.lancs.ac.uk/portal/en/datasets/search.html
    - See: http://www.aclweb.org/anthology/W17-1603
  - Kaggle: https://www.kaggle.com/datasets
  - https://corpus.byu.edu
  - Google n-grams: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
  - WaCKy: http://wacky.sslmit.unibo.it/doku.php?id=corpora
- Advantages:
  - Normally easy to process: clean and tidy.
  - Well defined, well designed.
  - Comparable research.
- Disadvantages:
  - Often restricted to purpose collected for.
  - Often cannot fill in missing data needed.
  - Cost?
  - Dated?
  - Representative of what?
  - Relying on integrity of data/owners.

Sources: APIs

- Increasingly offered by big vendors (and small) for interacting with service and data.
- Programmatically interact with data, instead of via webpages.
- Examples:
  - Twitter: https://developer.twitter.com/
  - Facebook: https://developers.facebook.com/
  - Reddit: https://pushshift.io

- Advantages:
  - Relatively easy to access.
  - Can define own queries.
  - Clean and tidy.
  - Extra data available.
  - Terms & conditions.
  - Documentation
- Disadvantages:
  - Sometimes restricted.
  - Rate limited.
  - Cost?
  - Representative?
  - Re-publishing?

Sources: Data dumps

- Large dumps of “everything”.
- Sometimes offered with permission, sometimes not.
- Examples:
  - archive.org
  - Reddit: https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/
  - Darknet archives: https://www.gwern.net/DNM-archives
- Advantages:
  - Masses of data.
  - Normally well structured.
- Disadvantages:
  - Representative of what?
  - What’s missing?
  - Legal?
  - Privacy concerns?
  - Dealing with scale.

Sources: web-scraping

- Using tools to mimic browsing websites, and pulling data.
- Examples:
  - Beautiful soup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
  - Requests: http://docs.python-requests.org/en/latest/user/quickstart/
  - Scrapy: https://doc.scrapy.org/en/latest/intro/overview.html
  - Selenium: http://seleniumhq.github.io/selenium/docs/api/py/
  - Justext: http://corpus.tools/wiki/Justext
- Advantages:
  - Access to data that otherwise would be difficult to collect.
- Disadvantages:
  - Can be messy
  - Reliant on well-formedness of webpages.
  - Technological barriers to some data.
  - Content changes
  - Session-based content.
  - Privacy?
  - Control over data collected.
  - Accuracy.

Ethical considerations

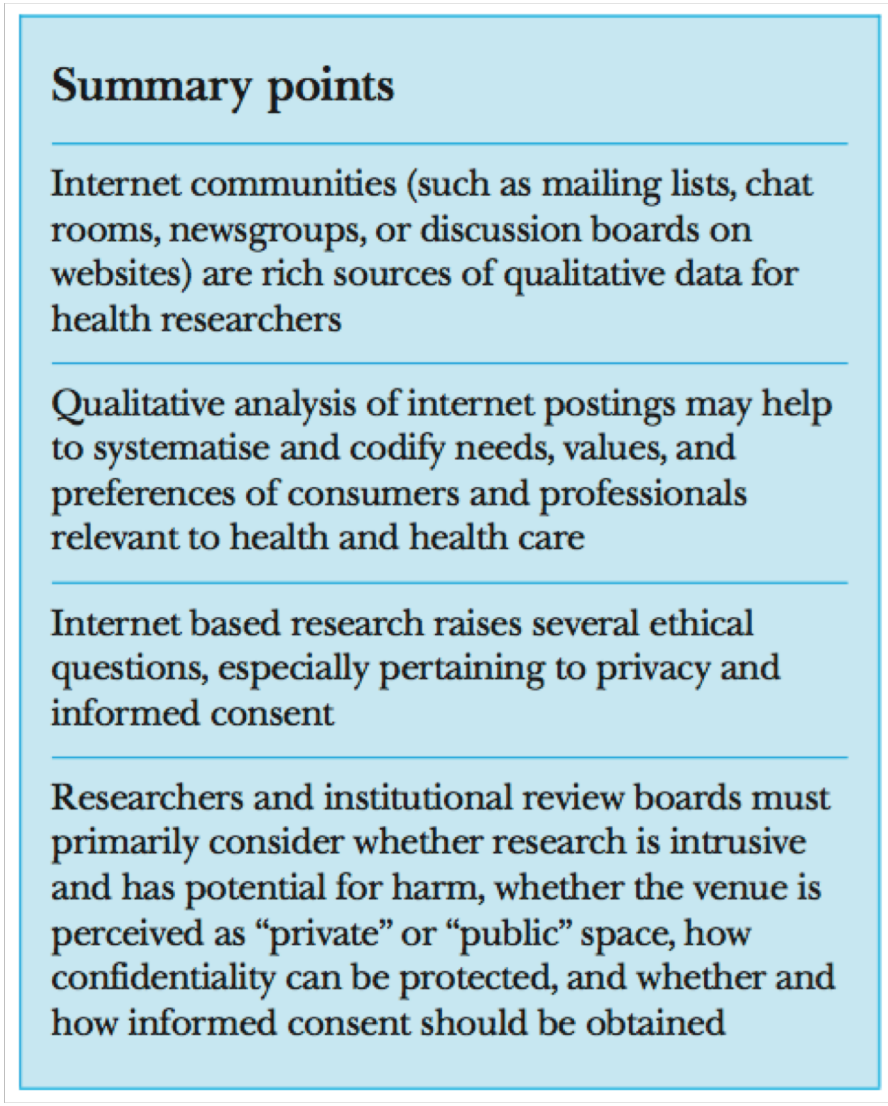
1. If the content of a subjects communication were to become known beyond the confines of the venue being studied would harm likely result?
2. Could analysis, publication, redistribution, or dissemination of content harm the subject in any way?
3. Does the author/participant consider personal network of connections sensitive information?
4. Does author/participant consider the presentation of information or venue to be private or public?
5. Do the terms of service conflict with ethical principles?
6. Is the author/subject a minor?

Association of Internet Researchers: https://aoir.org/ethics/

https://aoir.org/wp-content/uploads/2017/01/aoir_ethics_graphic_2016.pdf

Ethical considerations

- Eysenbach & Till (2001) Ethical issues in qualitative research on internet communities. BMJ.
  - http://dx.doi.org/10.1136/bmj.323.7321.1103
  - Strongly recommend you read this paper.
- 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

- Republishing the data is problematic, this causes problems for reproducibility.
- Ways around this:
  - Host the corpora and restrict concordances to fair use
  - Release URL lists / Tweet IDs
  - But what if data subsequently deleted?
- 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: http://commoncrawl.org
  - 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

Representativeness

- 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/

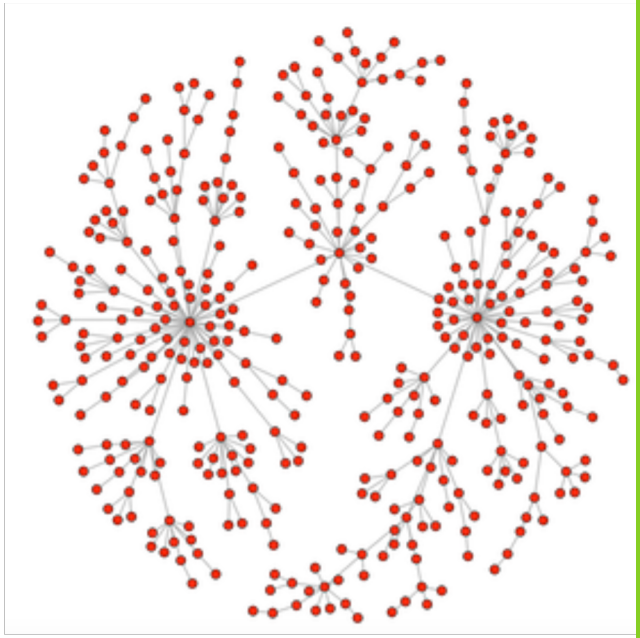
Metadata and corpus structure

- 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

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

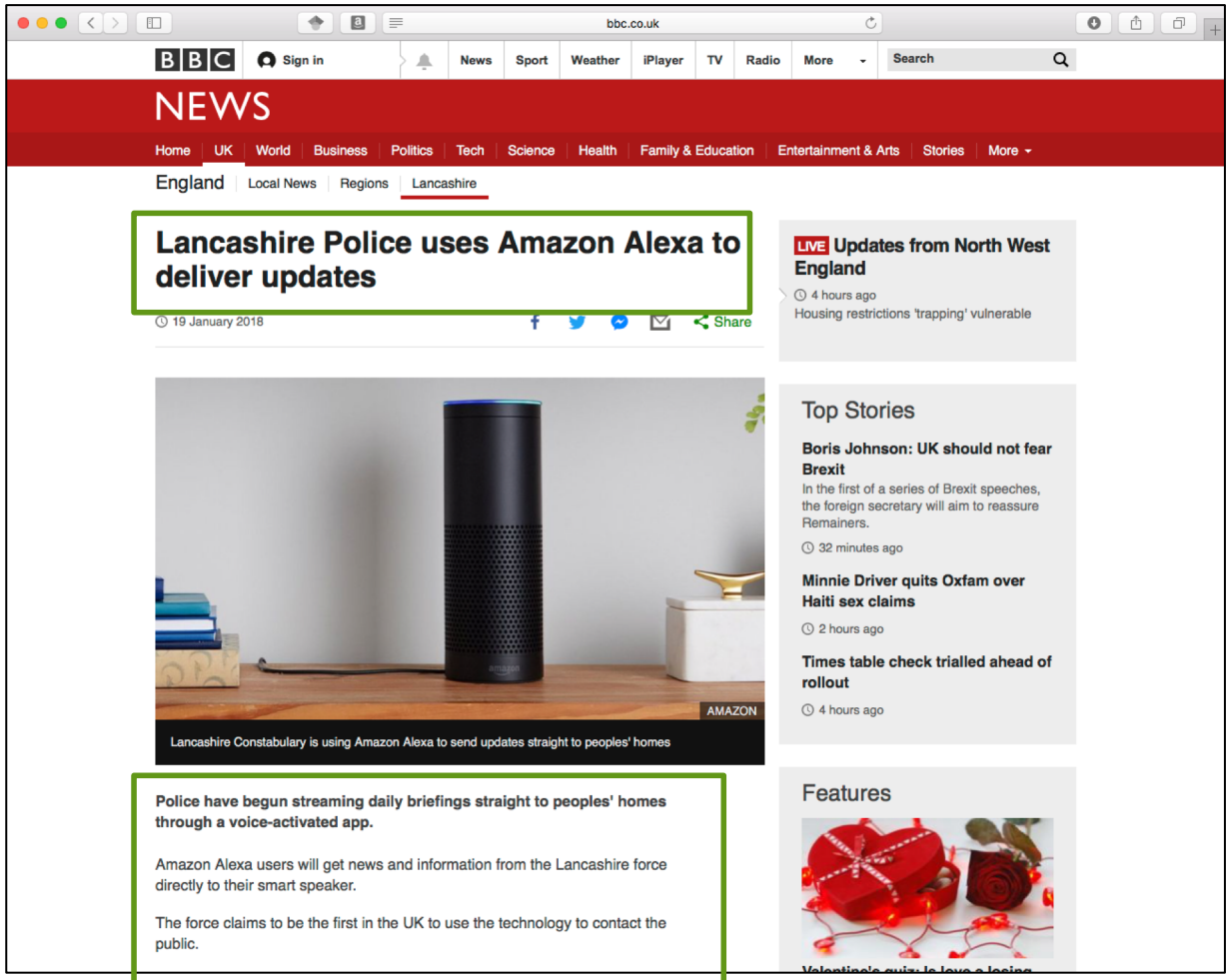
Approach: browsing/spidering and searching

Browsing / Spidering



Searching

Boilerplate



```
624
625
626                 <span class="off-screen">Image copyright</span>
627                 <span class="story-image-copyright">Amazon</span>
628
629             </span>
630
631             <figcaption class="media-caption">
632                 <span class="off-screen">Image caption</span>
633                 <span class="media-caption__text">
634                     Lancashire Constabulary is using Amazon Alexa to send updates straight to peoples' homes
635                 </span>
636             </figcaption>
637
638         </figure><p class="story-body__introduction">Police have begun streaming daily briefings straight to peoples' homes through a voice-activated app.</p><p>It
    •       was developed by PC Rob Flanagan, Lancashire Constabulary's innovations manager, who has worked with developers from Amazon to set it up.</p><div id=
639     <div class="bbccom_advert">
640         <script type="text/javascript">
641             /**/
642             (function() {
643                 if (window.bbcdotcom && bbcdotcom.adverts && bbcdotcom.adverts.slotAsync) {
644                     bbcdotcom.adverts.slotAsync('mpu', [1,2,3]);
645                 }
646             })();
647             /**/
648         </script>
649     </div>
650 </div><p>"As a police force we are always looking at ways to engage with our communities," he said.</p><p>"Alexa works alongside traditional policing methods
651             <span class="image-and-copyright-container">
652
653
654                 <div class="js-delayed-image-load" data-alt="Amazon will have to gain trust before businesses will feel confident inviting it into the boardr
655
656
657                  <span class="off-screen">Image copyright</span>
658                  <span class="story-image-copyright">Amazon</span>
659
660             </span>
661
662             <figcaption class="media-caption">
663                 <span class="off-screen">Image caption</span>
664                 <span class="media-caption__text">
665                     An estimated 11 million households had bought an Amazon Alexa device in the UK by the end of 2017
```

Boilerplate removal / web scraping

- Automate completely
  - justext: http://corpus.tools/wiki/Justext
- Write script for specific website / set of websites
  - Beautiful Soup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
  - Scrapy: https://doc.scrapy.org/en/latest/intro/overview.html
  - Selenium: http://seleniumhq.github.io/selenium/docs/api/py/

Corpus creation: recommendations

- 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?

Reading

- *** Eysenbach & Till (2001) Ethical issues in qualitative research on internet communities. BMJ.**
  - **http://dx.doi.org/10.1136/bmj.323.7321.1103**
- *** Hovy & Spruit (2016): The Social Impact of Natural Language Processing. ACL:**
  - **http://www.aclweb.org/anthology/P16-2096**
- Web Corpus Construction
- Applied text analysis with Python: Chapters 1 & 2.
- NLTK book: Chapters 1 & 2.