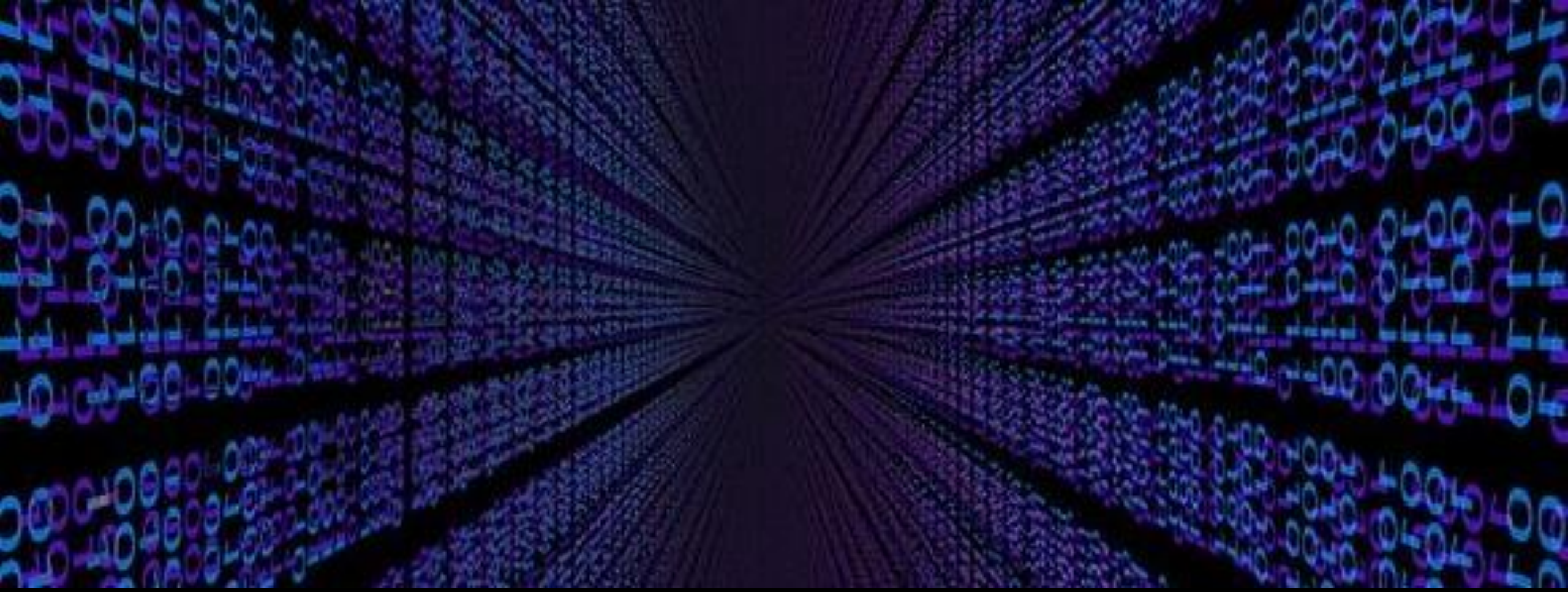


Week 1: Intro to Text Mining

XQuery Working Group

Text Mining at Scale

Fall 2019

What is it?

Using computational tools to analyze large volumes of text

Where impracticable for a human to read it all

Find patterns not noticed by humans

Common text mining goals

- Question answering
- Automatic summarization
- Named-entity recognition
- Sentiment analysis
- Language detection and machine translation
- Optical character recognition

Where do we see it at work?

CHAT BOTS

SPAM FILTERS

SOCIAL MEDIA MODERATION

AUTOCORRECT

VOICE ASSISTANTS

RISK ASSESSMENT

RESUME FILTERING

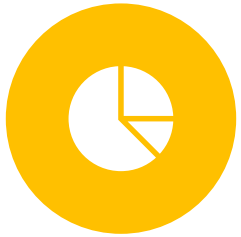
INFORMATION EXTRACTION

TREND ANALYSIS

Steps in the process

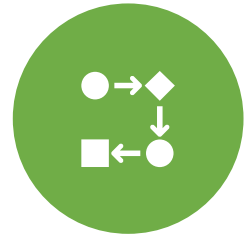
BUILD A CORPUS

PREPARE YOUR CORPUS



ANALYZE YOUR CORPUS

REVIEW RESULTS



REPEAT AS NEEDED

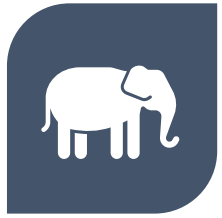
Where might you find text?

WEBPAGES & APIS

EMAILS

SOCIAL MEDIA

NEWSPAPERS



HATHI TRUST

GUTENBERG PROJECT

MEDICAL JOURNALS

COMPANY REPORTS

Challenges to Building a Corpus

- copyright restrictions
- licensing restrictions
- format limitations
- hard-to-navigate systems

Issues become more pronounced at scale

Language is ambiguous

"One morning I shot an elephant in my pajamas."

- Groucho Marx in 'Animal Crackers'

Text as data?

Language is semi-structured or unstructured data

ANTONY

Friends, Romans, countrymen, lend me your ears.
I come to bury Caesar, not to praise him.
The evil that men do lives after them;
The good is oft interrèd with their bones. 85
So let it be with Caesar. The noble Brutus
Hath told you Caesar was ambitious.
If it were so, it was a grievous fault,
And grievously hath Caesar answered it.
Here, under leave of Brutus and the rest 90
(For Brutus is an honorable man;

As TEI structured data

```
▼<sp xml:id="sp-1559" who="#Antony_JC">
  ▼<speaker xml:id="spk-1559">
     <w xml:id="fs-jc-0249800">ANTONY</w>
   </speaker>
  ▼<l xml:id="ftln-1559" n="3.2.82">
     <w xml:id="fs-jc-0249810" n="3.2.82" lemma="friend" ana="#n2">Friends</w>
     <pc xml:id="fs-jc-0249820" n="3.2.82">,</pc>
     <c></c>
     <w xml:id="fs-jc-0249840" n="3.2.82" lemma="Roman" ana="#n2-nn_j">Romans</w>
     <pc xml:id="fs-jc-0249850" n="3.2.82">,</pc>
     <c></c>
     <w xml:id="fs-jc-0249870" n="3.2.82" lemma="countryman" ana="#n2">countrymen</w>
     <pc xml:id="fs-jc-0249880" n="3.2.82">,</pc>
     <c></c>
     <w xml:id="fs-jc-0249900" n="3.2.82" lemma="lend" ana="#vvb">lend</w>
     <c></c>
     <w xml:id="fs-jc-0249920" n="3.2.82" lemma="i" ana="#pno">me</w>
     <c></c>
     <w xml:id="fs-jc-0249940" n="3.2.82" lemma="your" ana="#po">your</w>
     <c></c>
     <w xml:id="fs-jc-0249960" n="3.2.82" lemma="ear" ana="#n2">ears</w>
     <pc xml:id="fs-jc-0249970" n="3.2.82">.</pc>
   </l>
```

Vocabulary Digression

- Token / tokenize
- Part of speech (POS) tagging
- Stemming and lemmatization
- Named entity recognition (NER)
- N-grams and collocation
- Concordance (key word in context)
- Stop words
- Topic modelling

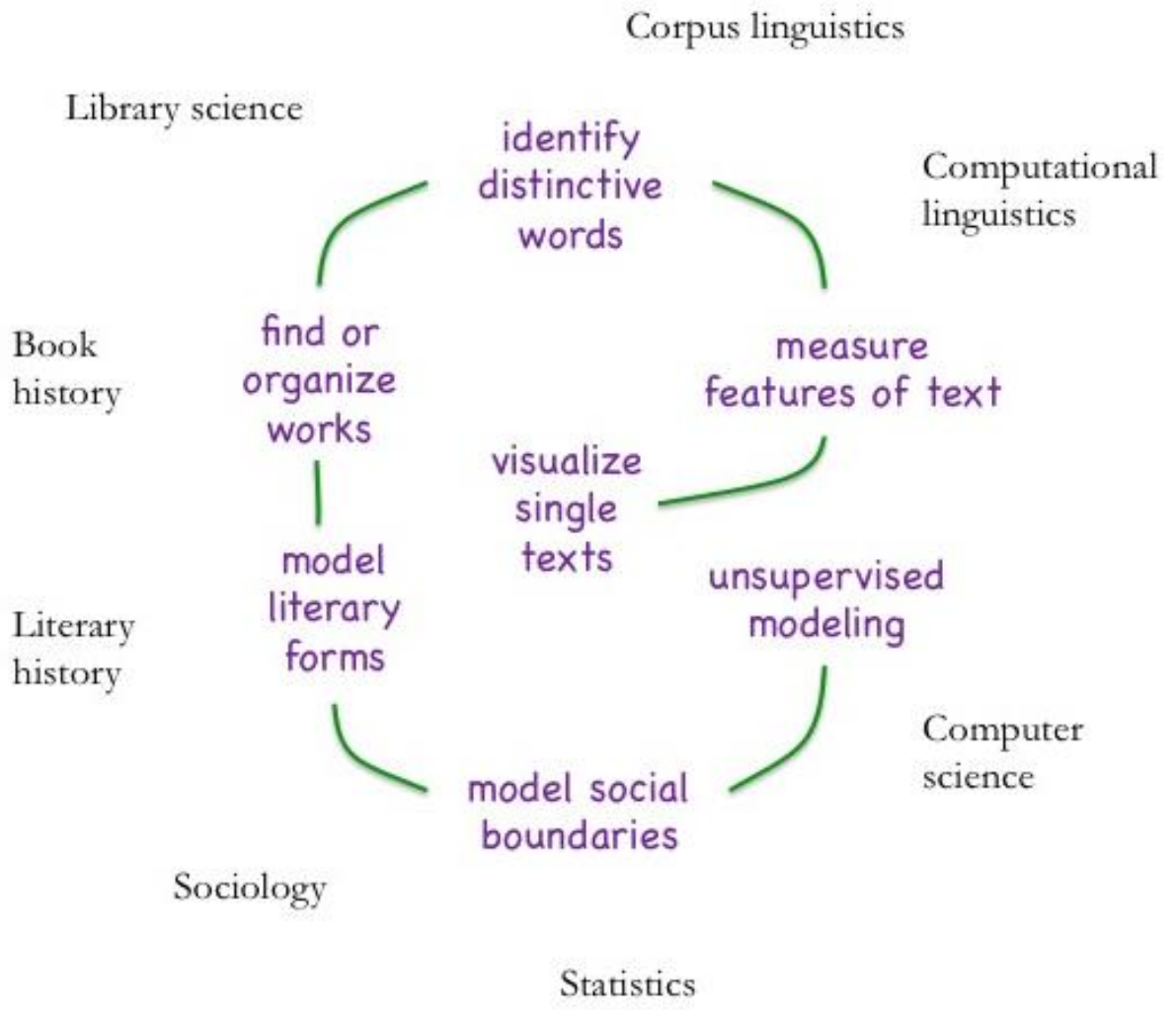
More on preparing your corpus

- Correct OCR errors
- Remove title, header information
- Remove html
- Split or combine files
- Remove certain words, punctuation
- Lowercase text
- Tokenize the words

Your text preparation can impact your results.

Ted Underwood: Seven Ways to Understand Text

https://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/



Examples

- https://books.google.com/ngrams
- https://hansard-interjections.herokuapp.com/tweets/
- https://dsl.richmond.edu/dispatch/pages/home
- https://voyant-tools.org/
- http://themacroscope.org/interactive/dcbtopicnet/

Deconstructing the text

Just bought a book from IKEA