Applied Quantitative Text Analysis @ Digital Methods Workshop

Joel Nothman
joel.nothman@sydney.edu.au

Sydney Informatics Hub
University of Sydney

8th June, 2018

Language can tell us a lot about people

- how do people feel about some product / event / situation?
- why do people share some media?
- how do people persuade or negotiate?
- how do people express their identity as a community member?
- how do people, communities and periods differ from one another?

- studying these questions: content analysis, discourse analysis, corpus linguistics . . .

- treating language as data might not be straightforward

Language is . . .

- structured: Jack kicked Jill $\neq$ Jill kicked Jack
- free to represent arbitrary meaning
- an efficient encoding of a communicative purpose given context, shared knowledge, and social convention
- efficiently decoded by humans despite rife ambiguity
- highly variable over time, place and genre/purpose

Ambiguity everywhere

- of lexical semantics: I walked to the bank
- of structure: I saw the girl on the hill with the telescope
- of named entity reference: Washington (George, Denzel, city, state, US government, university, sports teams, . . .)
- of semantic roles: Jill's care – is Jill giver or recipient?
- . . .

Language is statistically sparse

- infinitely many words/phrases/sentences; almost all are rare
- many ways to express the same thing
- we ideally want equivalent analyses of
  - Google bought YouTube for $1bn
  - The tech giant's acquisition of YouTube for $1bn
  - YouTube, which was acquired by Google for $1bn
  - Google didn't hesitate to snatch it up for a billion bucks
- non-literal language: Canberra announced; kicked the bucket

- How do ambiguity and variety of expression affect working with langugage as data?

Objectives of text processing in research

- description
  identify patterns in text
- retrieval
  find relevant content in a collection
- prediction
  automate labelling of textual phenomena

Description

- may involve comparing frequencies of some textual features
  - Does one president use *I*, *me* much more than another?
  - What distinguishes men's and women's tweets?
- may involve finding clusters of similar content
  - What political parties, internationally, have similar platforms?
- may involve visualising patterns in a text collection

Textual features

- usually our research question is about *content*/*meaning* or *style*
- but we need quantifiable properties of the text

- naive approaches to language tend to be most interpretable
  - comparing word (or n-gram) frequency
  - sentiment polarity, LIWC
  - syllables per word, words per sentence
- but always validate your quantitative conclusions
  - manual inspection: does the result mean what you think it means?
  - statistical validation: will the result hold on new data?

Findings should be validated by manual review

A concordance lists when a term appears, with surrounding context

Left	Term	Right
Netherfield. Such amiable qualities …	speak	for themselves. What a contrast
eyes. I never heard you	speak	ill of a human being
censuring anyone; but I always	speak	what I think." "I know
herself again. She longed to	speak	, but could think of nothing
explain the matter; Darcy must	speak	for himself." "You expect me
as she allowed him to	speak	. "You either choose this method
The person of whom I	speak	is a gentleman, and a
ring the bell—I must	speak	to Hill this moment." "It
she could not bear to	speak	of the day before was
inquiry. Mr. Wickham began to	speak	on more general topics, Meryton
It gives me pain to	speak	ill of a Darcy. But
though she did not often	speak	unnecessarily to Mr. Collins, she

From Voyant Tools

Retrieval

- avoid manual labour of finding texts with relevant phenomena
  - How did an idea spread through a social network?
- may need be able to compute how similar two texts are

Prediction

- avoid manual labour of labelling some phenomenon
  - Can we work out if someone is happy from their words?
  - How typical is a speech of a particular political affiliation?
  - Replicate human judgements of writing quality
- label a sample and try build an accurate system to label more
- model does not need to be interpretable, as long as it predicts well
- automated labelling may assist in description e.g. reducing the amount of manual coding in content analysis

Collecting text

- We don't tend to start with tidy paragraphs
  - Web forums with structure, quotation, signatures
  - Online news with boilerplate
  - Twitter
  - PDFs with headers and footers
  - Paper / microfiche
- Cleaning is inevitable
  - Web scraping
  - Boilerplate removal
  - Optical character recognition
  - Spam and duplicate removal
  - Spelling correction or normalisation
- Then need to tokenise text into sentences and words

Collections of cleaned text can be reused

- A fixed collection of texts is a corpus (plural corpora)
- To find unusually frequent words in someone's speech, you need to compare against an appropriate reference corpus

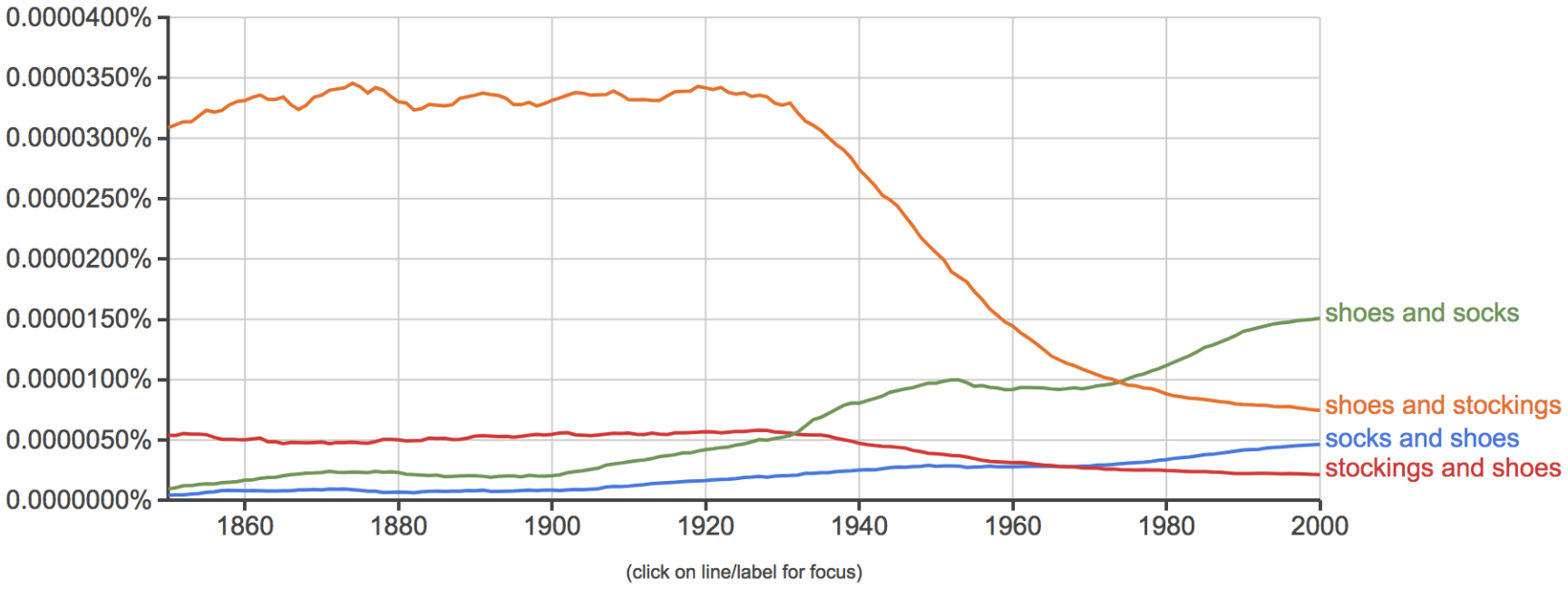
- General-purpose corpora select for:
  - dialect (e.g. British National Corpus)
  - genre (fiction, government, humour, news in Brown Corpus)
  - medium (newswire, blogs, web forums, conversational speech, broadcast speech)

Large scale corpora allow us to expore language variation

Google Books Ngram Viewer

Graph these comma-separated phrases: socks and shoes,stockings and shoes,shoes and socks,shoes and s case-insensitive

between 1850 and 2000 from the corpus English with smoothing of 10. Search lots of books



Manual annotation, e.g. for text categorisation

- Sometimes you want to split your data according to some metadata
  - Can you classify literature into the year it was published?
  - Can you classify tweets into the stated gender of the tweeter?
  - Can you classify tweets into the stated location of the tweeter?
- Sometimes you need to label a sample of documents with categories
  - ... and use these categories to split a corpus and compare
  - ... and use these categories to evaluate a classification rule
  - ... and use these categories to train logistic regression
- A gold standard is developed by manually labelling a sample of texts
  - Annotate each text multiple times, then measure inter-coder agreement
  - Pre-set a goal, and improve the category definitions if unmet

Reusing existing tools

- NLP focuses on accurately identifying and decoding aspects of language structure
- Fairly mature technologies (at least in English):
  - identify broad categories of sentiment expressed in a document
  - identify topics (words that tend to appear together) in a corpus
  - strip inflectional morphology from an word
  - label each word with its part of speech
  - identify syntactic dependencies between each word in a sentence
  - identify names of people, organisations, locations
  - disambiguate which famous entity is being referred to
  - transcribe speech to text
- but:
  - all these tools will make errors; use judiciously
  - may not work well on your language/medium/genre/dialect

The Sydney Informatics Hub

- Training and assistance in computational techniques for research
- `informatics.sydney.edu.au`
- a Core Research Facility

- Learn to code for research
- Attend a Hacky Hour for help with your research code
- Request assistance in collecting, analysing or modelling data
  - Or: contact Chao Sun, the FASS Data Scientist

- Are there specific techniques we should provide training in?