Opportunities and Challenges of Textual Big Data for the Humanities

Dr. Adam Wyner, Department of Computing

Prof. Barbara Fennell, Department of Linguistics

July 1, 2013

THiNK Network – Knowledge Exchange in the Humanities

RSA House, London, UK

Overview

- Introductions.
- Big data – resources.
- Text tools.
- Examples.
- Collaborative challenge.
- Knowledge exchange.

Big Data

Technological, resource, and economic changes present opportunities and challenges to the humanities. We live in a Big Data world of increasing bodies of textual data that are available on the Internet from libraries, government organisations, social websites, and blogs.

The Story is Out There

Big Data analysis in the news (since 2008):

- Obama's Open Government Law
- PRISM.
- http://www.guardian.co.uk/data
- Mayer-Schonberger and Cukier (2013). *Big Data.*
- Foreign Affairs. "The Rise of Big Data".

What are the consequences of leaving the tools in the hands of large organisations with social/commercial interests?

Big Data

- Lots being done in:
  - bioinformatics (searching articles to 'link up' knowledge).
  - legal patent analysis (newness).
  - commerical text mining (corporate blogs, Amazon, Facebook, Thomson-Reuters NER, etc).
  - security services.
  - medical records.

Samples – Open Source Data

- Open government data (UK, US, EU).
- Library collections that are out of copyright.
- Corpora, e.g. Public.Resource.Org, Legal Information Institutes, others....
- Blogs, websites, open websites, open journals, email communications....
- English and other languages.
- Value and benefits of text mining - JISC

Current Practice and Future Direction

- Current Big Data practice of working with the meta-data, explicit network information (e.g. linking friends to friends), or databases.
- Contrast with *information extraction* from *text*.
- Sentiment (positive and negative views) analysis is being done, but *coarse-grained*.
- How about *fine-grained textual content analysis*?

Some Research Questions

- From 1641 Depositions:
  - What is a deposition (commonalities across text)?
  - How is hearsay defined (how does it appear)?
  - How did the depositions change over time?
  - What are the interrelationships between depositions in terms of the content?
  - How is evidence manipulated by third parties (what are the textual indicators across text)?

Some Research Questions

- From Statutes and Regulations:
  - What are networks of laws?
  - How did the statutes and regulations change over time?
  - What are the relationships between laws, business rules, and compliance?
  - Cross jurisdictional variation in the realisation of statutes and regulations (disaster relief roles and actions).

Tools

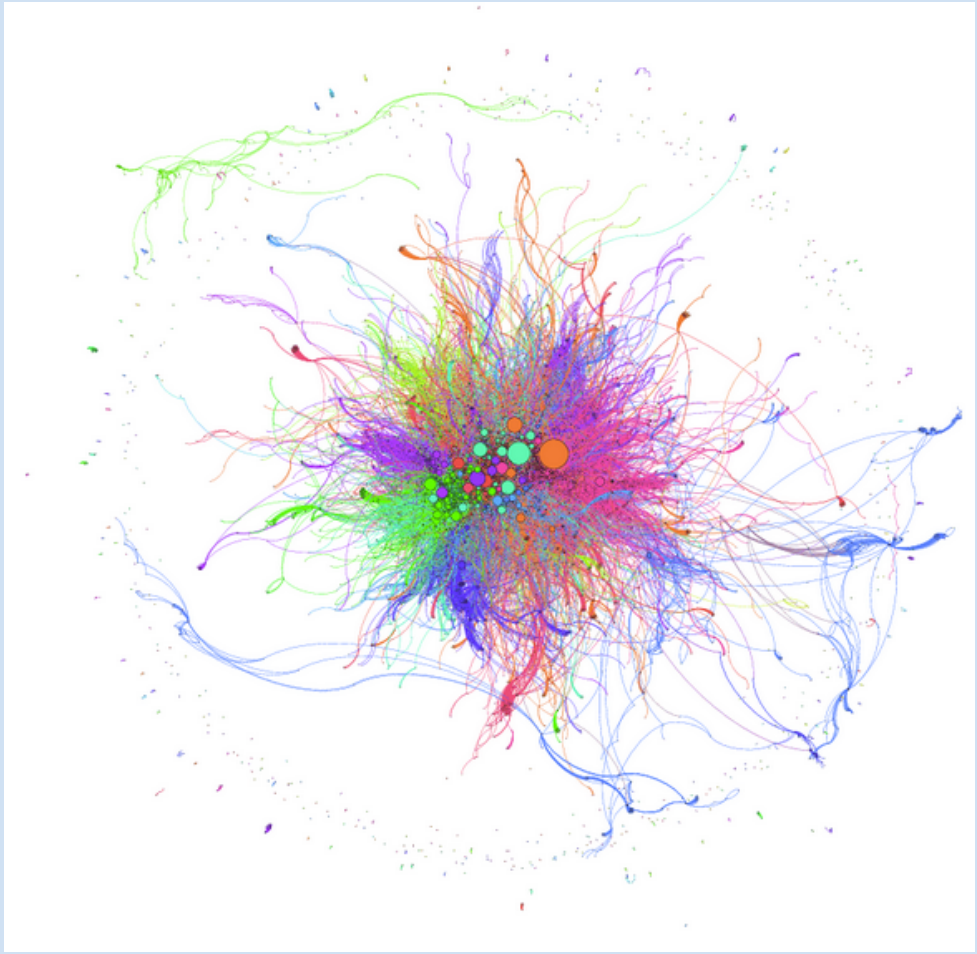
Not only do we have new resources, but we have new and powerful tools to search, compare, accumulate, share, and represent information about these data.

Outputs

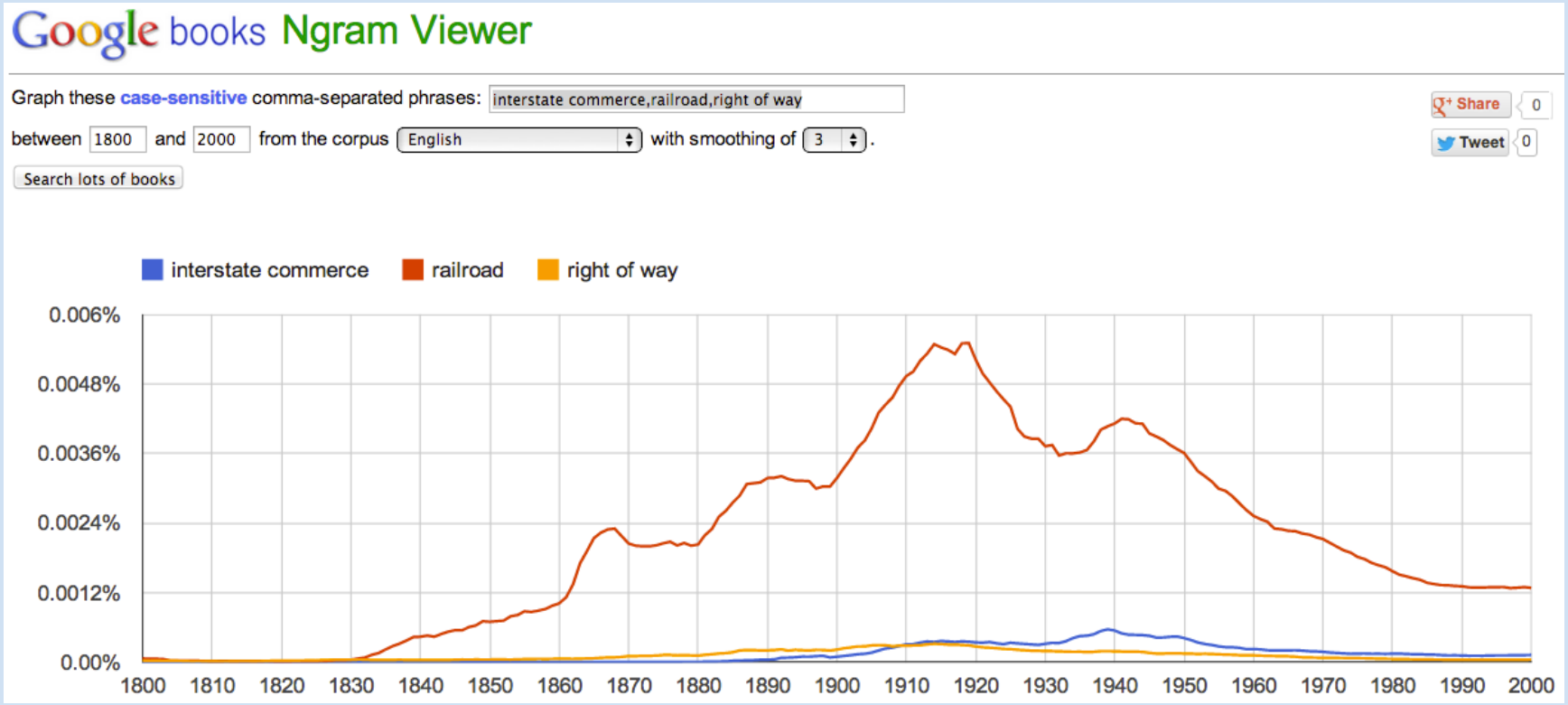
- Network graphs showing relationships (references, links) between web-based material.
- Google's Ngram Viewer and Legal Language Explorer.

Graphs of Dutch Legal Document References

Hoekstra, 2013.
"A Network Analysis of Dutch Regulations

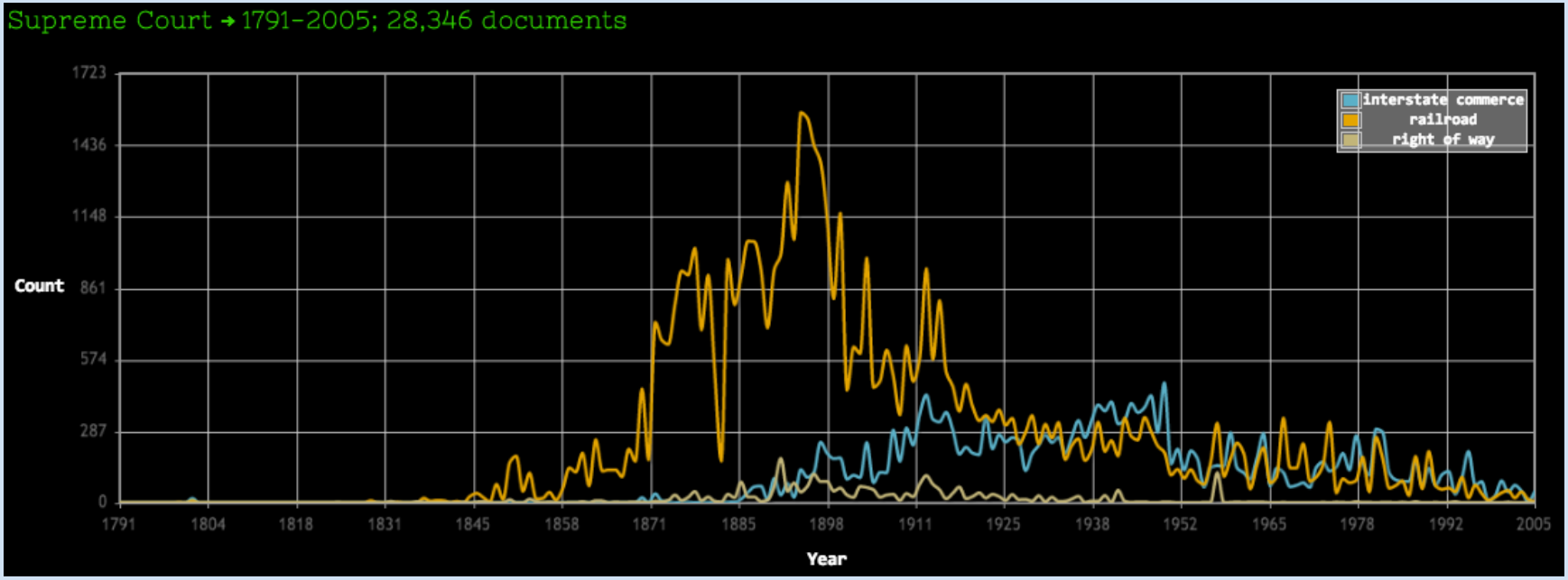


Google Ngram Viewer



interstate commerce, railroad, right of way

Legal Language Explorer



interstate commerce, railroad, right of way

Going Deeper – Where Knowledge Matters

- Going for structured semantic information contained in the texts.

Looking for?

- Named entity recognition (who, what, when, where).
- Coreference (associating entities across sentences and text).
- Fine-grained sentiment analysis (positive or negative dispositions on particulars).
- Word patterns (terminology that is descriptive).
- Semantically contentful information with annotations.
- Relationships, values,....

Tools

- http://en.wikipedia.org/wiki/Text_mining
- General Architecture for Text Engineering

Sample Applications

- Law (Legal case analysis, regulation, argumentation).
- Psychological analysis (Phil Gooch), associating patient narratives with psychopathologies.
- Anti-depressants, press, and influence (Nooreen Akhtar).

GATE Example on Argumentation

- Objective: identify and extract arguments from a corpus
- Web-based discussion forums on Amazon about cameras (papers by Wyner et al.).
- Could do this on the BBC's *Have Your Say* or similar.

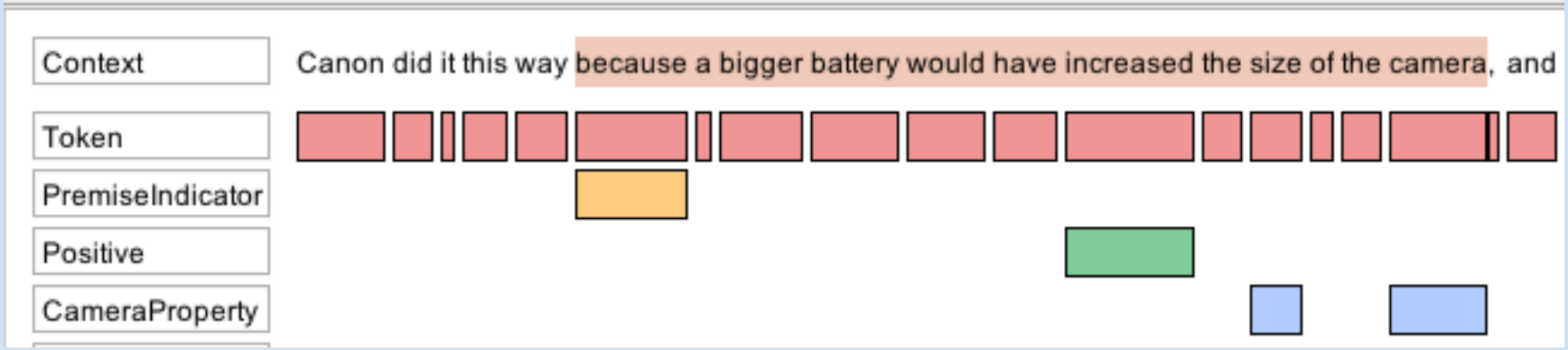
Terminological Annotations

- Rhetorical structure information (*premise*, *conclusion*, etc).
- Domain terminology (*camera features*, etc).
- Contrast (*poorly*, *not*, etc.).

I adore this camera. Why? Because I get a higher percentage of good shots out of it than from any other camera I have, and that includes a few DSLRs, and the video is ridiculously good. Is the image quality as good as a DSLR or even a micro 4/3 system camera? No, but it's not far off, and a shot that's perfectly exposed and focused and free of camera shake on a smaller sensor like this beats not getting a picture at all by a long way. The SX220HS is my constant companion and lives in a pouch on my belt.

Query for patterns

{PremiseIndicator}({Token})*10{Positive}({Token})*10{CameraProperty}



Structure and extract an argument for buying the camera

Premises:

The pictures are perfectly exposed.

The pictures are well-focused.

No camera shake.

Good video quality.

Each of these properties promotes image quality.

Conclusion:

(You, the reader,) should buy the CanonSX220.

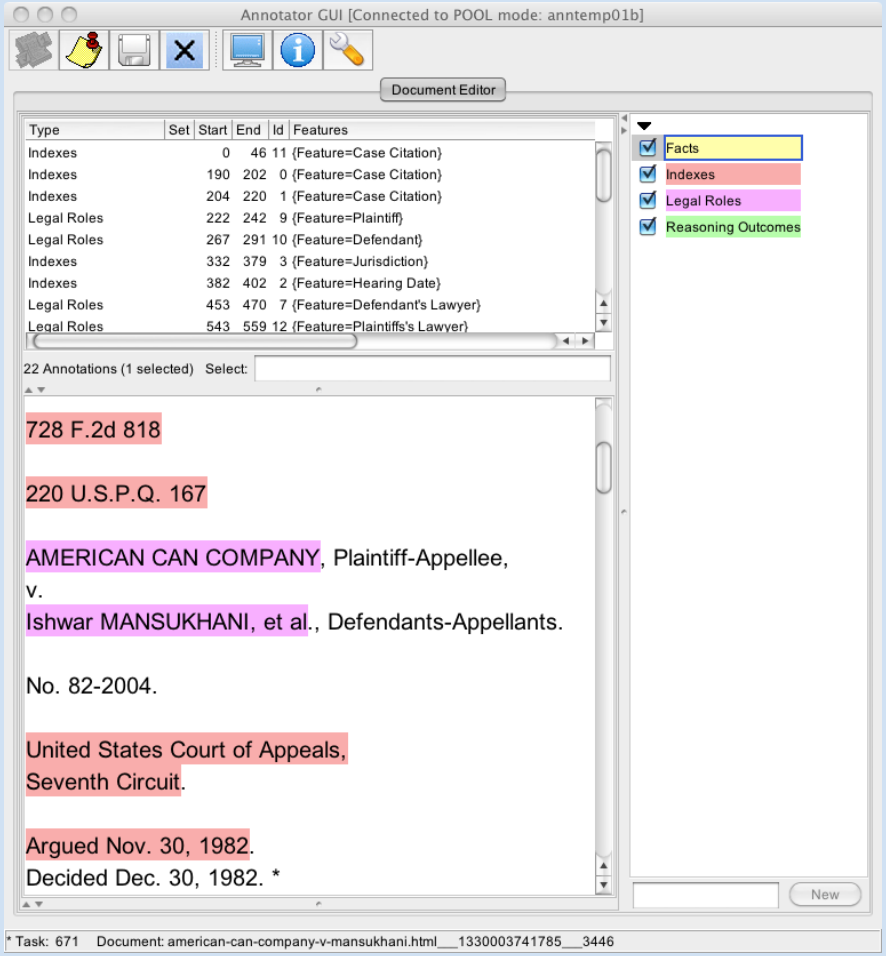
Teamware

- Tool for distributed, collaborative semantic annotation.
- Makes a corpus searchable by semantic concepts.
- Collective introspection - making subjective evaluations *objective*, *comparable*, *measurable*, *generalisable*, and *retestable*.

Teamware Example

Crowdsourced Legal Case Annotation, Wyner
Like manual annotation, but online and automatically compared.

Can create online, collaborative annotation tasks for lots of text and concepts – argumentation, story roles, newspaper elements, political positions,....



Collaborative Challenge

The challenge is to develop not only the tools, which we largely have to hand, but more importantly the human resources to work with them to carry out distributed, collaborative projects.

Collaborative Challenge

- The *interface* – computer people bring x, humanities people bring y, combined they produce z.
- *Specialist knowledge* is built into something that is *machine processable*, e.g. lists and rules in GATE.
- Collaboratively building gold standards; refining the lists and rules.

Knowledge Exchange

- Putting tech in hands of humanities scholars.
- Team collaboration in development – humanities scholars provide their subject specific knowledge; tech provide tools, support, development, frameworks, analysis.
- Creating, growing, and maintaining a *common language*.
- Lawyers, Linguists, Arts, Social/Political Scientists, Policy-makers....
- Specific tools, how-tos, statistics, auxiliary coding....

Other Tools for Various Users

- Other tools to explore – 'Mining the Social Web', data mining, visual analytics....
- Variety of users with different skill levels.

Thanks for your attention!

- Questions?
- Contacts:

  - Adam Wyner azwyner@abdn.ac.uk
  - Barbara Fennell b.a.fennell@abdn.ac.uk