Documents and text: Introduction to natural language processing

Russell J. Funk

Carlson School of Management
University of Minnesota

Roadmap

- Motivation
- Background
- Computation
- Organizational research
- Frontiers

Motivation

Why documents and text?

Perhaps more than ever, social life is captured in documents and text...

- social media posts (Bail, 2016; Bakshy et al., 2015)
- newspapers (DiMaggio et al., 2013; Baker et al., 2016)
- political speeches (Rule et al., 2015)
- court cases (Danescu-Niculescu-Mizl et al., 2012)

That includes organizations, where networks matter for...

- patented inventions (Kaplan and Vakili, 2015; Kelly et al., 2018)
- product descriptions (Hoberg and Phillips, 2016)
- email messages (Aral and Van Alstyne, 2011; Srivastava and Banaji, 2011)
- press releases (Kennedy, 2008)

To study documents and text, we'll use natural language processing...

What is natural language processing?

- Natural language processing is a set of tools and techniques for analyzing unstructured, natural language data (e.g., speech, text).
- In contrast to formal or constructed languages (e.g., programming languages), natural languages are more complex and less controlled.
- Natural language processing is particularly challenging due to the inherent ambiguity of language.

What do these sentences mean?

"Time flies like an arrow."

"The horse raced past the barn."

What about these?

"Time flies like an arrow. Fruit flies like a banana."

"The horse raced past the barn fell."

Background

Where was natural language processing developed?

- Historically, development of natural language processing has been by two communities.
- These different communities sometimes have divergent goals.

Linguistics. . .

- Developed much of the conceptual machinery and resources underpinning modern NLP (e.g., Penn Treebank, Brown Corpus, WordNet).
- Primary interest is in the scientific understanding of language.

Computer science (and statistics). . .

- Developed much of the computational machinery (i.e., algorithms) underpinning modern NLP (e.g., topic models).
- Primary interest is in the scientific study of algorithms.

Computation

What's the connection to computational social science?

At a surface level...

- Many applications in NLP are computationally intensive (e.g., due to complex operations on matrices).
- Text is basically the canonical example of high dimensional data (Gentzkow et al., 2019).
- In addition, the scale of many NLP data sets makes even seemingly large social science data sets (e.g., patents) look trivial.
- But many NLP techniques depend on massive amounts of data to perform well (c.f., "The unreasonable effectiveness of data").
- Consequently, research using NLP tends to be computational in nature.

At a deeper level...

- As we discussed with networks, NLP tends to be very data driven.
- Particularly as currently practiced in social science, NLP is best viewed as a "method in search of a theory."
- Many popular techniques (e.g., topic models, word embeddings, clustering) are inductive in nature, focused on finding latent structure.

Organizational research

Natural language processing and organizational research

- Overall, natural language processing has seen (comparatively) limited application in organizational research.
- That's starting to change, but we're still behind, even relative to other social sciences.
- Here are some examples of application areas.

Innovation

- novel patents (Kaplan and Vakili, 2015; Kelly et al., 2018)
- influential papers (Gerow et al., 2018)

Sentiment

- social media and the stock market (Bollen et al., 2011)
- scientific citations (Catalini et al., 2015)

Culture

- disciplinary boundaries (Vilhena et al., 2014)
- socialization (Srivastava et al., 2018)
- authenticity (Buhr et al., 2020)
- mood (Golder and Macy, 2011)

Frontiers

Where is natural language processing going next?

My $0.02

Deep learning

- So far, most applications of deep learning have been in industry.
- We'll likely start to see more applications in science, especially for prediction tasks (e.g., matching, instruments).

Theory building

- We're seeing increasing adoption of natural language processing techniques among traditionally qualitative, inductive researchers.
- There is growing interest in using natural language (e.g., topic models, word embeddings) and related techniques for theory building.

Opening the black box

- Many advanced NLP models perform well, but what happens under the hood is a black box (to mix two metaphors).
- That may limit their value from a scientific perspective.

Where is natural language processing going next?

My $0.02

Structured data extraction

- NLP techniques are becoming much better at extracting (structuring) data from raw, unstructured text.
- These techniques will less likely picture in the analysis phase, but more likely in the data collection phase of research.

Unpacking meaning

- Recent advances in NLP (e.g., word embeddings, topic models, sentiment analysis) are increasing our ability to extract meaning from text.
- These techniques create new possibilities for analysis, including for example, how meaning changes over time, across organizations, and so forth.

Appendix