Text Analysis

1. Interpreting large volumes of text (distant reading)
2. Highlighting aspects of texts that are not obvious to humans
3. Determining authorship
4. Finding relevant texts
5. Cleaning up scanned texts
Human input

Machine learning:
What topic definitions best fit this corpus?
Human input

Machine learning:
What topic definitions best fit this corpus?

Inference:
Which of the topics best fit these texts?
Human input

Machine learning:
What topic definitions best fit this corpus?

Inference:
Which of the topics best fit these texts?

Which texts are similar to each other?

How does the occurrence of topics change over time?

Which topics tend to occur together?

Which texts are likely to have come from a given corpus?

In what sense is this word being used?

…
What data sets can I use?

- The input is a directory of multiple .txt files
  - These can be books, articles, chapters, pages, etc.
- Some places to find files to use:
  - [http://gutenberg.org](http://gutenberg.org)
  - [https://umich.app.box.com/s/nfdp6hz228qtbl2hwhhb](https://umich.app.box.com/s/nfdp6hz228qtbl2hwhhb)
  - [https://github.com/iulibdcs/tei_text](https://github.com/iulibdcs/tei_text)
Running MALLET

- Download (mallet.cs.umass.edu)
- Decompress zip file (might take a few minutes on lab computers)
- Set up environment variable
  - Control Panel > System > Advanced System Settings
  - MALLET_HOME = C:\Users\<username>\Downloads\mallet-2.0.8RC2
- Run “cmd”
- Get into the right directory
  - cd Downloads\mallet-2.0.8RC2\sample-data
- Importing the data
  - ..\bin\mallet import-dir --help
  - ..\bin\mallet import-dir --input web --output web.mallet --keep-sequence --remove-stopwords
- Topic modeling
  - ..\bin\mallet train-topics --input web.mallet
  - ..\bin\mallet train-topics --input web.mallet --num-topics 25
  - ..\bin\mallet train-topics --input web.mallet --num-topics 25 --output-topic-keys topic-keys.txt --output-doc-topics doc-topics.txt
Learning further

- Look at the MALT options—try different tokenization and stopword settings
- Pre-processing texts
  - Some texts must be split up by chapters
  - Try stripping out headers, etc.
  - Python is useful for this
- Topic model visualization software
- Bringing topic models together with other ways of looking at a text
  - Publication dates and other metadata
  - Looking at how the model has handled specific texts
  - Comparing the topic model to other forms of representation (indexes, etc.)
- Look at what people have written about interpreting topic models
  - Ben Schmidt—looking deep into models
  - Lisa Rhody—thinking about how these programs work differently with different types of text
  - Binder & Jennings—think of topic models as a historically contingent medial form
- Classification and other forms of text analysis