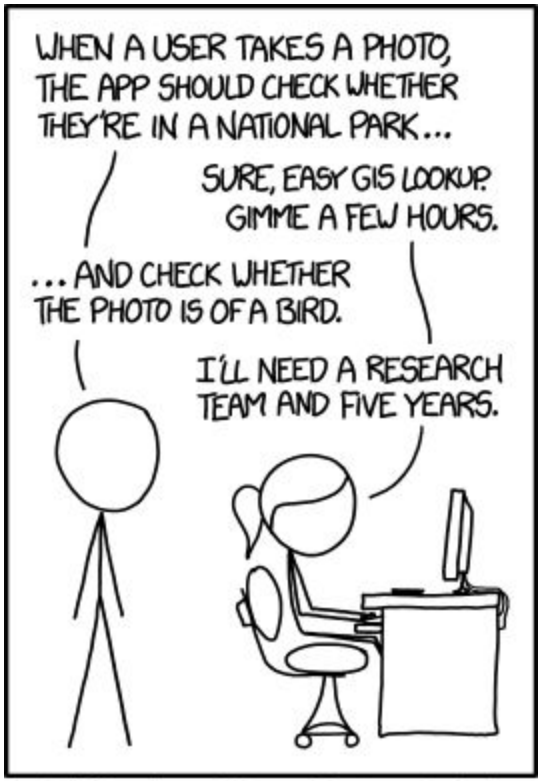
Integrating Technological Tools into Textual Analysis

Anupam Basu

Interdisciplinary Project in the Humanities



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Sources for Text Corpora:

- JSTOR data <http://about.jstor.org/service/data-for-research >
- EEBO-TCP < http://earlyprint.wustl.edu/ >
- ECCO-TCP < http://www.textcreationpartnership.org/tcp-ecco/ >
- Project Gutenberg < http://www.gutenberg.org/ >
- Hathi Trust < https://www.hathitrust.org/ >
- Social media / web datasets
    a. Downloadable as corpus: e.g. Wikipedia
    b. Download via API: e.g. Twitter

What kind of corpus do we have?

- Scale of the corpus
  - Can limit choice of tools and models that are useful
- Is it plaintext or marked up (XML/HTML etc.)? *
  - Should we simply strip away markup?
  - Can we use any information available on structure?
- Does it need preprocessing or regularization? *
- What metadata is available?
  - Is it in a regularized / parseable format? *

Online Tools for Text Analysis:

- Voyant < http://voyant-tools.org/ >
- TAPoR < http://tapor-test.artsrn.ualberta.ca/home >
- SEASR < http://www.seasr.org/ >
- WordHoard < http://wordhoard.northwestern.edu/ >
- Google Ngram Browser < https://books.google.com/ngrams >
- Mark Davies' Corpora at BYU < http://corpus.byu.edu/ >
- EarlyPrint < http://earlyprint.wustl.edu/ >

Advanced Toolkits:

- Desktop / Command line tools
  - Gephi
  - Mallet
  - Weka
  - Mahout
- Programming Libraries
  - NLTK
  - Gensim
  - Scikit-Learn
  - tm
  - Stanford NLP
  - OpenNLP

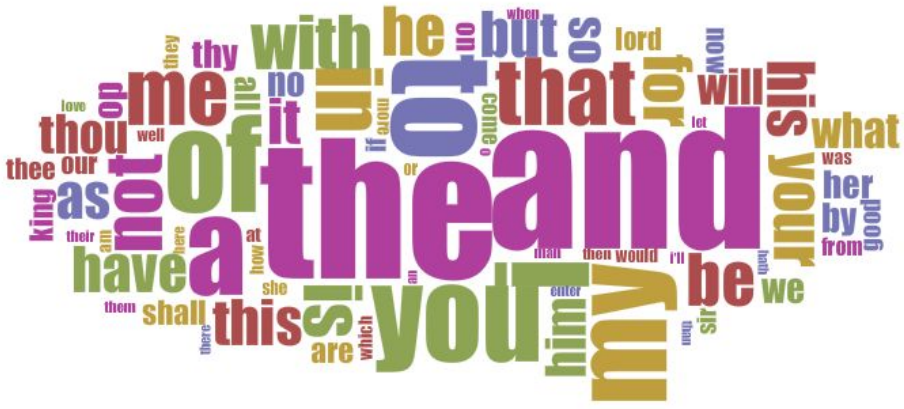
What can we do with Texts? *

- Find frequent words
- Keywords in context
- Identify distinctive words
- Contrast vocabularies
- Ngram usage over time

- Entity extraction
  - Names *
  - Places (geotagging) *
  - Network visualization of relations between entities *
- Stylometry / Authorship attribution
- Sentiment analysis
- Group texts – Clustering
- Categorize texts – Classification

Voyant Tools: Reveal Your Texts

Cirrus



Summary

- There are 37 documents in this corpus with a total of **890,366 words** and **28,750 unique words**.
- **Longest documents** (by words): 1599 Hamlet (32,212), 1593 King Richard III (31,663). Shortest documents: 1591 Comedy of Errors (16,251), 1594 Midsummer Night's...; (17,195). All...
- Highest **vocabulary density** ( ): 1611 Tempest (187.9), 1605 Macbeth (186.7). Lowest density: 1593 King Richard III (132.1), 1600 Much Ado about Nothi...; (133.3). All...
- Most **frequent words** in the corpus: the (28,323), and (26,090), i (20,468), to (19,220), of (17,626). More...
- Words with **notable peaks in frequency** across the corpus: is (9,122), most (1,170), then (2,170), i (20,468), speak (1,176). More...
- **Distinctive words** (compared to the rest of the corpus)
  1. 1590 Love's Labour's Lost: biron (194), the (860), princess (130), ferdinand (127), armado (121). More...
  2. 1591 Comedy of Errors: of (619), syracuse (232), dromio (221), antipholus (219), ephesus (169). More...
  3. 1591 Two Gentlemen of...;: proteus (218), valentine (216), julia (144), i (558), speed (128). More...
  4. 1594 Merchant of Venice: the (859), portia (148), i (656), bassanio (118), shylock (106). More...
  5. 1594 Midsummer Night's...;: lysander (100), demetrius (98), hermia (96), and (576), love (102). More...

  Next 5 of 32 remaining

Corpus Reader

1) 1590 Love's Labour's Lost.txt

LOVE'S LABOUR'S LOST

DRAMATIS PERSONAE

FERDINAND king of Navarre.

BIRON |

|

LONGAVILLE | lords attending on the King.

|

DUMAIN |

BOYET |

| lords attending on the Princess of France.

MERCADE |

DON

ADRIANO DE ARMADO a fantastical Spaniard.

SIR NATHANIEL a curate.

HOLOFERNES a schoolmaster.

DULL a constable.

COSTARD a clown.

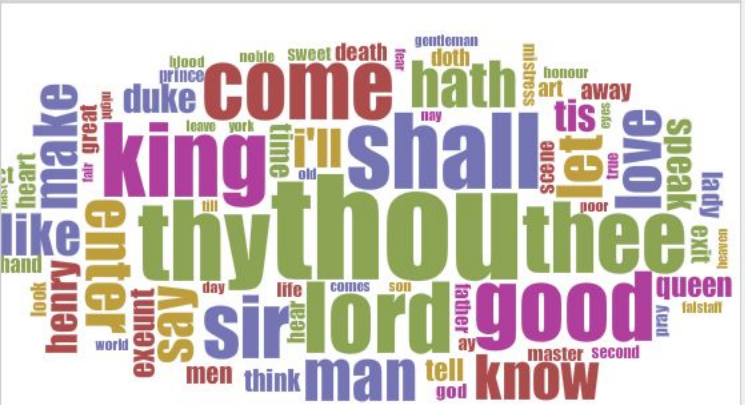
Search

Words in the Entire Corpus

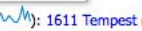
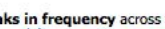
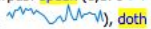
Corpus

Voyant Tools: Reveal Your Texts

Cirrus



Summary

- There are 37 documents in this corpus with a total of **890,366 words** and **28,750 unique words**.
- **Longest document** (by word count): 1599 Hamlet (32,212), 1593 King Richard III (31,663). Shortest d... (16,251), 1594 Midsummer Night's...; (17,195). All...
- Highest **vocabulary density** (): 1611 Tempest (187.9), 1605 Macbeth (186.7). Lowest density: 1593 King Richard III (132.1), 1600 Much Ado about Nothi...; (133.3). All...
- Most **frequent words** in the corpus: thou (5,354), thy (3,807), shall (3,585), lord (3,336), king (3,301). More...
- Words with **notable peaks in frequency** across the corpus: speak (1,176 ), tell (1,067 ), exit (976 ), comes (624 ), doth (866 ). More...
- **Distinctive words** (compared to the rest of the corpus)
  1. 1590 Love's Labour's Lost: biron (194), princess (130), ferdinand (127), armado (121), costard (116). More...
  2. 1591 Comedy of Errors: syracuse (232), dromio (221), antipholus (219), ephesus (169), adriana (90). More...
  3. 1591 Two Gentlemen of...;: proteus (218), valentine (216), julia (144), speed (128), silvia (123). More...
  4. 1594 Merchant of Venice: portia (148), bassanio (118), shylock (106), antonio (100), launcelot (83).

Click to see a table of documents

Words in the Entire Corpus

Corpus Reader

FERDINAND king of Navarre.

BIRON |

|

LONGAVILLE | lords attending on the King.

|

DUMAIN |

BOYET |

| lords attending on the Princess of France.

MERCADE |

DON

ADRIANO DE ARMADO a fantastical Spaniard.

SIR NATHANIEL a curate.

HOLOFERNES a schoolmaster.

DULL a constable.

COSTARD a clown.

MOTH page to Armado.

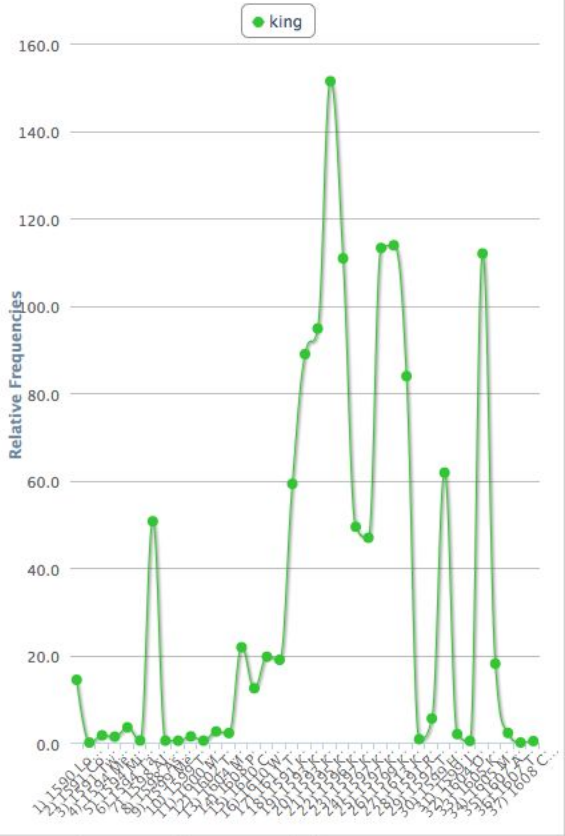
A Forester.

The PRINCESS of France: (PRINCESS:)

king

Corpus

Word Trends



Relative | Segments | king | collapse terms

Keywords in Context

Words in Documents