Day 0: Touching Base

Kenneth Benoit

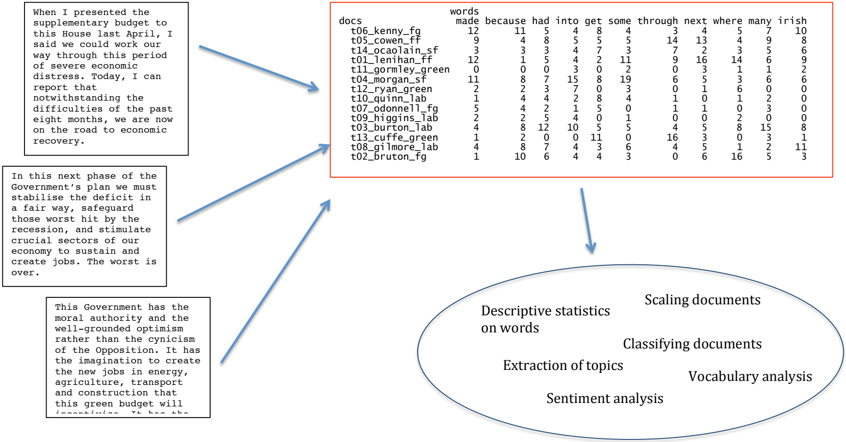
Quantitative Analysis of Textual Data

September 16, 2014

Targets

- Whom this class is for
- Learning objectives
  - fundamentals
  - availability and consequences of *choices*
  - practical ability to work with texts
  - issues of text for social science
- Prequisites
  - quantitative methods
  - familiarity with R
  - ability to use a text editor
  - (optional) ability to process text files in a programming language such as Python

Basic QTA Process: Texts → Feature matrix → Analysis



```
When I presented the
supplementary budget to
this House last April, I
said we could work our
way through this period
of severe economic
distress. Today, I can
report that
notwithstanding the
difficulties of the past
eight months, we are now
on the road to economic
recovery.
```

```
In this next phase of the
Government's plan we must
stabilise the deficit in
a fair way, safeguard
those worst hit by the
recession, and stimulate
crucial sectors of our
economy to sustain and
create jobs. The worst is
over.
```

```
This Government has the
moral authority and the
well-grounded optimism
rather than the cynicism
of the Opposition. It has
the imagination to create
the new jobs in energy,
agriculture, transport
and construction that
this green budget will
incentivise. It has the
```

```
                   words
docs                made because had into get some through next where many irish
  t06_kenny_fg        12       11   5    4   8    4       3    4     5    7    10
  t05_cowen_ff         9        4   8    5   5    5      14   13     4    9     8
  t14_ocaolain_sf      3        3   3    4   7    3       7    2     3    5     6
  t01_lenihan_ff      12        1   5    4   2   11       9   16    14    6     9
  t11_gormley_green    0        0   0    3   0    2       0    3     1    1     2
  t04_morgan_sf       11        8   7   15   8   19       6    5     3    6     6
  t12_ryan_green       2        2   3    7   0    3       0    1     6    0     0
  t10_quinn_lab        1        4   4    2   8    4       1    0     1    2     0
  t07_odonnell_fg      5        4   2    1   5    0       1    1     0    3     0
  t09_higgins_lab      2        2   5    4   0    1       0    0     2    0     0
  t03_burton_lab       4        8  12   10   5    5       4    5     8   15     8
  t13_cuffe_green      1        2   0    0  11    0      16    3     0    3     1
  t08_gilmore_lab      4        8   7    4   3    6       4    5     1    2    11
  t02_bruton_fg        1       10   6    4   4    3       0    6    16    5     3
```

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

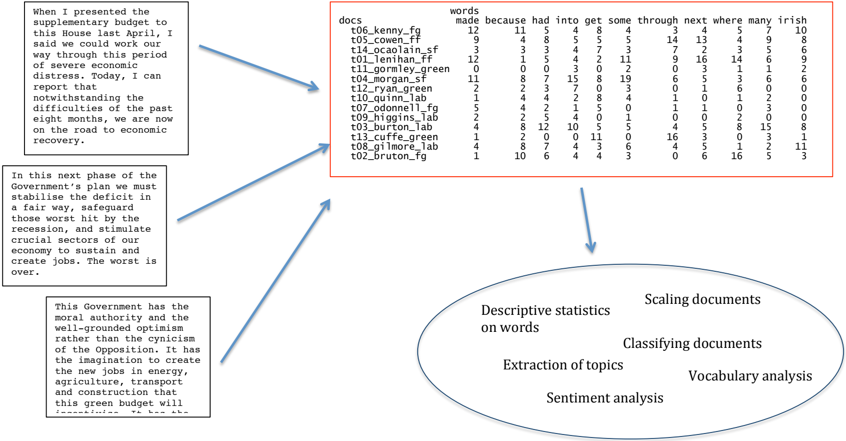
Sentiment analysis

What role for “qualitative” analysis in QTA?

- Ultimately all reading of texts is qualitative, even when we count elements of the text or convert them into numbers
- QTA may involve human judgment in the construction of the feature-document matrix
- But quantitative text analysis differs from more qualitiative approaches in that it:
    - Involves large-scale analysis of many texts, rather than close readings of few texts
    - Requires no interpretation of texts in a non-positivist fashion
    - Does not explicitly concern itself with the social or cultural predispositions of the analysts (not critical or constructivist)
- Uses a variety of statistical techniques to extract information from the document-feature matrix

Key feature of quantitative text analysis (cont.)

- Conversion of textual features into a quantitative matrix. Features can mean:
- A quantitative or statistical procedure to extract information from the quantitative matrix
- Summary and interpretation of the quantitative results



When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will incentivise. It has the

```
                     words
docs                  made because had into get some through next where many irish
  t06_kenny_fg          12      11   5    4   8    4       3    4     5    7    10
  t05_cowen_ff           9       4   8    5   5    5      14   13     4    9     8
  t14_ocaolain_sf        3       3   3    4   7    3       7    2     3    5     6
  t01_lenihan_ff        12       1   5    4   2   11       9   16    14    6     9
  t11_gormley_green      0       0   0    3   0    2       0    3     1    1     2
  t04_morgan_sf         11       8   7   15   8   19       6    5     3    6     6
  t12_ryan_green         2       2   3    7   0    3       0    1     6    0     0
  t10_quinn_lab          1       4   4    2   8    4       1    0     1    2     0
  t07_odonnell_fg        5       4   2    1   5    0       1    1     0    3     0
  t09_higgins_lab        2       2   5    4   0    1       0    0     2    0     0
  t03_burton_lab         4       8  12   10   5    5       4    5     8   15     8
  t13_cuffe_green        1       2   0    0  11    0      16    3     0    3     1
  t08_gilmore_lab        4       8   7    4   3    6       4    5     1    2    11
  t02_bruton_fg          1      10   6    4   4    3       0    6    16    5     3
```

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

LOGISTICS

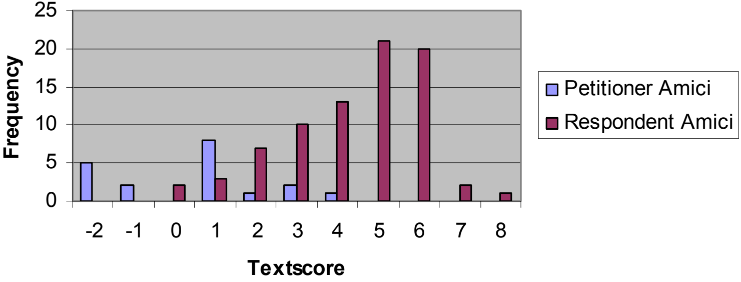
quanteda: R package

Course resources

- Syllabus: describes class, lists readings, links to reading, and links to exercises and datasets
- Web page on `http://www.kenbenoit.net/nyu2014qta`
  - Contains course handout
  - Slides from class
  - In-class exercises and supporting materials
  - Texts for analysis
  - (links to) Software tools and instructions for use
- Main readings
  - Lots of articles
  - Some other texts or on-line articles linked to the course handout (downloadable online)

EXAMPLES

Legal document scaling: “Wordscores”



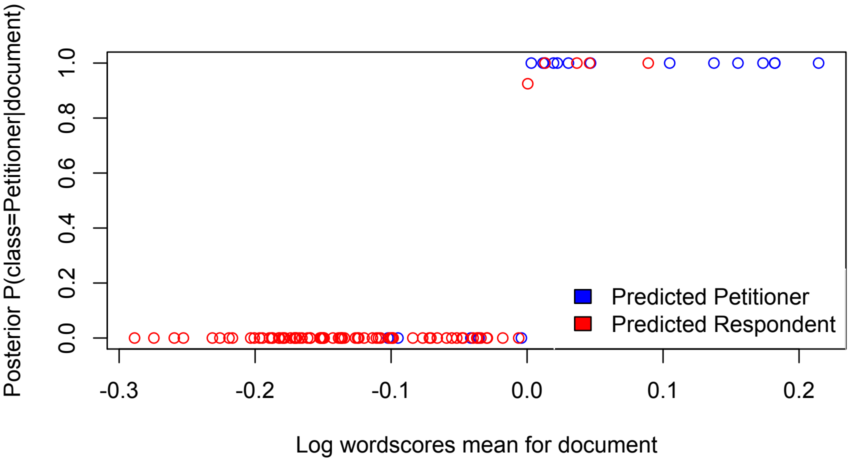
Amicus Curiae Textscores by Party

Using Litigants' Briefs as Reference Texts

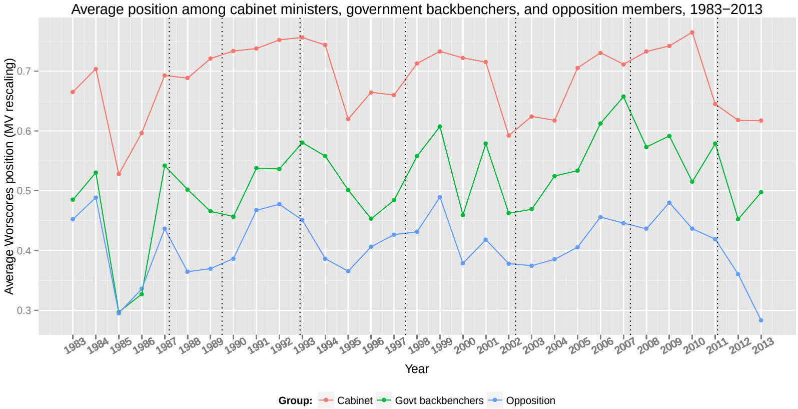
(Set Dimension: Petitioners = 1, Respondents = 5)

(from Evans et. al. 2007)

Document classification: “Naive Bayes” classifier



Government v. Opposition in yearly budget debates



(from Herzog and Benoit EPSA 2013)