Analysing texts with R

(and writing a package to do so)

Adam Obeng

About me: Adam Obeng

Computational Social Scientist (i.e. Data Scientist, Research Scientist, etc.)

ABD PhD in Sociology at Columbia

Jared taught me R

adamobeng.com

About me: Adam Obeng

Computational Social Scientist (i.e. Data Scientist, Research Scientist, etc.)

ABD PhD in Sociology at Columbia

Jared taught me R

Lucasarts

adamobeng.com

quanteda and readtext

Kenneth Benoit [aut, cre],
Paul Nulty [aut],
Kohei Watanabe [ctb],
Benjamin Lauderdale [ctb],
Adam Obeng [ctb],
Pablo Barberá [ctb],
Will Lowe [ctb]

Quantitative Text Analysis

Google "quantitative text analysis"

All Images News Videos Shopping More Search tools

About 11,900 results (0.34 seconds)

Scholarly articles for **"quantitative text analysis"**

Quantitative text analysis - Mehl - Cited by 59

... interest group influence using **quantitative text analysis** - Klüver - Cited by 148

A conceptual framework for **quantitative text analysis** - Roberts - Cited by 117

Quantitative Text Analysis (TCD 2016) | Kenneth Benoit's Home Page

www.kenbenoit.net/**quantitative-text-analysis**-tcd-2016/

Feb 4, 2016 - The course is also designed to cover many fundamental issues in **quantitative text analysis** such as inter-coder agreement, reliability, validation ...

Quantitative Text Analysis (TCD) | Kenneth Benoit's Home Page

www.kenbenoit.net/tcd2014qta/

Jan 8, 2014 - This course is aimed to provide a practical foundation to and a working knowledge of the main applied techniques of **quantitative text analysis** ...

Quantitative Text Analysis

Text as data:

- Linguistics
- Computer science
- Social sciences -> QTA

Roberts, Carl W. "A conceptual framework for quantitative text analysis." *Quality and Quantity* 34.3 (2000): 259-274.

QTA assumptions

- Texts reflect characteristics
- Texts represented by features
- Analysis estimates characteristics

QTA: Documents -> Document-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

docs	words 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

Ken Benoit, [The Quantitative Analysis of Textual Data (NYU Fall 2014)]

Outline

- Loading texts (descriptive stats)
- Extracting features
- Analysis: supervised scaling

+ Digressions about the process of writing an R package

QTA Step 1: Loading texts

Demo

Digression #1: how do we make it simple?

- v1.0 API changes to meet ROpenSci guidelines
  - namespace collisions
- Introducing readtext

Digression #1: readtext

```
readtext(
    file, ignoreMissingFiles = FALSE,
    textfield = NULL,
    docvarsfrom = c("metadata", "filenames"),
    dvsep = "_", docvarnames = NULL,
    encoding = NULL, ...)
```

Digression #1: readtext

- plaintext
- delimited text
- doc
- docx
- pdf
- JSON, line-delimited JSON, Twitter API output
- XML
- HTML
- zip, .tar, and .gz archives
- remote files
- glob paths

any (possible) combination of those

“any” encoding

```
> readtext('path/to/whatever')
```

just works™

Digression #1: listMatchingFiles

From a pseudo-URI, return all matching files

Given that:

- A URI can resolve to zero or more files (e.g. `'/path/to/*.csv'`, ‘https://example.org/texts.zip’)
- Globbing is platform-dependent (e.g. `'/path/to/\*.tsv'` escaping)
- Recursion

Digression #1 sub-digression #1

Some people, when confronted with a problem, think “I know, I'll use regular expressions.” Now they have two problems. — jzw

Digression #1 sub-digression #1

Some people, when confronted with a problem, think “I know, I'll use regular expressions.” Now they have two problems. — jzw

Digression #1: listMatchingFiles

- If it’s a remote file, download it
- If it’s an archive, extract it, glob the contents
- If it’s a directory, glob the contents

-> Call listMatchingFiles() on the result

Termination condition: was it a glob last time? (a glob cannot resolve to a glob)

https://github.com/kbenoit/readtext/blob/98dbccc9a3ac07f387ef94bcfecab0eb5282dc5b/R/utils.R#L87-L222

QTA Step 2: Extracting features
text -> dfm

- Feature creation (NLP)
  - tokenizing
  - removing stopwords
  - stemming
  - skip-ngrams
  - dictionaries
- Feature selection
  - Document frequency
  - Term frequency
  - Purposive selection
  - Deliberate disregard

Demo: extracting features

QTA Step 3: Analysis
Supervised scaling

Goal: differentiate document characteristics

e.g. where do they (or their authors) fall on the political spectrum

QTA Step 3: Analysis
Supervised scaling

Like ML classification, but continuous outcome:

- Get training (reference) texts
- Generate word scores in training texts
- Score test (virgin) texts
- Evaluate performance

Wordscores

Laver, Michael, Kenneth Benoit, and John Garry. "Extracting policy positions from political texts using words as data." American Political Science Review 97.02 (2003): 311-331.

QTA Step 3: Analysis

Supervised scaling demo

Digression #2: Testing

“Do you want your results to be correct or plausible?” — Greg Wilson

True for ML and for code

Digression #2: Testing

- Use CI as source of truth, not local tests (even with --as--cran)
  - (Still might not match CRAN)
- Enforce test coverage
- Test coverage is per-line

https://travis-ci.org/kbenoit/readtext
https://travis-ci.org/kbenoit/quanteda
https://codecov.io/gh/kbenoit/readtext
https://codecov.io/gh/kbenoit/quanteda

Digression #2: Testing

We discovered a lot of our own bugs

```
R CMD check out logs                                                                                    out
$ for name in $(find "${RCHECK_DIR}" -type f -name "*out");do echo ">>> Filename: ${name} <<<";cat ${name};done   0.01s
>>> Filename: readtext.Rcheck/00install.out <<<
* installing *source* package ‘readtext’ ...
** R
Warning in strsplit(msg, "\n") :
  input string 1 is invalid in this locale
Error in parse(outFile) : /home/travis/build/kbenoit/readtext/readtext.Rcheck/00_pkg_src/readtext/R/get-functions.R:167:38:
unexpected input
166:                            XML::xmlValue)
167:     txt <- txt[!grepl('^\\s*$', txt)]
                                          ^
ERROR: unable to collate and parse R files for package ‘readtext’
* removing ‘/home/travis/build/kbenoit/readtext/readtext.Rcheck/readtext’

R CMD check failed
                                                                                                        Top
```

Digression #2: Testing
Sometimes it’s R’s fault

`base::tempfile()`: (usually) different filenames within the same session

`base::tempdir()`: always the same directory name within the same session

`readtext::mktemp()` behaves like GNU coreutils mktemp

Digression #2: Testing
Sometimes it’s R’s fault

Jun 06 Adam Obeng readlines() truncates text file with Codepage 437 encoding – Hello r-devel, The

Jun 08 Martin Maechler Appended is the file -- you need to tell your e-mail software to use one of the MIME types that

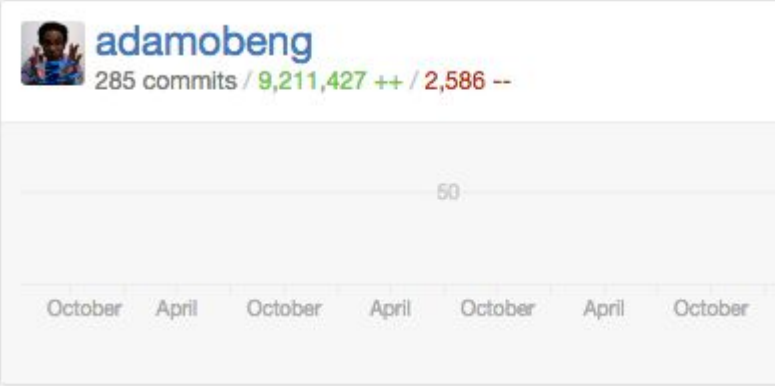
Jun 09 Martin Maechler I can reproduce the issue on Linux (Fedora F22), R 3.3.0 patched of today. Here's code for exp

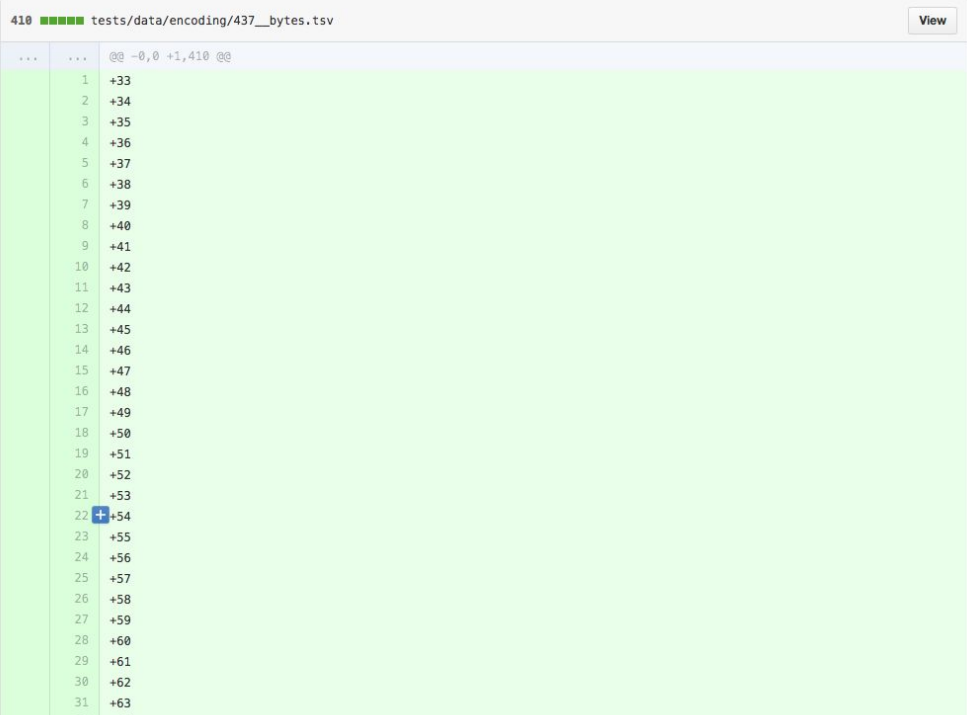
crickets

If you know what’s going on:
http://r.789695.n4.nabble.com/readlines-truncates-text-file-with-Codepage-437-encoding-td4721527.html

Digression #2 sub-digression #1: how to win at GitHub

Digression #2 sub-digression #1: how to win at GitHub





Thanks!

Slides and code: adamobeng.com

References:

- Ken Benoit, The Quantitative Analysis of Textual Data (NYU Fall 2014)
- — , Quantitative Text Analysis (TCD)

HERE BE DRAGONS

(Additional slides)

QTA Step 3: Analysis
Unsupervised scaling

Problems with Wordscores:

1. “the positions themselves are abstract concepts that cannot be observed directly”
2. the set of words may change over time

Wordfish

Slapin, Jonathan B., and Sven-Oliver Proksch. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52.3 (2008): 705-722.

QTA Step 3: Analysis

Unsupervised scaling: Wordfish

Naive Bayes with Poisson distributional assumption

QTA Step 3: Analysis

Unsupervised scaling demo

Digression #1: non-breaking spaces

```
R CMD check out logs                                                                                    out
$ for name in $(find "${RCHECK_DIR}" -type f -name "*out");do echo ">>> Filename: ${name} <<<";cat ${name};done   0.01s
>>> Filename: readtext.Rcheck/00install.out <<<
* installing *source* package ‘readtext’ ...
** R
Warning in strsplit(msg, "\n") :
  input string 1 is invalid in this locale
Error in parse(outFile) : /home/travis/build/kbenoit/readtext/readtext.Rcheck/00_pkg_src/readtext/R/get-functions.R:167:38:
unexpected input
166:                            XML::xmlValue)
167:     txt <- txt[!grepl('^\\s*$', txt)]
                                          ^
ERROR: unable to collate and parse R files for package ‘readtext’
* removing ‘/home/travis/build/kbenoit/readtext/readtext.Rcheck/readtext’

R CMD check failed
                                                                                                        Top
```

Digression #1: non-breaking spaces

Digression #1: non-breaking spaces

⌥ Opt+3 -> #

⌥ Opt+Space -> \xa0

```
3 ## csv format¬
4 get_csv <- function(path, textfield, ...) {¬
5 ¬
```

Solution: [pre-commit hook]

Back to the demo: loading text and descriptive stats

Digression #4: Git is a literal genie

Restore lost history (squashed commit:) #261

Merged **kbenoit** merged 1 commit into `kbenoit:master` from `adamobeng:fix_git` on Oct 26

Conversation 2 | Commits 1 | Files changed 20

Changes from **all commits** | **20 files** | +145 −126

Squashed commit of the following:

```
commit 9dd74ad
Author: Kenneth Benoit <kbenoit@lse.ac.uk>
Date:   Mon Oct 24 14:05:07 2016 +0100

    Correct build errors

commit b69a89b
Author: Kenneth Benoit <kbenoit@lse.ac.uk>
Date:   Mon Oct 24 12:23:49 2016 +0100
```

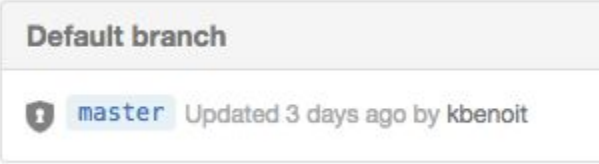
Digression #4: Git is extremely elegant

Git for Computer Scientists

But the porcelain is equally difficult to use

Digression #4: Git needs additional constraints

Don’t allow commits to master:



git-flow?

Documents

Usually texts, but also paragraphs, etc.

Features

- words
- n-grams
- skip-grams
- dictionaries
- phrases
- manual coding
- etc.

Analysis

- Descriptive stats
- Supervised scaling and classification
- Unsupervised scaling
- Clustering and topic models