
R-Ladies
R-Ladies Jozi Nov meetup

About R-Ladies

- Worldwide organisation whose mission is to promote gender diversity in the R community
- R-Ladies' primary focus is on supporting minority gender R enthusiasts to achieve their programming potential, by building a collaborative global network
- Our aim is to create a warm and welcoming place for R users to exchange ideas, learn and collaborate

Connect with us

Meetup

https://www.meetup.com/rladies-johannesburg/

@RLadiesJozi
@RLadiesGlobal
@WeAreRLadies

https://github.com/orgs/rladies/teams/jozi/repositories

slack

https://rladies-community-slack.herokuapp.com/

December meetup

Year-end event:

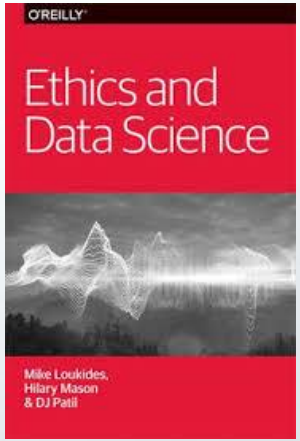
The Computers documentary
11 December
Braamfontein

More info about the project:
http://eniacprogrammers.org/

Topics for 2019

Data science and ethics

Data visualisation



Please share ideas!

THANK YOU

CoE for Mathematical and Statistical Sciences

International Relations

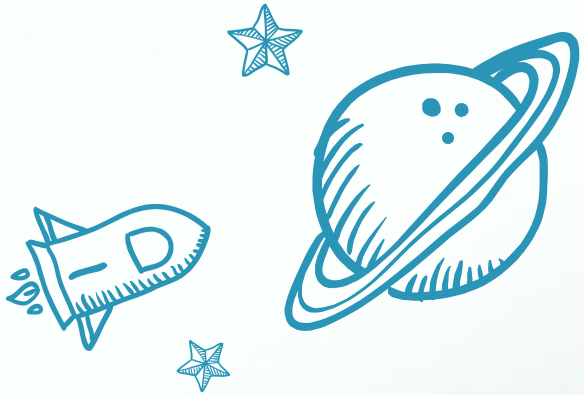
Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the CoE-MaSS.

Introduction to text mining in r

Retha Langa

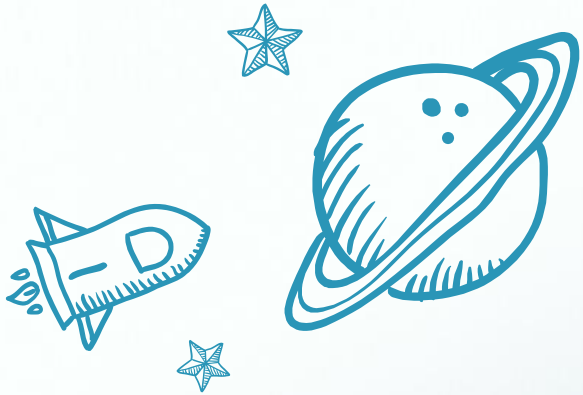

Contents

- 3 main ideas for tonight
- Text mining
- *Dubula ibhunu* trial case study
- A practical example
- Resources

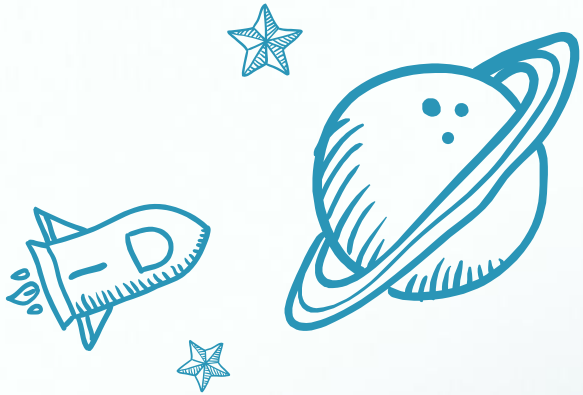
Know your data

Knowing your data and having solid industry/domain knowledge will lead to better questions and richer findings



Think about your ?

"The art and science of asking questions is the source of all knowledge" – Thomas Berger



Broaden your horizons

Applying ways of doing things from different fields can unlock some seriously cool – and ethical – stuff.

Process of analyzing collections of textual materials in order to capture key concepts and themes and uncover hidden relationships and trends

Source: IBM

RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5,*,†] Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17,*,†]

Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014

Alix Rule[a], Jean-Philippe Cointet[b], and Peter S. Bearman[a,1]

Article

Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy

Social Science Computer Review
2015, Vol. 33(1) 3-20

DOI: 10.1177/0894439314521983
ssc.sagepub.com
SAGE

Andrea Ceron[I], Luigi Curini[I], and Stefano M. Iacus[I]

Content analysis of 150 years of British periodicals

Thomas Lansdall-Welfare[a], Saatviga Sudhahar[a], James Thompson[b], Justin Lewis[c], FindMyPast Newspaper Team[d,1], and Nello Cristianini[a,2]

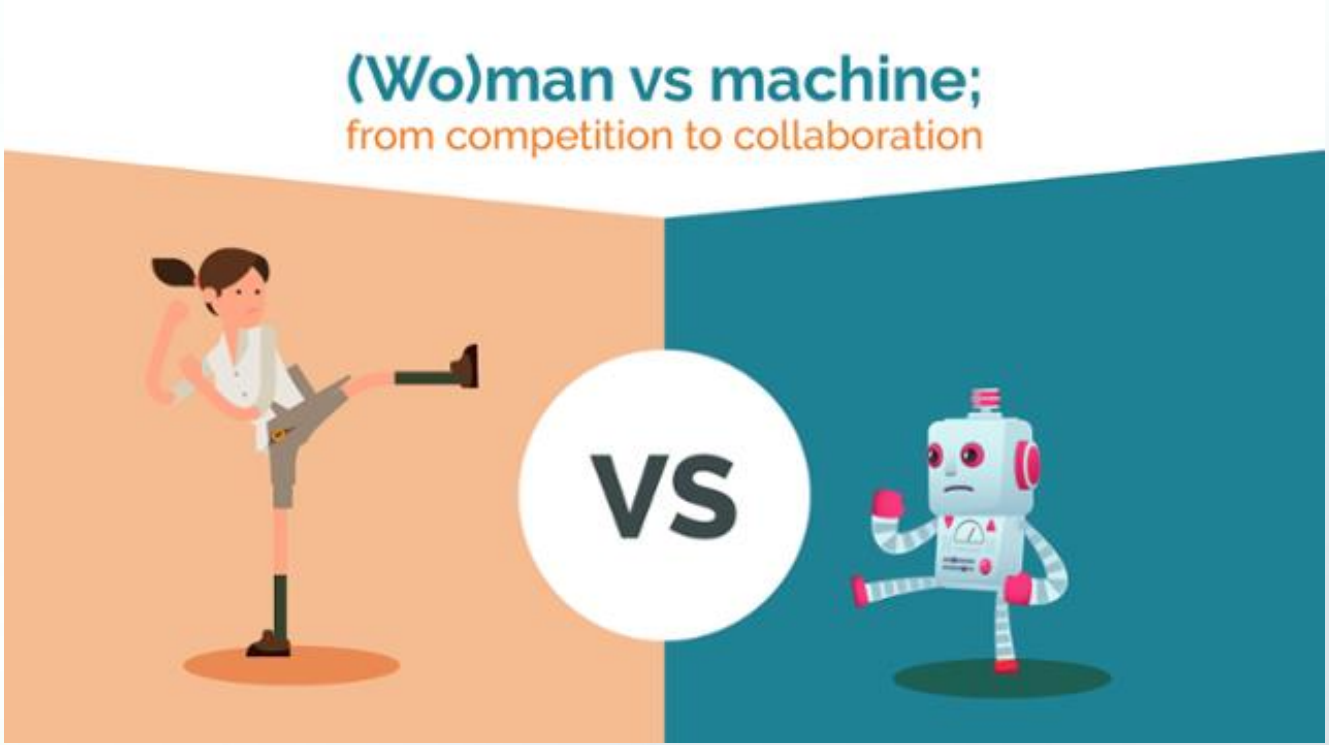
Social media competitive analysis and text mining: A case study in the pizza industry

Wu He[a,*], Shenghua Zha[b,1], Ling Li[a,c,2]

Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews

Katerina Berezina, Anil Bilgihan, Cihan Cobanoglu & Fevzi Okumus

Dubula ibhunu Trial case study


(Wo)man vs machine;
from competition to collaboration
VS

How did the different organisations and witnesses deploy the land issue in court?

111 267

Words of testimony

3 organisations

ANC, AfriForum & TAU SA

11

Witness testimonies

Derek Hanekom

Collins Chabane

Dr Mongane Wally Serote

Gwede Mantashe

Julius Malema

Photos: Tourism.gov; SABC; City Press; Wikipedia

3 METHODOLOGICAL CHOICES

Bag of words

Words are treated as single tokens.

Only the frequencies of words per text are used and their positions are ignored.

Dictionary

Manual approach to compiling the list of keywords.

Sentiment analysis

Testimonies were characterised by a high level of emotion.

Witnesses often became emotional during the trial, especially when testifying about land reform.

3 THEMES

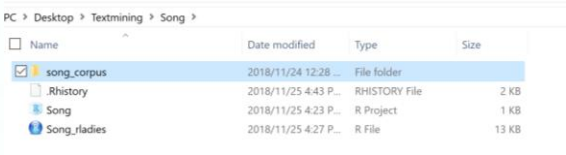
Photos: The Heritage Portal; Independent

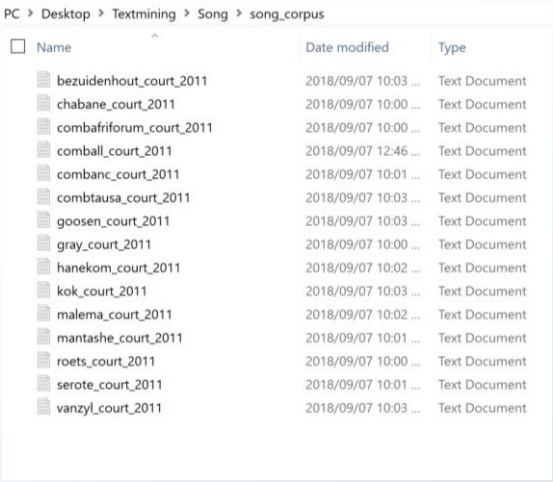
Turning to R

Text mining steps

Importing text

STEP 1: Create a corpus





A corpus in text mining is a database for holding and managing the text documents to be analysed – i.e. a text document collection.

Packages

```
library(readtext)
library(stringr)
library(quanteda)
library(tidytext)
library(ggplot2)
library(dplyr)
library(tidyr)
```



Pre-processing

Common steps include:

- Using string operations to modify or remove parts of the text
- Splitting the text into tokens (tokenization)
- Normalization of the text (for example, converting to lower-case and stemming)
- Removing stopwords, punctuation and numbers

Reading-in your data

```
21
22 court <- readtext("song_corpus/*_court_2011.txt")
23 court
24
25 random <- c("lordship", "lord", "sir", "indistinct", "inaudible")
26
27 court$text <- tolower(court$text)
28 random.regex <- paste(random, collapse = "|")
29 random.regex
30
31 court$text <- str_remove_all(court$text, pattern = random.regex)
32 court
33
34
```

Pre-processing and creating a DTM

What is a Document-Term Matrix?

- Document term matrix is one of the most common formats for representing a text corpus in a bag-of-words format.
- A DTM is a matrix in which rows are documents, columns are terms, and cells indicate how often each term occurred in each document.

Pre-processing and creating a DTM

```
#Creating document-term matrix

court.corpus <- corpus(court)
court.dtm <- dfm(court.corpus,
                 tolower = TRUE,
                 stem = TRUE,
                 remove_punct = TRUE,
                 remove_numbers = TRUE,
                 remove_hyphens = TRUE,
                 remove_symbols = TRUE,
                 remove = stopwords("english"))

```

What does it look like?

```
> court.dtm[, 1:10]
Document-feature matrix of: 15 documents, 10 features (27.3% sparse).
15 x 10 sparse Matrix of class "dfm"
                           features
docs                        read correct okay music general probabl biggest tool influenc human
  bezuidenhout_court_2011.txt   13       4    9    22       5       4       1    8       21     2
  chabane_court_2011.txt         2       3    6     3       3       0       0    0        0     0
  combafriforum_court_2011.txt  18      24   18    52       8      11       1    5        6     1
  comball_court_2011.txt        80     243   84    78      54      26       5   16       31    20
  combanc_court_2011.txt        41     128   57     3      33      10       3    2        4    17
  combtausa_court_2011.txt      21      91    9    23      13       5       1    9       21     2
  goosen_court_2011.txt          3      33    0     1       2       0       0    0        0     0
  gray_court_2011.txt            4       4    3    50       2       5       0    5        4     0
  hanekom_court_2011.txt        18      60   27     0      12      10       2    2        2     5
  kok_court_2011.txt             1      24    0     0       0       0       0    1        0     0
  malema_court_2011.txt          8      13    4     0       7       0       0    0        2     8
  mantashe_court_2011.txt        5      16   10     0       8       0       0    0        0     3
  roets_court_2011.txt          14      20   15     2       6       6       1    0        2     1
  serote_court_2011.txt          8      36   10     0       3       0       1    0        0     1
  vanzyl_court_2011.txt          4      30    0     0       6       1       0    0        0     0
>
```

Analysing your DTM – A Dictionary approach

```
#Creating a dictionary of keywords that belong in the three clusters (land, reconciliation & transformation)

my_Dict <- dictionary(list(land = c("land*"), farm = c("farm*"), farmer = c("farmer*"),
                           transform = c("transform*"), freedom = c("freedom*"),
                           revolution = c("revolu*"),
                           cohesion = c("cohes*"),
                           reconciliation = c("reconcil*"), redress = c("redress*"),
                           mandela = c("mandela*"), dialogue = c("dialog*"),
                           nationalisation = c("nationali*"), compensation = c("compens*"),
                           rainbow = c("rainbow*"), agriculture = c("agricultur*"),
                           charter = c("charter*"),
                           constitution = c("constituti*")))
dict_court.dtm <- dfm_lookup(court.dtm, my_Dict, nomatch = "unmatched")
dict_court.dtm

```

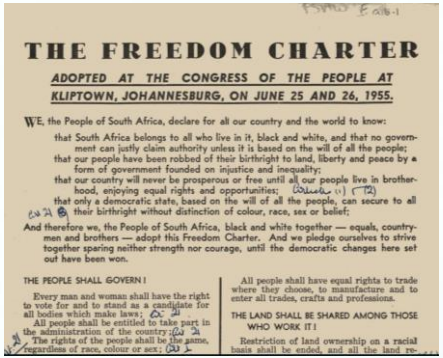
The result

```
> dict_court.dtm
Document-feature matrix of: 15 documents, 18 features (44.4% sparse).
15 x 18 sparse Matrix of class "dfm"
                              features
docs                           land farm farmer transform freedom revolution cohesion reconciliation redress mandela
  bezuidenhout_court_2011.txt     2   25      6         2       2          4        0              5       0       0
  chabane_court_2011.txt          1    0      0         0       1          0        0              0       0       0
  combafriforum_court_2011.txt   11   45     18         0       2          0        3              9       0       5
  comball_court_2011.txt         82  219    126        15      63         40        3             23      13      23
  combanc_court_2011.txt         66  128     92        12      47         36        0              9      13      18
  combtausa_court_2011.txt        5   46     16         3      14          4        0              5       0       0
  goosen_court_2011.txt           0    2      2         0       9          0        0              0       0       0
  gray_court_2011.txt            11    7      1         0       0          0        0              7       0       5
  hanekom_court_2011.txt          2   70     50         1      11          1        0              3       0       3
  kok_court_2011.txt              0    1      1         1       0          0        0              0       0       0
  malema_court_2011.txt          51   44     29         8      18         28        0              0       0      11
  mantashe_court_2011.txt         7    6      5         0       5          2        0              2       0       1
  roets_court_2011.txt            0   38     17         0       2          0        3              2       0       0
  serote_court_2011.txt           5    8      8         3      12          5        0              4      13       3
  vanzyl_court_2011.txt           3   18      7         0       3          0        0              0       0       0
```

RECONCILIATION

	Frequency	%
Cohesion	3	0.5
Constitution	45	6.9
Dialogue	66	10.1
Mandela	23	3.5
Rainbow	1	0.2
Reconciliation	23	3.5
Redress	13	2
TOTAL	**174**	**26.6**

Transformation



THE FREEDOM CHARTER

ADOPTED AT THE CONGRESS OF THE PEOPLE AT KLIPTOWN, JOHANNESBURG, ON JUNE 25 AND 26, 1955.

WE, the People of South Africa, declare for all our country and the world to know:

that South Africa belongs to all who live in it, black and white, and that no government can justly claim authority unless it is based on the will of all the people;

that our people have been robbed of their birthright to land, liberty and peace by a form of government founded on injustice and inequality;

that our country will never be prosperous or free until all our people live in brotherhood, enjoying equal rights and opportunities;

that only a democratic state, based on the will of all the people, can secure to all their birthright without distinction of colour, race, sex or belief;

And therefore we, the People of South Africa, black and white together — equals, countrymen and brothers — adopt this Freedom Charter. And we pledge ourselves to strive together sparing neither strength nor courage, until the democratic changes here set out have been won.

THE PEOPLE SHALL GOVERN!

Every man and woman shall have the right to vote for and to stand as a candidate for all bodies which make laws;

All people shall be entitled to take part in the administration of the country.

The rights of the people shall be the same, regardless of race, colour or sex;

All people shall have equal rights to trade where they choose, to manufacture and to enter all trades, crafts and professions.

THE LAND SHALL BE SHARED AMONG THOSE WHO WORK IT!

Restriction of land ownership on a racial basis shall be ended, and all the land re-...

	Frequency	%
Charter	20	3.1
Freedom	63	9.6
Nationalisation	6	0.9
Revolution	40	6.1
Transformation	15	2.3
TOTAL	144	22.1

LAND

	Frequency	%
Agriculture	29	4.4
Compensation	5	0.8
Farm	93	14.2
Farmer	126	19.3
Land	82	12.6
TOTAL	**335**	**51.3**

ANC, TAU SA and Afriforum

	Reconciliation	Transformation	Land
ANC	115	120	208
AfriForum	38	3	66
TAU SA	21	21	61

Derek Hanekom

Collins Chabane

Dr Mongane Wally Serote

Gwede Mantashe

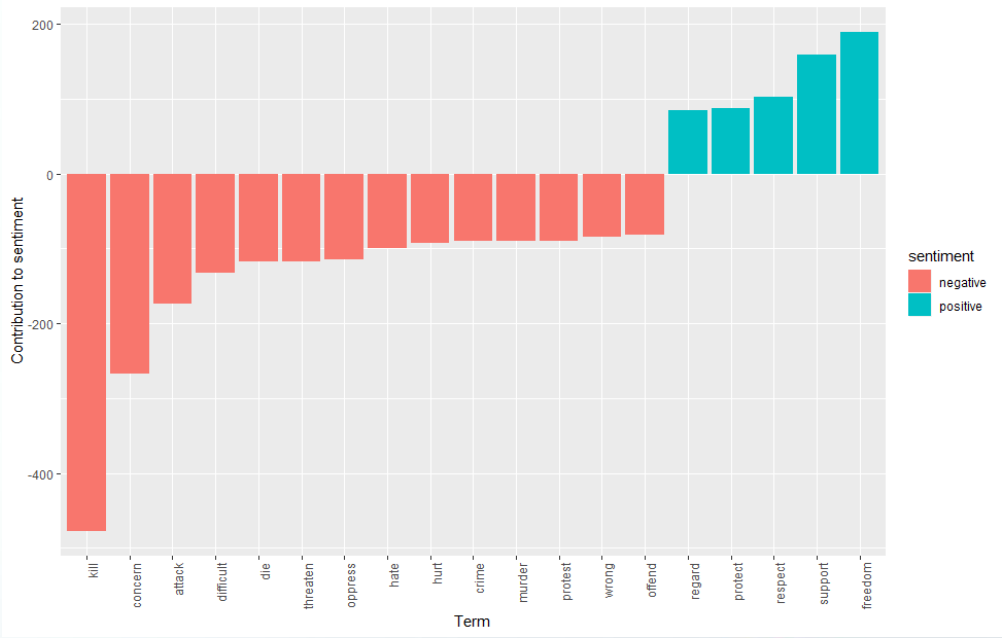
Julius Malema

Photos: Tourism.gov; SABC; City Press; Wikipedia

Malema and the rest of the ANC

	Reconciliation	Transformation	Land
Malema	29	76	112
ANC veteran	42	29	42

Sentiment analysis

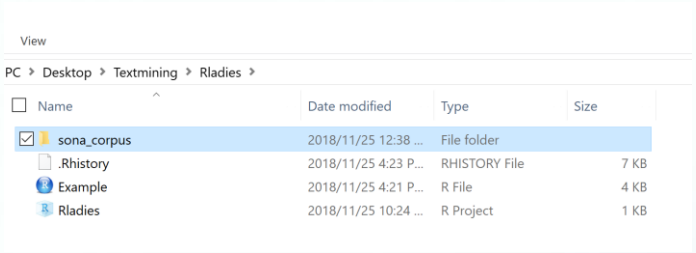


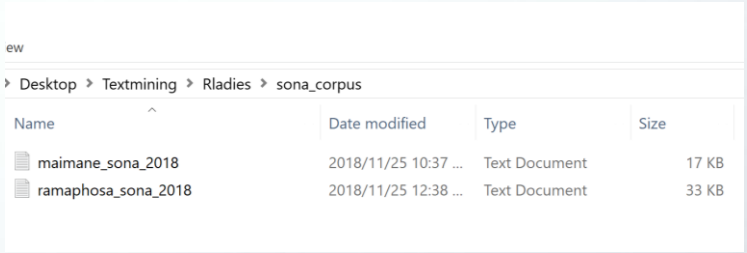
Lets practise

2018 STATE OF THE NATION AND OPPOSITION REPLY

Importing text

STEP 1: Create a corpus





RESOURCES

The quest for better questions: https://bit.ly/2OZUifa

Ethical AI? It's All About Perspective: https://bit.ly/2AnvPuL

A wolf in sheep's clothing: Disruption is overrated in terms of innovation: https://bit.ly/2BzVwdH

Biography of an uncharted people: https://unchartedpeople.org/

What is the role of the Arts and Humanities in the age of Data Science? Two short proposals: https://bit.ly/2ByRylj

RESOURCES

Text mining with R: A tidy approach by Julia Silge & David Robinson

https://www.tidytextmining.com/

Julia Silge:

https://juliasilge.com/

@juliasilge

Quanteda package materials:
https://quanteda.io/

Text Analysis in R article (hand-out)

Kaggle datasets:

https://bit.ly/2THxFzq

RESOURCES

General resources:

https://github.com/stepthom/text_mining_resources

Conversion Vignette:
https://bit.ly/2P2UJ8i

Plotting: https://bit.ly/2PUm1Dk




Thank you!

Any questions?

You can find me at:

- @HeritageAffair
- rethag@gmail.com