

A taste of text mining

PEETER TINITS

#DIGMET SUMMER SCHOOL, TARTU

28.08.2019

Billboards dataset

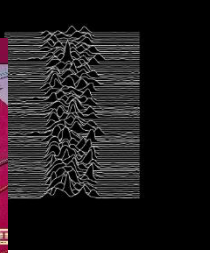
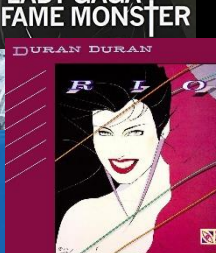
Billboard HOT 100

FOR WEEK ENDING OCTOBER 8, 1988

Billboard **HOT 100** **SINGLES™**

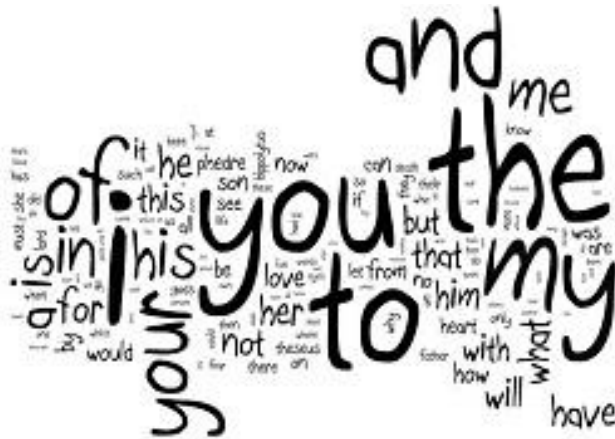
Compiled from a national sample of retail store and one-stop sales reports and radio playlists.

THIS WEEK	LAST WEEK	2 WKS AGO	WKS ON CHART	TITLE PRODUCER (SONGWRITER)	ARTIST LABEL & NUMBER/DISTRIBUTING LABEL
1	2	5	9	★ ★ No. 1 ★ ★ LOVE BITES R.LANGE (CLARK, COLLEN, ELLIOTT, LANGE, SAVAGE)	◆ DEF LEPPARD (C) MERCURY 870 402-7/POLYGRAM
2	5	13	24	RED RED WINE UB40, R.FALCONE (N.DIAMOND)	◆ UB40 (C) A&M 1244
3	1	1	11	DON'T WORRY, BE HAPPY (FROM "COCKTAIL") L.GOLDSTEIN (B.MCFERRIN)	◆ BOBBY MC FERRIN (C) EMI-MANHATTAN 50146
4	6	10	11	DON'T BE CRUEL R.ZITO (D.BLACKWELL, E.PRESLEY)	◆ CHEAP TRICK (C) EPIC 34-07965/E.P.A.
5	4	7	12	ONE GOOD WOMAN P.LEONARD, P.CETERA (P.CETERA, R.LEONARD)	◆ PETER CETERA (C) (CD) FULL MOON 7-27824/WARNER BROS.
6	14	21	6	GROOVY KIND OF LOVE P.COLLINS, A.DUDLEY (T.WINE, C.BAYER BACHARACH)	◆ PHIL COLLINS (T) (C) ATLANTIC 7-89017
7	3	3	18	I'LL ALWAYS LOVE YOU R.WAKE (J.GEORGE)	◆ TAYLOR DAYNE (T) (C) ARISTA 1-9700
8	8	12	16	I HATE MYSELF FOR LOVING YOU D.CHILD, K.LAGUNA (J.JETT, D.CHILD)	◆ JOAN JETT AND THE BLACKHEARTS (C) BLACKHEART 4-07919/E.P.A.
9	10	17	13	WHAT'S ON YOUR MIND (PURE ENERGY) F.MAHER (PROBB, K.VLAQUE)	◆ INFORMATION SOCIETY (T) (C) (M) TOMMY BOY 7-27826/REPRISE
10	12	14	16	PLEASE DON'T GO GIRL M.STARR (M.STARR)	◆ NEW KIDS ON THE BLOCK (T) (C) COLUMBIA 38-07700
11	15	19	12	DON'T BE CRUEL L.A.BABYFACE (BABYFACE, L.A.REID, D.SIMMONS)	◆ BOBBY BROWN (T) (C) MCA 53327
12	16	18	11	FALLEN ANGEL T.WERMAN (B.DALL, C.C.DEVILLE, B.MICHAELS, R.ROCKETT)	◆ POISON (C) ENIGMA 44191/CAPITOL
13	18	20	8	DON'T YOU KNOW WHAT THE NIGHT CAN DO? S.WINWOOD, T.LORD-ALGE (S.WINWOOD, W.JENNINGS)	◆ STEVE WINWOOD (T) (C) VIRGIN 7-99290



Stopwords

Common words that carry little content



associated
at
available
away
awfully
b
back
backward
backwards
be
became
because
become
becomes
becoming
been
before
beforehand
begin
behind
being
believe
below
beside
besides
best
better
between
beyond

keeps
kept
know
known
knows
l
last
lately
later
latter
latterly
least
less
lest
let
let's
like
liked
likely
likewise
little
look
looking
looks
low
lower
ltd
m
made

thanx
that
that'll
thats
that's
that've
the
their
theirs
them
themselves
then
thence
there
thereafter
thereby
there'd
therefore
therein
there'll
there're
theres
there's
thereupon
there've
these
they
they'd
they'll

Keywords

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

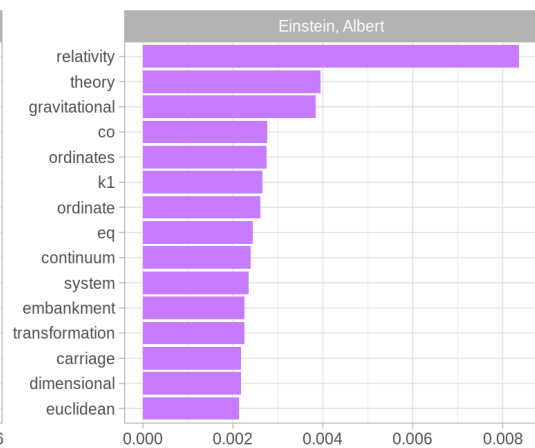
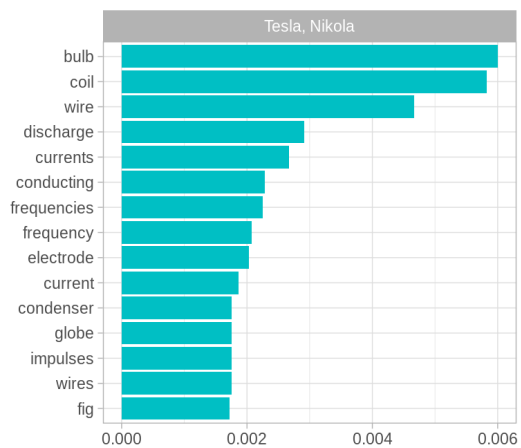
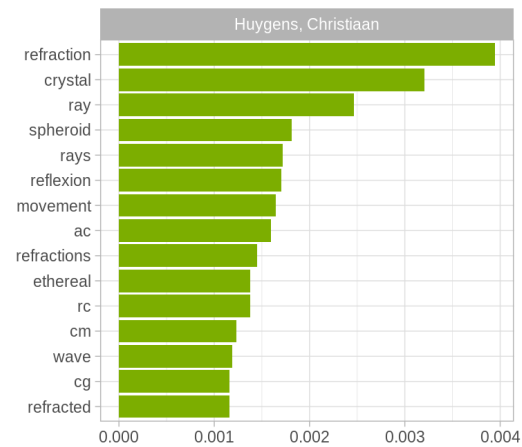
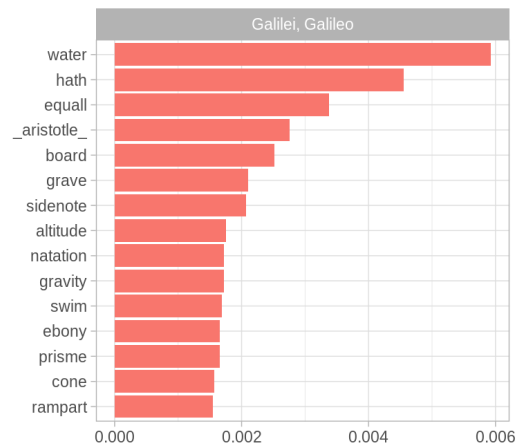
Term frequency

Number of times term t appears in a doc, d

Inverse document frequency


$$\log \frac{1 + \overset{\text{\# of documents}}{n}}{1 + \underset{\text{Document frequency of the term } t}{df(d, t)}} + 1$$

Keywords



tf-idf

Sentiments

 API TEST TOOL

English

Sentiment

Graphical

I ¹really enjoyed using the ¹Canon Ixus in Madrid on March 4. The ²Panasonic Lumix ²is a bit disappointing, but the ³Canon ³camera is ³not bad at all. All I want when taking photos is point it and then just press the button. For only 200 dollars, a ⁴really fair ⁴price, this ⁵camera is ⁵perfect for me. Besides, I have had a ⁶good ⁶customer ⁶service ⁶experience. ⁷John Faraday was ⁷very nice!

LEGEND color key

SENTIMENT

Sentiment topic

Positive sentiment text

Negative sentiment text

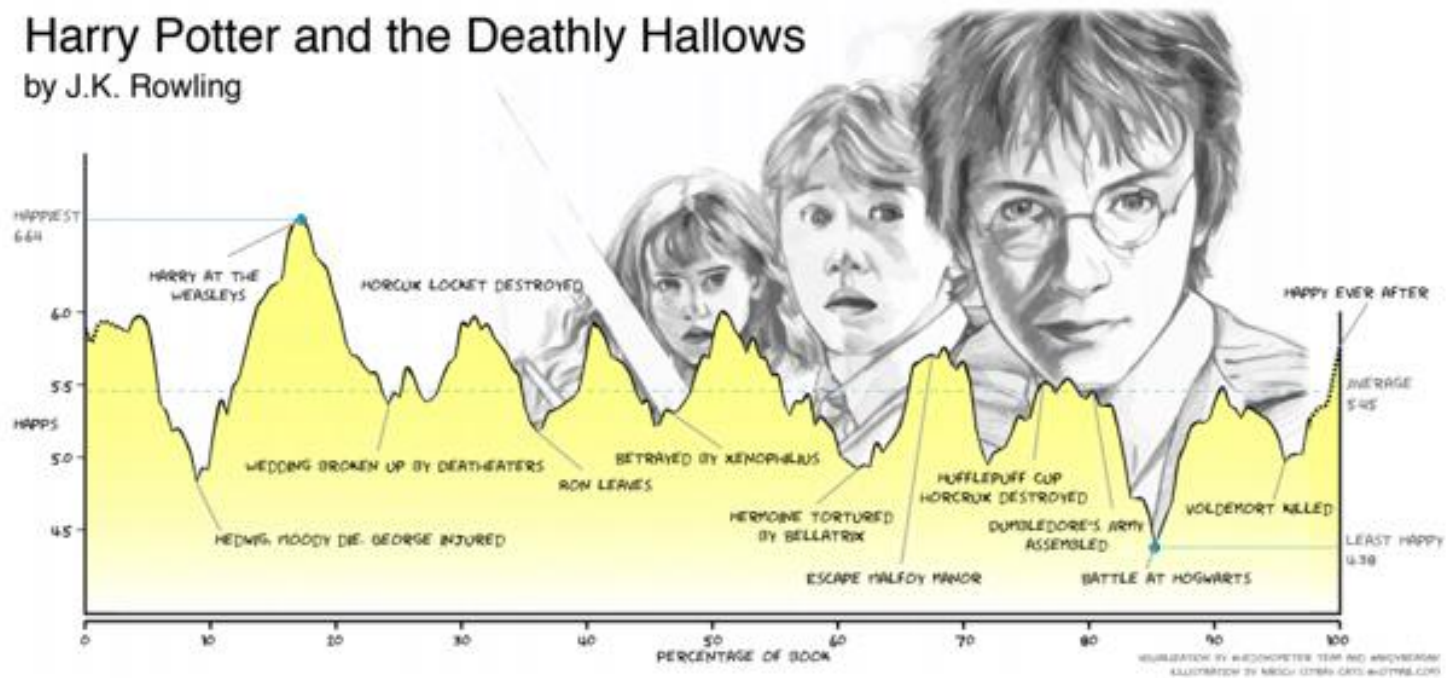
¹ Text and topic link

ANALYZE TEXT ▶

RESET ↺

Sentiments

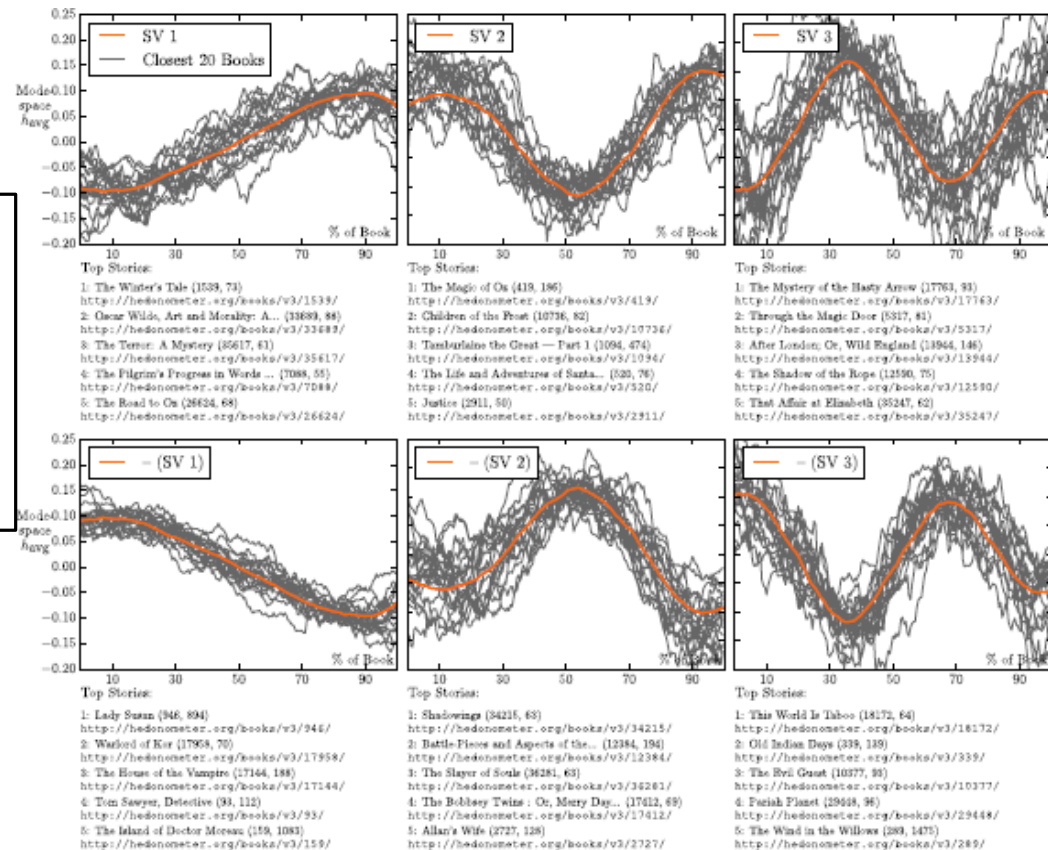
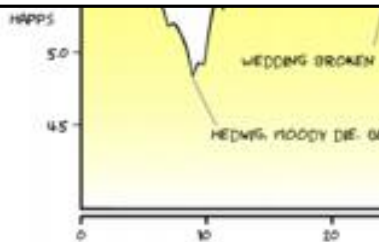
Harry Potter and the Deathly Hallows by J.K. Rowling



Sentiments

Harry Potter and

1. Rags to Riches (rise)
2. Riches to Rags (fall)
3. Man in a Hole (fall then rise)
4. Icarus (rise then fall)
5. Cinderella (rise then fall then rise)
6. Oedipus (fall then rise then fall)



TED talks dataset



TED Ideas worth spreading

Questions to look at

How many words per minute?

Which positive words are frequent?

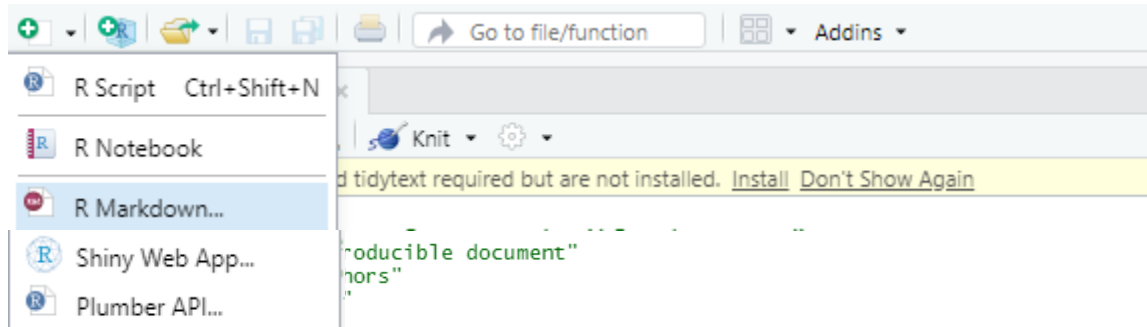
Who get the most laughs?

- Do laughs correlate with views?

What do top talks talk about?

What were the keywords each year?

Reproducible documents



```
6 ---
7
8 {r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 lapply(c("tidytext", "tidyverse", "here", "ggrepel"),
11        function(x) if(!is.element(x, installed.packages())) {
12
13
14 library(tidyverse)
15 library(tidytext)
16 library(ggrepel)
17
18
19 ## TED talks corpus study
20
21 Here we study the corpus of TED talks. We are really happy th
22
23 ### Data
24
```

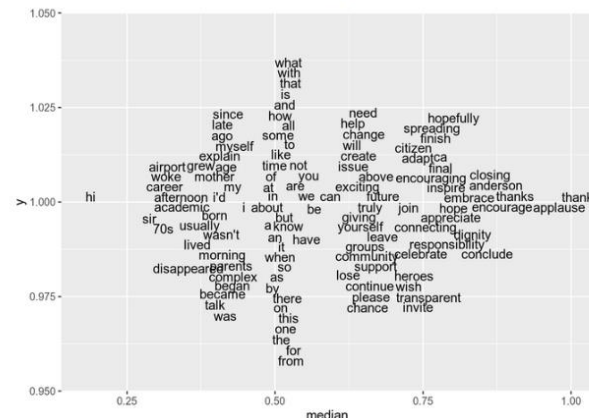
tidytext required but are not installed. [Install](#) [Don't Show Again](#)

```
mutate(count_n(), in_talks=n_distinct(talkid)) %>%
ungroup()
```

And we can add absolute counts of how many observations and talks there are per word

```
ted_tokens <- ted_tokens %>%
  group_by(word) %>%
  mutate(count_n(), in_talks=n_distinct(talkid)) %>%
  ungroup()
```

And we found out something interesting



<https://rmarkdown.rstudio.com>

Some summary

Reproducible research is a good way

- but also a lazy way

Code is not meant to be unreadable.

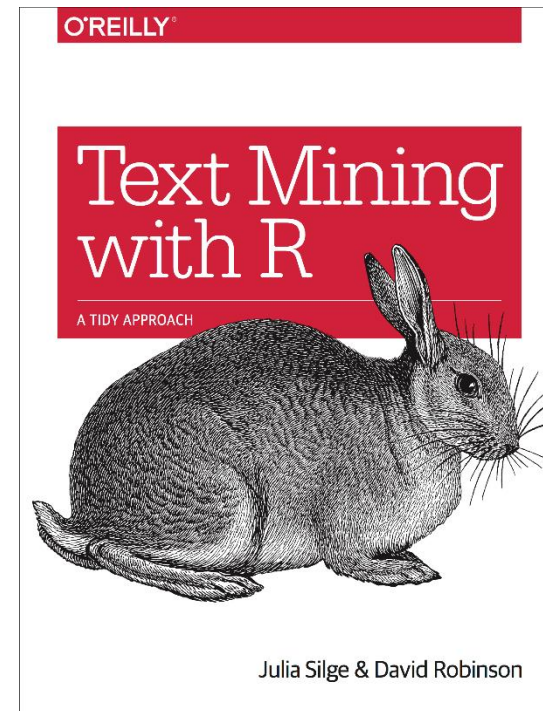
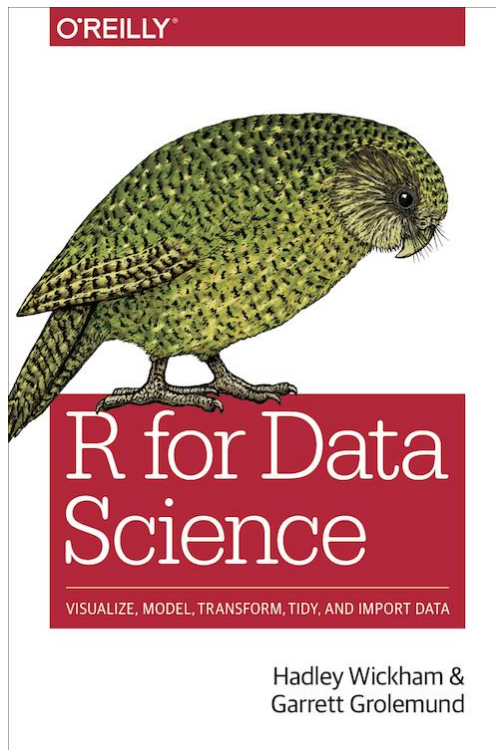
- Literate programming for readers and yourself

Some text mining is simple 😊

- But beware caveats and assumptions!

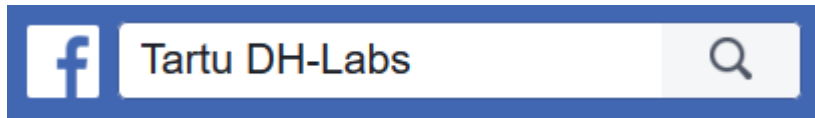
No right answers but good to find standards

Follow-ups



[Golemund, G.; Wickham, H. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data](#)
[Silge, J.; Robinson, D. 2017. Text Mining with R. A tidy approach.](#)

Follow-ups



Tartu DH-Labs

Grupp

30 liiget

✓ Liitunud

Tartu DH-Lab: <http://bit.do/Tartu-DH-labs>

- Open room Fridays in UT library
- Activities planned for Fall