BAN432
Applied Textual Data Analysis for Business and Finance

Keywords in Context

Christian Langerfeld and Maximilian Rohrer

30 September 2022

Packages we use today

- quanteda
- dplyr
- tidytext
- wordcloud
- stopwords

Data:

firm_dataset.Rdata from the last lecture

Structure of the course

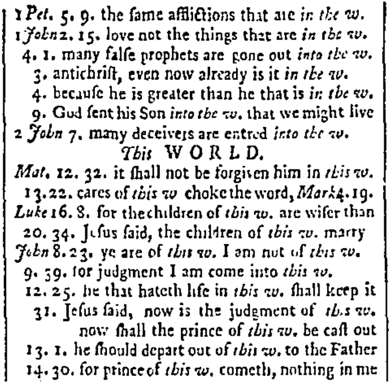
1	Introduction to course & basic R	Introduction
2	Introduction to R, specific to textual analysis	
3	Collecting textual data: APIs	Collecting data
4	Collecting textual data: EDGAR	
5	Preprocessing and cleaning, part I	Preprocessing data
6	Preprocessing and cleaning, part II	
7	Guest Lecture: Gisle Andersen (NHH)	
8	Regex-based application, Geography	Analyses
9	Regex-based application, Keyword in Context	
10	Automatic text summarization	
11	Sentiment: Twitter & Critical understanding	
12	Sentiment: Finance application	
13	Doc-Clustering: Cosine similarity & k-means	
14	Doc-clustering: Topic models	
15	Doc-Clustering: Multinominal Inverse Regression	
16	Guest Lecture: Vegard Larsen (Norges Bank)	
17	Contemporaneous papers in Finance	
18	Recap	

Plan for this lecture:

- **K**ey **W**ords **i**n **C**ontext: KWIC
- keywords in annual reports

Examples for Concordances

- in pre-digital times
  - in the middle ages: concordances of the Bible; words in the Bible listed in alphabetical order with context
  - literature: concordances of important texts by poets such as Homer or Shakespeare
  - in libraries: concordances of book titles
- in the digital age:
  - in research, mostly in the humanities

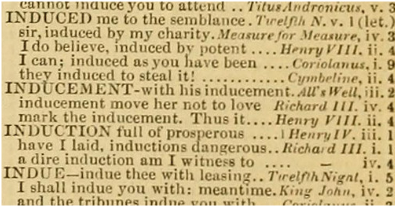


1 *Pet.* 5. 9. the ſame afflictions that are *in the w.*
1 *John* 2. 15. love not the things that are *in the w.*
4. 1. many falſe prophets are gone out *into the w.*
3. antichriſt, even now already is it *in the w.*
4. becauſe he is greater than he that is *in the w.*
9. God ſent his Son *into the w.* that we might live
2 *John* 7. many deceivers are entred *into the w.*
This WORLD.
Mat. 12. 32. it ſhall not be forgiven him in *this w.*
13. 22. cares of *this w* choke the word, *Mark* 4. 19.
Luke 16. 8. for the children of *this w.* are wiſer than
20. 34. Jeſus ſaid, the children of *this w.* marry
John 8. 23. ye are of *this w.* I am not of *this w.*
9. 39. for judgment I am come into *this w.*
12. 25. he that hateth life in *this w.* ſhall keep it
31. Jeſus ſaid, now is the judgment of *this w.*
now ſhall the prince of *this w.* be caſt out
13. 1. he ſhould depart out of *this w.* to the Father
14. 30. for prince of *this w.* cometh, nothing in me

Figure 1: Bible concordance, source: Wikipedia

Examples for Concordances

- in pre-digital times
  - in the middle ages: concordances of the Bible, words in the Bible listed in alphabetical order with context
  - literature: concordances of important texts by poets such as Homer or Shakespeare
  - in libraries: concordances of book titles
- in the digital age:
  - in research, mostly in the humanities

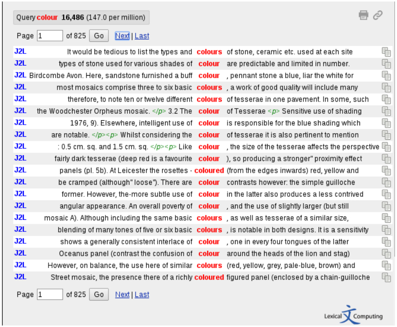


cannot induce you to attend .. *Titus Andronicus*, v. 3
INDUCED me to the semblance . *Twelfth N.* v. 1 (let.)
sir, induced by my charity. *Measure for Measure*, iv. 3
I do believe, induced by potent .... *Henry VIII.* ii. 4
I can; induced as you have been .... *Coriolanus*, i. 9
they induced to steal it! .... *Cymbeline*, ii. 4
INDUCEMENT-with his inducement. *All's Well*, iii. 2
inducement move her not to love *Richard III.* iv. 4
mark the inducement. Thus it.... *Henry VIII.* ii. 4
INDUCTION full of prosperous .... 1 *Henry IV.* iii. 1
have I laid, inductions dangerous.. *Richard III.* i. 1
a dire induction am I witness to .... — iv. 4
INDUE—indue thee with leasing.. *Twelfth Nigat*, i. 5
I shall indue you with: meantime. *King John*, iv. 2
and the tribunes indue you with .... *Coriolanus*, ii. 3

Figure 2: Shakespeare concordance, source: flickr.com

Examples for Concordances

- in pre-digital times
  - in the middle ages: concordances of the Bible, words in the Bible listed in alphabetical order with context
  - literature: concordances of important texts by poets such as Homer or Shakespeare
  - in libraries: concordances of book titles
- in the digital age:
  - in research, mostly in the humanities



Query colour 16,486 (147.0 per million)

Page 1 of 825 Go Next | Last

J2L	It would be tedious to list the types and	colours	of stone, ceramic etc. used at each site
J2L	types of stone used for various shades of	colour	are predictable and limited in number.
J2L	Birdcombe Avon. Here, sandstone furnished a buff	colour	, pennant stone a blue, liar the white for
J2L	most mosaics comprise three to six basic	colours	, a work of good quality will include many
J2L	therefore, to note ten or twelve different	colours	of tesserae in one pavement. In some, such
J2L	the Woodchester Orpheus mosaic. </p> 3.2 The	colour	of Tesserae <p> Sensitive use of shading
J2L	1976, 9). Elsewhere, intelligent use of	colour	is responsible for the blue shading which
J2L	are notable. </p><p> Whilst considering the	colour	of tesserae it is also pertinent to mention
J2L	: 0.5 cm. sq. and 1.5 cm. sq. </p><p> Like	colour	, the size of the tesserae affects the perspective
J2L	fairly dark tesserae (deep red is a favourite	colour	), so producing a stronger" proximity effect
J2L	panels (pl. 5b). At Leicester the rosettes -	coloured	(from the edges inwards) red, yellow and
J2L	be cramped (although" loose"). There are	colour	contrasts however: the simple guilloche
J2L	former. However, the-more subtle use of	colour	in the latter also produces a less contrived
J2L	angular appearance. An overall poverty of	colour	, and the use of slightly larger (but still
J2L	mosaic A). Although including the same basic	colours	, as well as tesserae of a similar size,
J2L	blending of many tones of five or six basic	colours	, is notable in both designs. It is a sensitivity
J2L	shows a generally consistent interlace of	colour	, one in every four tongues of the latter
J2L	Oceanus panel (contrast the confusion of	colour	around the heads of the lion and stag)
J2L	However, on balance, the use here of similar	colours	(red, yellow, grey, pale-blue, brown) and
J2L	Street mosaic, the presence there of a richly	coloured	figured panel (enclosed by a chain-guilloche

Page 1 of 825 Go Next | Last

Figure 3: Sketch Engine, source: Wikipedia

KWIC

How could concordances and KWIC-tables be useful for text mining?

KWIC

How could concordances and KWIC-tables be useful for text mining?

- qualitative analysis
  - evaluate if a given search term is well suited for a task or if another term should be preferred
- quantitative analysis
  - analyze the words in the surrounding of the search term

Introduction

- in the last lecture we used regex to extract keywords from texts
- geographic dispersion score based on frequencies of keywords
- logic: the more frequent a search term appears, the higher the score
- today:
  - similar approach (regex based search for a keyword)
  - data: business descriptions from 10-Ks (as in last week's lecture)
  - **validity**: to which degree does an analysis of keyword counts measure what it claims to measure?

Introduction

- in the last lecture we used regex to extract keywords from texts
- geographic dispersion score based on frequencies of keywords
- logic: the more frequent a search term appears, the higher the score
- today:
  - similar approach (regex based search for a keyword)
  - data: business descriptions from 10-Ks (as in last week's lecture)
  - **validity**: to which degree does an analysis of keyword counts measure what it claims to measure?

Today's research questions:

(1) **How do companies write about the environment in their annual reports?**

(2) **Does an "environment score" based on word counts reflect a companies environmental effort?**

Introduction

Background for today's analysis:

- A growing number of investors is concerned about environmental aspects when making an investment
- "Impact Investing" is a term that reflects this trend: *"Impact investing refers to investments that intend to generate measurable social and/or environmental impacts, as well as a financial return." (European Commission)*

Introduction

Goal for today's lecture:

- create KWIC-table for the search term "environment.*"
- get an overview of the data by reading what companies write about the "environment"
- evaluate if the search term is suited to measure the environmental effort of companies

pre	keyword	post
of Labor , the	Environmental	Protection Agency and various
We are subject to	environmental	laws and regulations in
order to comply with	environmental	laws or regulations .
of compliance with existing	environmental	laws and regulations (
and liability for known	environmental	conditions) will not
we cannot predict what	environmental	laws or regulations will
to comply with such	environmental	laws or regulations or
to respond to such	environmental	claims . Under the
in a highly competitive	environment	and face global competition

Outline of the approach

(1) Read the business descriptions into R
(2) Tokenize the business descriptions
(3) Create a KWIC-table using the `kwic()` function from `quanteda`
(4) Make a wordcloud of the words in the surrounding of "`environment.*`"
(5) Refine the approach

Elements of a Key Word in Context table

A KWIC-table usually includes these components:

(1) left context of a specified length

(2) keyword

(3) right context of a specified length

(4) document identifier

Qualitavie analysis

Occurrences of “environment” do not capture the word sense “natural environment”:

docname	pre	keyword	post
text1	of Labor , the	Environmental	Protection Agency and various
text1	We are subject to	environmental	laws and regulations in
text1	order to comply with	environmental	laws or regulations .
text1	of compliance with existing	environmental	laws and regulations (
text1	and liability for known	environmental	conditions) will not
text1	we cannot predict what	environmental	laws or regulations will
text1	to comply with such	environmental	laws or regulations or
text1	to respond to such	environmental	claims . Under the
text2	in a highly competitive	environment	and face global competition
text2	agencies influence our operating	environment	. Monetary policy conducted

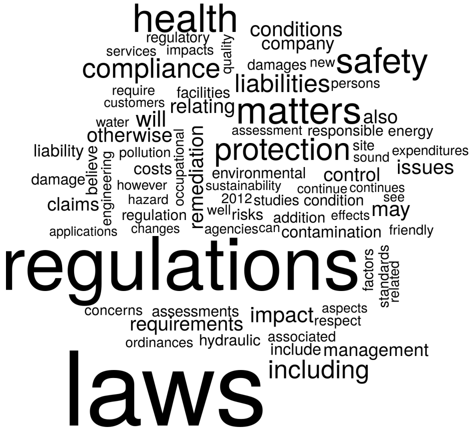


Figure 4: Wordcloud of the words in the right context

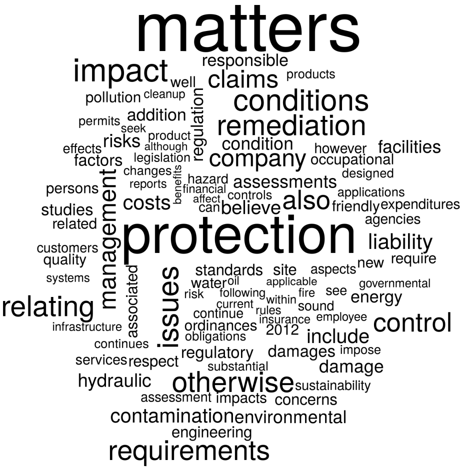


Figure 5: Wordcloud of the words in the right context, some frequent words refering to law and regulations removed

The phrase “environmental protection”

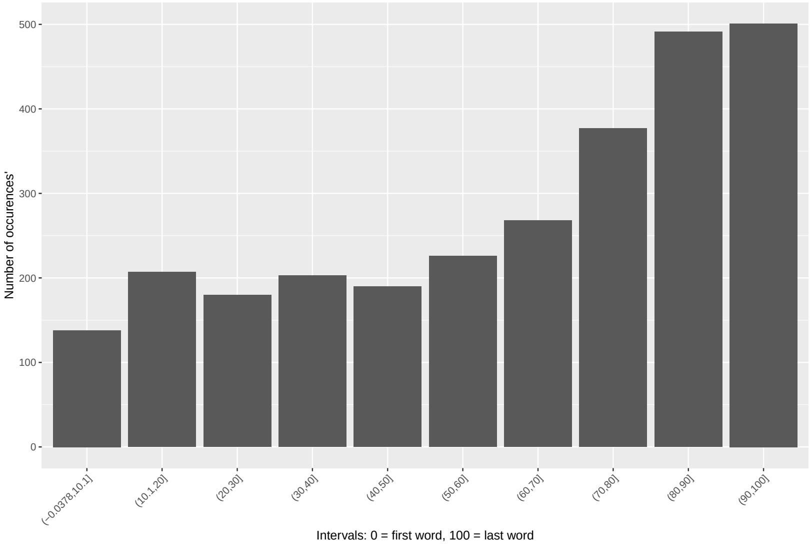
- After removing frequent words that obviously refer to laws and regulations, the word “protection” stands out as one of the most frequent ones.
- Does it refer to the companies efforts to improve environmental protection?

We generate another KWIC-table for the phrase “`environmental protection`”:

docname	pre	keyword	post
text1	of Labor , the	Environmental Protection	Agency and various other
text7	EPA , that the	Environmental Protection	Agency (" EPA
text8	. The U.S .	Environmental Protection	Agency (" EPA
text12	CERCLA also authorizes the	Environmental Protection	Agency (the "
text16	and Irritrol brands achieved	Environmental Protection	Agency (" EPA
text17	or proposed , concerning	environmental protection	or the discharge of
text22	CERCLA also authorizes the	Environmental Protection	Agency and , in
text22	. The United States	Environmental Protection	Agency (the "
text31	and / or local	environmental protection	laws . Many of
text31	, the United States	Environmental Protection	Agency (" USEPA

Where in the text does the search term “environment.*” appear?

Occurences of 'environment.*' within the documents

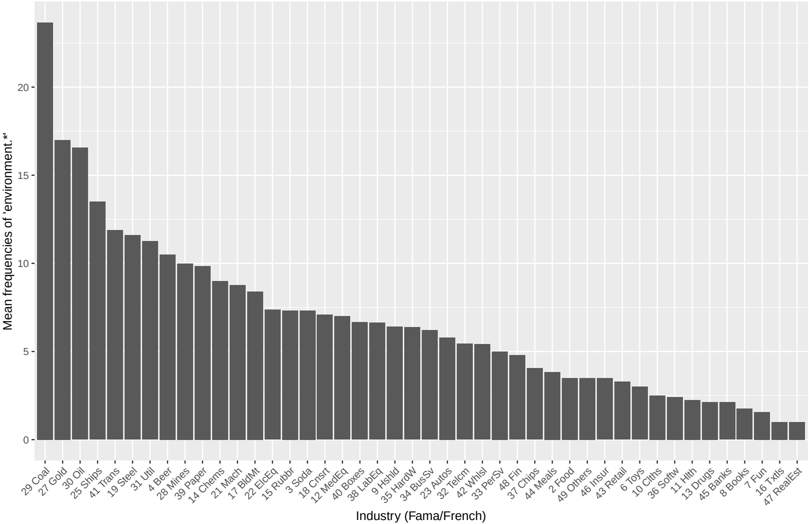


- How to interpret the plot?
- Hypothesis: “environment” is part of a phrase that contains references to law regulations.
- These phrases occur more often towards the end of a document

- Can we use the frequencies of the search term “environment.*” as a proxy for how big the environmental effort of a company is?

Aggregated mean frequencies of 'environment.*' for different industries

Data: Section 'Item 1' from 500 10−Ks



Answer to the research questions

Q1 How do companies write about the environment in their annual reports?

A1 The most frequent words in the surrounding of the keyword refer to laws and regulations. This implies that the companies' motivation to mention the environment is to meet the legal requirements.

Q2 Does an "environment score" based on word counts reflect a companies environmental effort?

A2 Because of the semantic ambiguity of the search term "environment.*", this term does not seem to be suited to quantify a companies environmental effort.

Is there an alternative term that could be used to find companies that intend to generate social or environmental impact?

Summary

- KWIC-tables as a basis for qualitative analysis of textual data:
  - convenient way of getting an overview of how a keyword is used in the data
- KWIC-tables as a basis for a quantitative analysis of textual data
  - find words that frequently appear in the surrounding of a keyword
- a score based on keyword counts can be problematic if the keyword has several meanings
- in this case the measure the criterion of *validity* is not fulfilled (we are not measuring what we claim to measure)
- the companies that write most about “environment.*” are the ones that are most affected by legal regulation of their activities
- these companies tend to mention phrases that refer to legal regulations with respect to the environment towards the end of