BAN432 Applied Textual Data Analysis for Business and Finance

Collecting textual data: crawling EDGAR

Maximilian Rohrer and Christian Langerfeld

September 13, 2022

Overview

- Collecting textual data
    - Introduction
    - API
    - Company Disclosure

Plan for this lecture

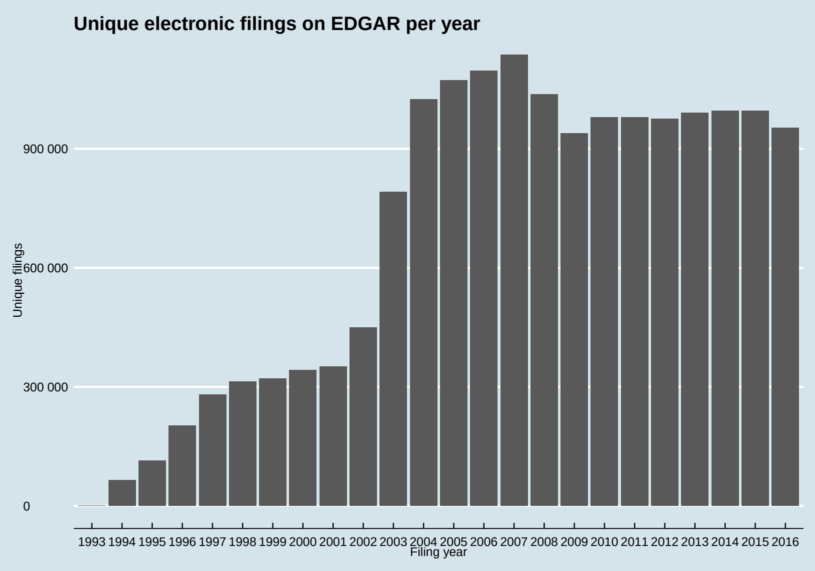
- Electronic Data Gathering, Analysis, and Retrieval (EDGAR)
  - Regulatory set-up
  - Descriptives
- Accessing and structuring EDGAR in *R*
  - Goal: writing a small crawler

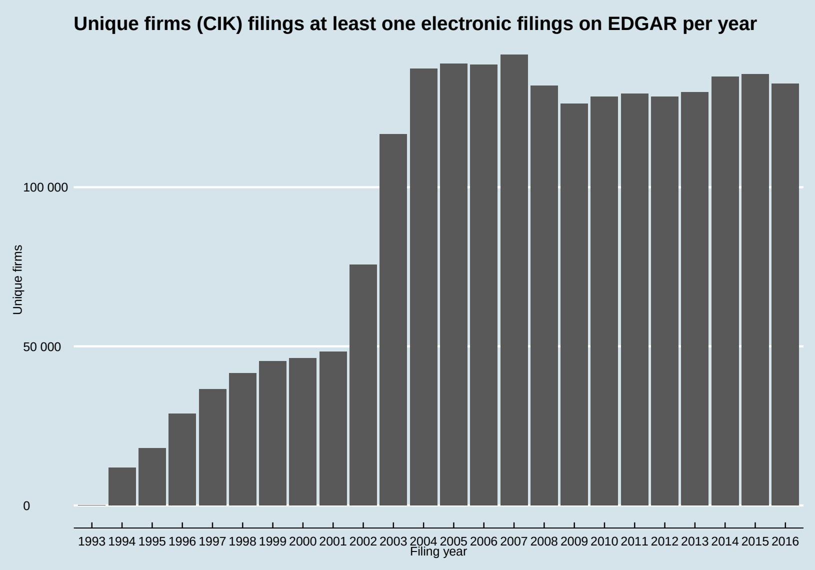
Regulatory set-up

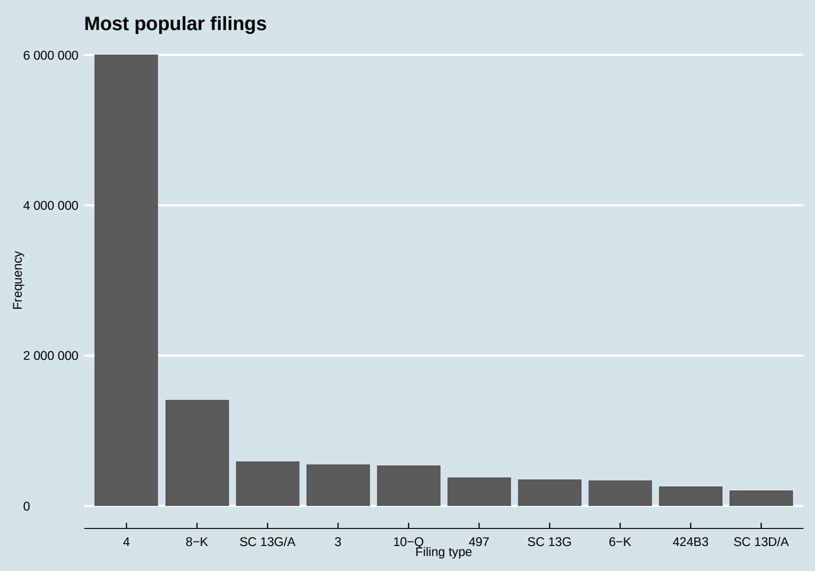
- Companies with public securities are required by law to file a number of different forms with the U.S. Securities and Exchange Commission (SEC).
- Examples are: annual reports (10-K), quarterly reports (10-Q), transaction by insiders and blockholders (Form 4), material information (8-K), etc.
- The purpose is to make information available to investors and companies, and by that improve efficiency of security markets.
- SEC developed Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system to handle electronic form filing.
- As of May 6, 1996 all public U.S. companies were required to make all their filings, with a few exceptions, on EDGAR.
- See: Garc?a, D./Norli, ?., 2012, Crawling EDGAR. The Spanish Review of Financial Economics 10. 1-10.

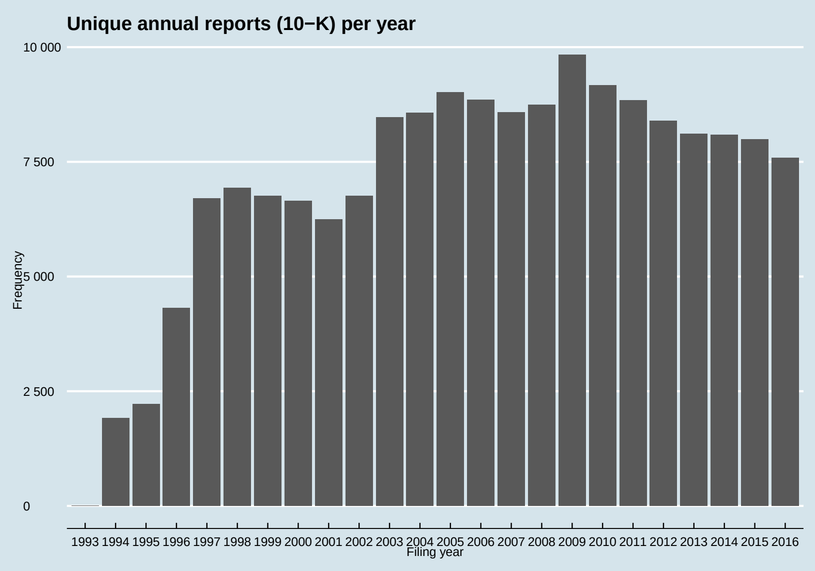
Why care?

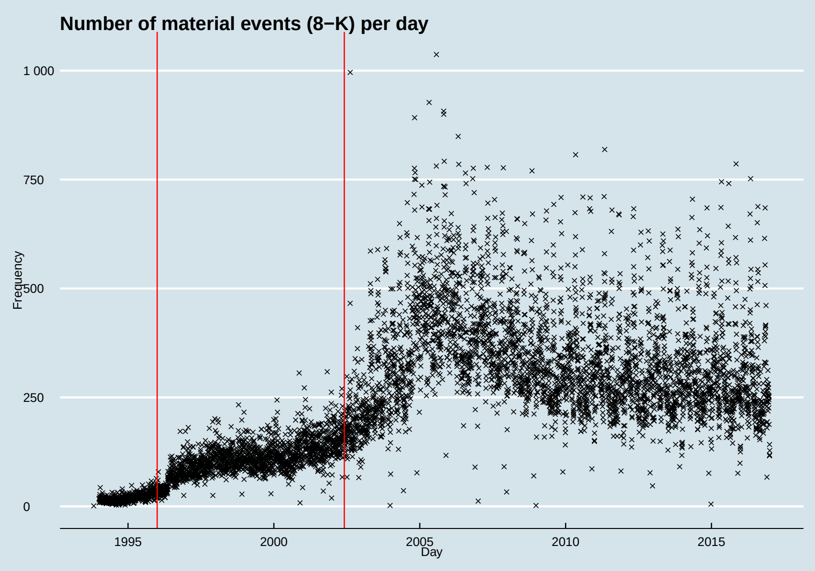
- Structured access to important filings: e.g. access to all annual reports at one place in "one" format.
- Identify different events: shareholder involvement in the annual meeting, insider transaction, etc.
- Sentiment of filings: earnings release, press releases, etc.
- Construct high level firm-characteristics: e.g. which hedging instruments does a company use? When do stock option plans for the management mature?
- Detailed information of firm events: e.g. merger prospectus describes the exact process of merger negotiations with the parties.


Unique electronic filings on EDGAR per year
Unique filings
900 000
600 000
300 000
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Filing year


Unique firms (CIK) filings at least one electronic filings on EDGAR per year
Unique firms
100 000
50 000
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Filing year


Most popular filings
6 000 000
4 000 000
2 000 000
0
Frequency
4
8−K
SC 13G/A
3
10−Q
497
SC 13G
6−K
424B3
SC 13D/A
Filing type


Unique annual reports (10−K) per year
10 000
7 500
5 000
2 500
0
Frequency
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Filing year


Number of material events (8−K) per day
1 000
750
500
250
0
Frequency
1995
2000
2005
2010
2015
Day

Descriptives: Background

- Firms for which annual reports are referred to as 10-Ks changes over time
- In mid-1996 filing became compulsory
- Sarbanes-Oxley Act of July 2002
- Regulation Fair Disclosure (Reg FD) of August 2000

Important forms

- Annual reports (10-K) and quarterly reports (10-Q)
- Changes in ownership (Form 4)
- Material events (8-K) such as press releases
- ...
- Full list https://www.sec.gov/forms

Working task: accessing Apple's most recent annual report

- Go to: https://www.sec.gov/edgar.shtml
- "Company Filings Search"
  1. When did Apple Inc. file the most recent annual report (10-K)?
  2. Open the 10-K file and investigate a bit
  3. Open the Complete submission text file, try to understand the structure

How to access EDGAR with R?

- There always seems to be a new R-package
- SEC information: https://www.sec.gov/os/accessing-edgar-data
- Index files and individual urls
    - More transferrable learning
    - Kind of always end up here anyways
- WRDS (library, but NHH not access)
- Annual reports nicely pre-coded: https://sraf.nd.edu/data/

Read EDGARs information

- Go to: https://www.sec.gov/os/accessing-edgar-data
- Read following chapters:
  - Data APIs
  - Using the EDGAR index files
  - CIK
  - Paths and directory structure

Access through the webpage

- Instead of searching for the filings of a given company individually, we can access an index file listing all filings during a given period.
- Follow the link: https://www.sec.gov/Archives/edgar/full-index/
- Download master.idx file for any given quarter (best one of the earlier years... pre 2000), and open it with any text editor (such as notepad on windows).
- Note: structure how url is constructed!
- We can access this file directly from R.

Accessing EDGAR master file with R

Constructing url to download index for filings in Q1 of 2015,

```
# define the relevant quarter
q <- 1
# define the relevant year
y <- 2015

# define web.url
web.url <- paste(
  "https://www.sec.gov/Archives/edgar/full-index/",
   y,
   "/QTR", q,
   "/master.idx", sep ="")

# check URL
print(web.url)
```

```
## [1] "https://www.sec.gov/Archives/edgar/full-index/2015/QTR1/master.
```

Accessing EDGAR master file with R

Downloading index file,

```
# Download the index file
download.file(web.url,
              destfile = paste0("EdgarIndexFileYear",
                                y,
                                "QTR",
                                q),
              headers = c("User-Agent"=
                            "YOUR_MAIL_ADRESSE@nhh.no"))
```

Accessing EDGAR master file with R

Can you detect a structure?

```r
# load the initial 100 lines
print(readLines(paste0("EdgarIndexFileYear", y, "QTR", q), n = 20))
```

```
##  [1] "Description:           Master Index of EDGAR Dissemination Fee
##  [2] "Last Data Received:    March 31, 2015"
##  [3] "Comments:              webmaster@sec.gov"
##  [4] "Anonymous FTP:         ftp://ftp.sec.gov/edgar/"
##  [5] "Cloud HTTP:            https://www.sec.gov/Archives/"
##  [6] ""
##  [7] " "
##  [8] " "
##  [9] " "
## [10] "CIK|Company Name|Form Type|Date Filed|Filename"
## [11] "-------------------------------------------------------------------
## [12] "1000032|BINCH JAMES G|4|2015-03-03|edgar/data/1000032/00012091
## [13] "1000045|NICHOLAS FINANCIAL INC|10-Q|2015-02-09|edgar/data/1000
## [14] "1000045|NICHOLAS FINANCIAL INC|8-K|2015-02-04|edgar/data/10000
## [15] "1000045|NICHOLAS FINANCIAL INC|CORRESP|2015-02-18|edgar/data/1
## [16] "1000045|NICHOLAS FINANCIAL INC|CORRESP|2015-02-27|edgar/data/1
## [17] "1000045|NICHOLAS FINANCIAL INC|SC 13G/A|2015-02-17|edgar/data/
## [18] "1000045|NICHOLAS FINANCIAL INC|SC 13G|2015-03-27|edgar/data/10
```

Working example: construct EDGAR master file R

Try to structure/load the index file correctly. Use `read.delim()` or `read_delim()` [preferred, from tidyverse package].

The final result should look like this:

```
head(edgar.index)
```

```
## # A tibble: 6 x 5
##       CIK `Company Name`     `Form Type` `Date Filed` Filename
##     <dbl> <chr>              <chr>       <date>       <chr>
## 1 1000032 BINCH JAMES G      4           2015-03-03   edgar/data/100
## 2 1000045 NICHOLAS FINANCIA~ 10-Q        2015-02-09   edgar/data/100
## 3 1000045 NICHOLAS FINANCIA~ 8-K         2015-02-04   edgar/data/100
## 4 1000045 NICHOLAS FINANCIA~ CORRESP     2015-02-18   edgar/data/100
## 5 1000045 NICHOLAS FINANCIA~ CORRESP     2015-02-27   edgar/data/100
## 6 1000045 NICHOLAS FINANCIA~ SC 13G/A    2015-02-17   edgar/data/100
```

Coding Recap

- We have downloaded a master file and structured it.
- The remainder we will talk about. . .
  - writing a small functioning crawler
  - the structure of filings (10-K in specific)

A simple crawler

Construct a crawler, that downloads all 10-Q filings of Apple (SIC = 0000320193) in the year 2008

Step 2: Structure the individual steps

1. Download index file
2. Limit to Apple and 10-Q
3. Download

Note: different options of how to iterate

Structure of downloaded EDGAR data

- Documents associated with Apple's 2016 10-K: https://www.sec.gov/Archives/edgar/data/320193/000162828016020309/0001628280-16-020309-index.htm
- XML (loaded as text) file captures all documents: https://www.sec.gov/Archives/edgar/data/320193/000162828016020309/0001628280-16-020309.txt
- Structure:
  - Header
  - All individual documents seperated by: `<DOCUMENT> ... </DOCUMENT>`
  - For each document, there is a small header, and the text, seperated by `<TEXT> ... </TEXT>`
  - ...unfortunately the structure is not absolute, especially when using older files

Structure of downloaded EDGAR data

```
download.file("https://www.sec.gov/Archives/edgar
              /data/320193/000162828016020309/0001
              628280-16-020309.txt",
              destfile = "randomEdgarFile.txt",
              headers = c("User-Agent"=
                            "YOUR_MAIL_ADRESSE@nhh.no"))

temp <- readLines("randomEdgarFile.txt",
                  encoding = "UTF-8")
```

url form previous page

Structure of downloaded EDGAR data

```
## <SEC-DOCUMENT>0001628280-16-020309.txt : 20161026
## <SEC-HEADER>0001628280-16-020309.hdr.sgml : 20161026
## <ACCEPTANCE-DATETIME>20161026164216
## ACCESSION NUMBER:        0001628280-16-020309
## CONFORMED SUBMISSION TYPE:   10-K
## PUBLIC DOCUMENT COUNT:       96
## CONFORMED PERIOD OF REPORT:  20160924
## FILED AS OF DATE:        20161026
## DATE AS OF CHANGE:       20161026
##
## FILER:
##
##  COMPANY DATA:
##      COMPANY CONFORMED NAME:         APPLE INC
##      CENTRAL INDEX KEY:          0000320193
##      STANDARD INDUSTRIAL CLASSIFICATION: ELECTRONIC COMPUTERS [3571]
##      IRS NUMBER:             942404110
##      STATE OF INCORPORATION:         CA
##      FISCAL YEAR END:            0924
##
##  FILING VALUES:
##      FORM TYPE:      10-K
##      SEC ACT:        1934 Act
```

Structure of downloaded EDGAR data

```
## <DOCUMENT>
## <TYPE>10-K
## <SEQUENCE>1
## <FILENAME>a201610-k9242016.htm
## <DESCRIPTION>10-K
## <TEXT>
## <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http
## <html>
##  <head>
##      <!-- Document created using Wdesk 1 -->
##      <!-- Copyright 2016 Workiva -->
##      <title>Document</title>
##  </head>
##  <body style="font-family:Times New Roman;font-size:10pt;">
## ...
## </TEXT>
## </DOCUMENT>
## <DOCUMENT>
## <TYPE>EX-10.18
## <SEQUENCE>2
```

Filtering tasks

- How many individual files were submitted? (“SEQUENCE”)
- What types are those files? (“TYPE”)
- What content does the file have? (“DESCRIPTION”)
- What is the file name? (“FILE NAME”)
- Where does the actual text of the document start? (“TEXT”)

Summary of this lecture

- Regulatory set-up and descriptives for EDGAR
- Structure of EDGAR data-base
- Accessing and structuring EDGAR in R
- Screening filings in R and accessing them directly
- Last lecture on collecting textual data
- Next week: preprocessing and cleaning of obtained data