R For Everyone: Basic Computational Journalism with R - Text Analysis

...

Raden Muhammad Hadi

Before We Start

Make sure you have installed the latest version of R and RStudio or just go to RStudio Cloud.

View this presentation here:

http://bit.ly/2YjRYFN

Who I am

bit.ly/radenmhadi

github.com/hadimaster65555

Raden Muhammad Hadi

- Mathematical Modeler - R&D @ Quantus Telematika Indonesia
- Mathematics Graduate - Specializing in Algebra (a.k.a Abstract Nonsense)
- R user for 6 years
- Also speak Java, Python, and Javascript (#ForcedByCompany)

Outline

- What is Computational Journalism
- Text Analysis
  - Get data from web
  - Basic pre-processing
  - Basic visualization
  - Sentiment Analysis (Lexicon Based)

What is Computational Journalism?
(Is it a … cake?)

Computational Journalism %>% glimpse()

Defined as the **application of computation** to the activities of journalism such as

- **information gathering**,
- **organization**,
- **sensemaking**,
- **communication** and
- **dissemination of news information**, while **upholding values of journalism** such as **accuracy** and **verifiability**.

- **Nick Diakopoulos in "A functional Roadmap for Innovation in Computational Journalism"**

Stanford Computational Journalism Lab

Using computational methods to uncover accountability stories that would otherwise go untold.

Stanford Computational Journalism Lab: http://cjlab.stanford.edu/

Computational Journalism - What They Do

- Text Analysis
- Data Visualization
- Filtering Algorithms
- Algorithmic Accountability and Discrimination
- Quantitative Fairness
- Randomness and Significance (p-hacking, causality)
- Network Analysis
- Knowledge Representation
- Truth and Trust (computational propaganda, fake news detection, etc)
- Privacy, Security, and Censorship

Computational Journalism vs Data Journalism

Computational Journalism is **application of computational thinking into journalism**

Data Journalism is a **workflow**

Text Analysis

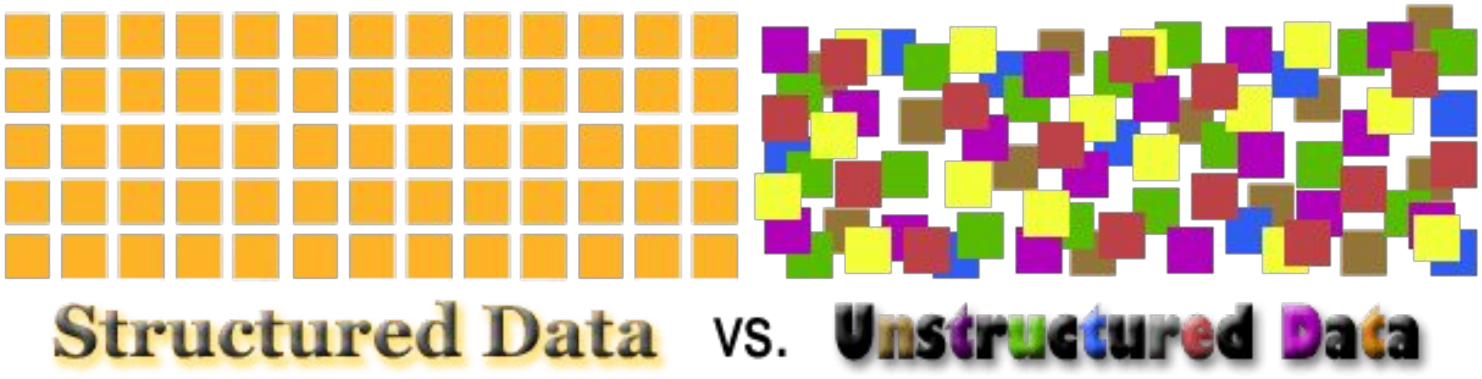
Text Analysis (also called as Text Mining)

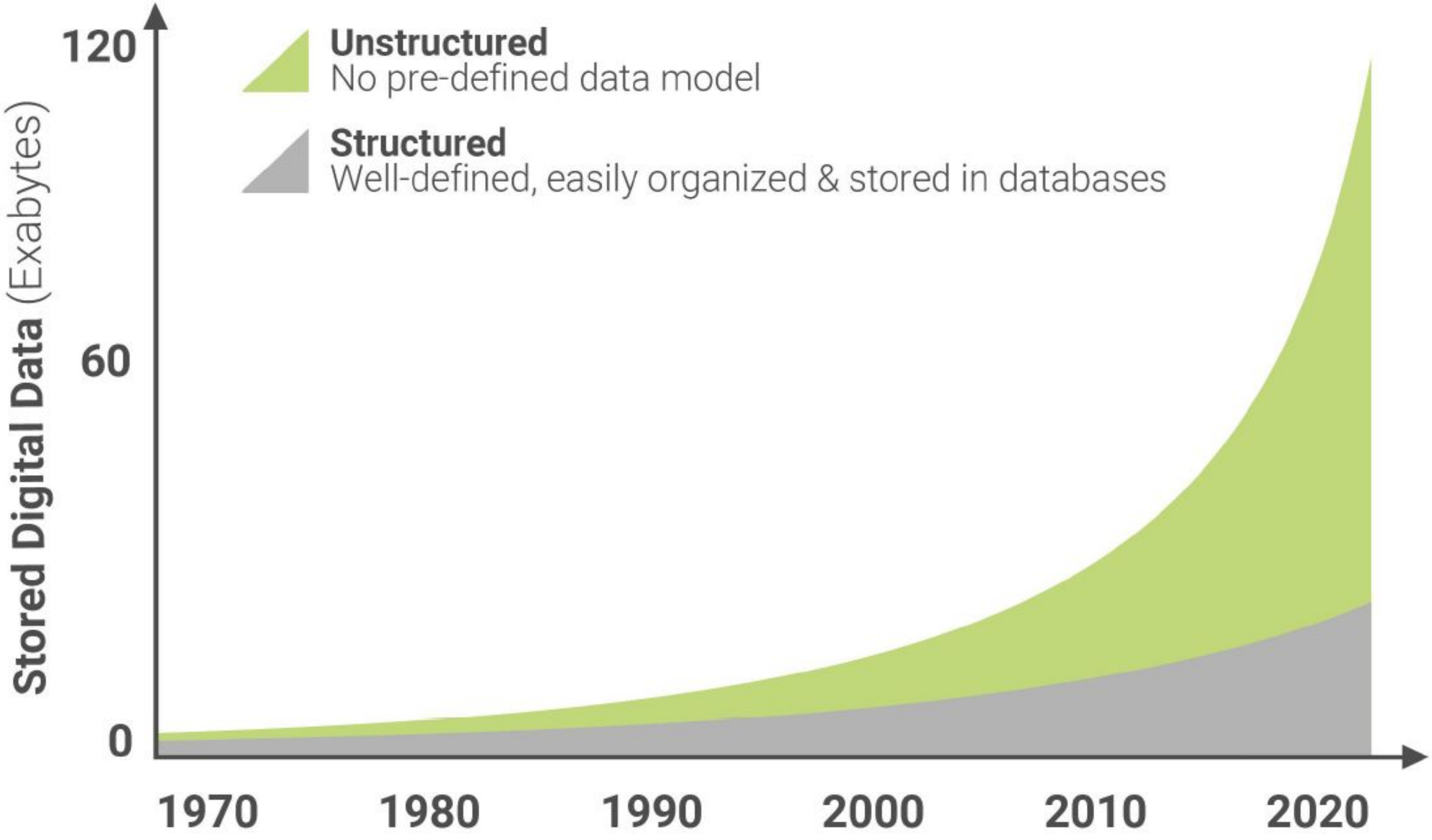
Text Analysis is about **parsing texts** in order to **extract machine-readable facts** from them. The purpose of Text Analysis is to ***create structured data out of free text content***.

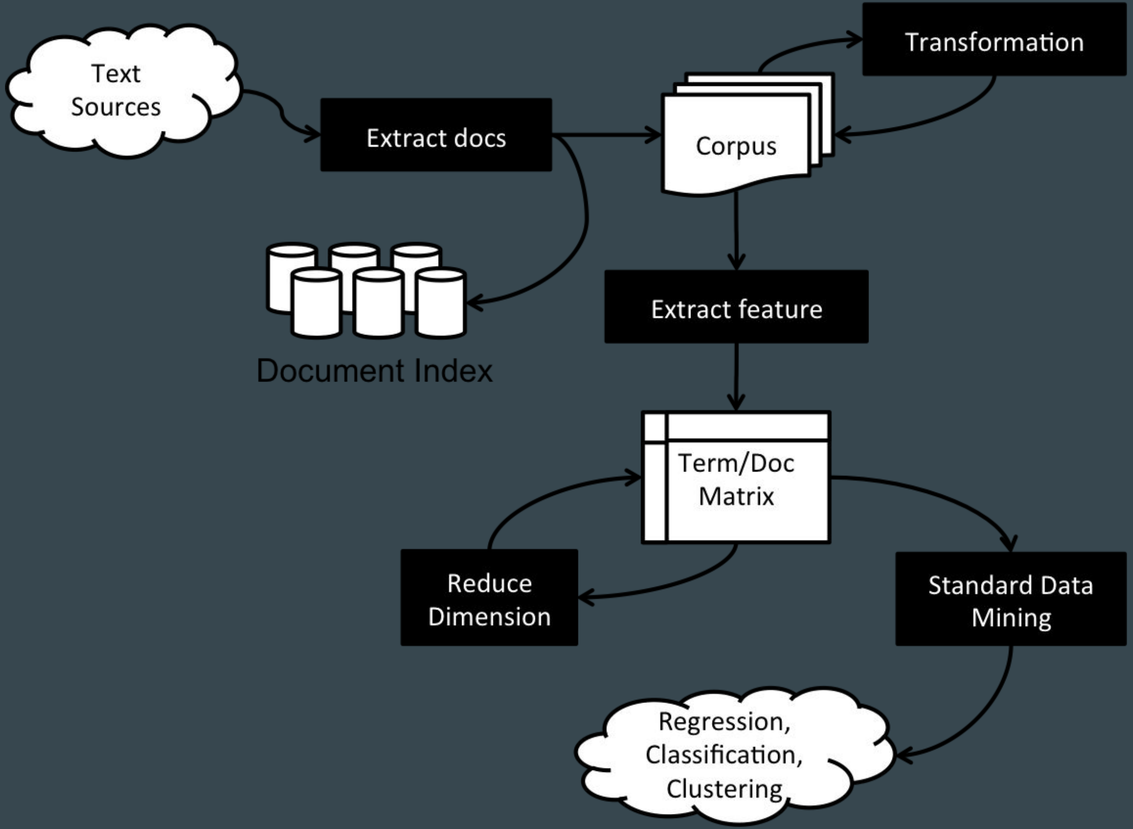
Source: https://www.ontotext.com/

Why care about Text Data?

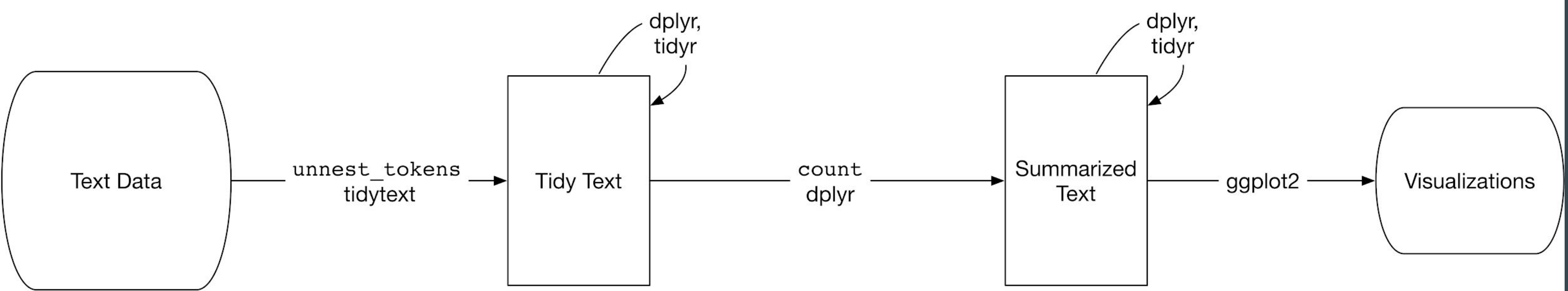
"80% of business-relevant information originates in unstructured form, primarily text."




Unstructured
No pre-defined data model
Structured
Well-defined, easily organized & stored in databases
Stored Digital Data (Exabytes)
120
60
0
1970
1980
1990
2000
2010
2020



Typical Text Mining Workflow

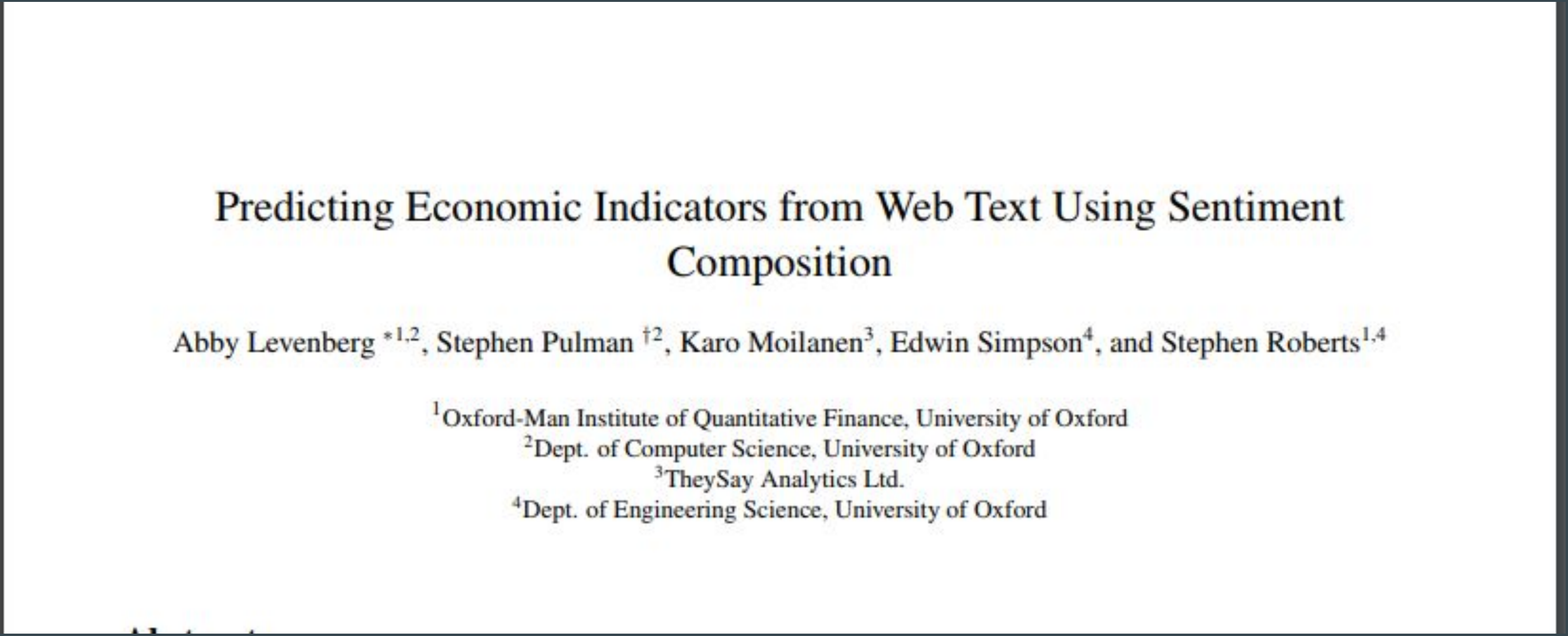


Source: https://www.tidytextmining.com/images/tidyflow-ch-1.png

Our Simple Workflow

But I’m from Econ, why I need to learn it?

A large amount of unstructured text is also generated in economic environments: company reports, policy committee deliberations, court decisions, media articles, political speeches, etc

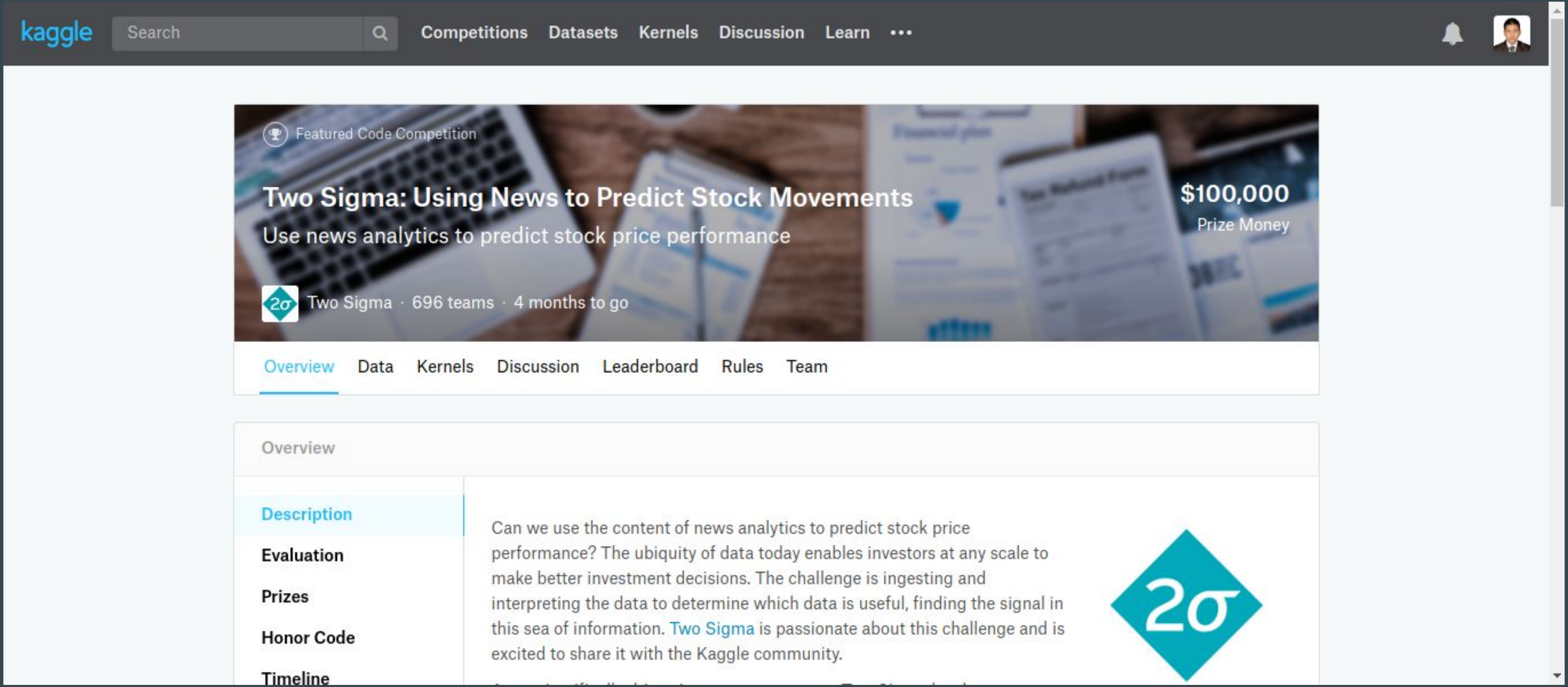


Predicting Economic Indicators from Web Text Using Sentiment Composition

Abby Levenberg [*1,2], Stephen Pulman [†2], Karo Moilanen[3], Edwin Simpson[4], and Stephen Roberts[1,4]

[1]Oxford-Man Institute of Quantitative Finance, University of Oxford
[2]Dept. of Computer Science, University of Oxford
[3]TheySay Analytics Ltd.
[4]Dept. of Engineering Science, University of Oxford

Source: http://www.robots.ox.ac.uk/~parg/pubs/sentiment_ICICA2014.pdf



Two Sigma: Using News to Predict Stock Movements

Use news analytics to predict stock price performance

$100,000 Prize Money

Two Sigma · 696 teams · 4 months to go

Can we use the content of news analytics to predict stock price performance? The ubiquity of data today enables investors at any scale to make better investment decisions. The challenge is ingesting and interpreting the data to determine which data is useful, finding the signal in this sea of information. Two Sigma is passionate about this challenge and is excited to share it with the Kaggle community.

Source: https://www.kaggle.com/c/two-sigma-financial-news

Why R?

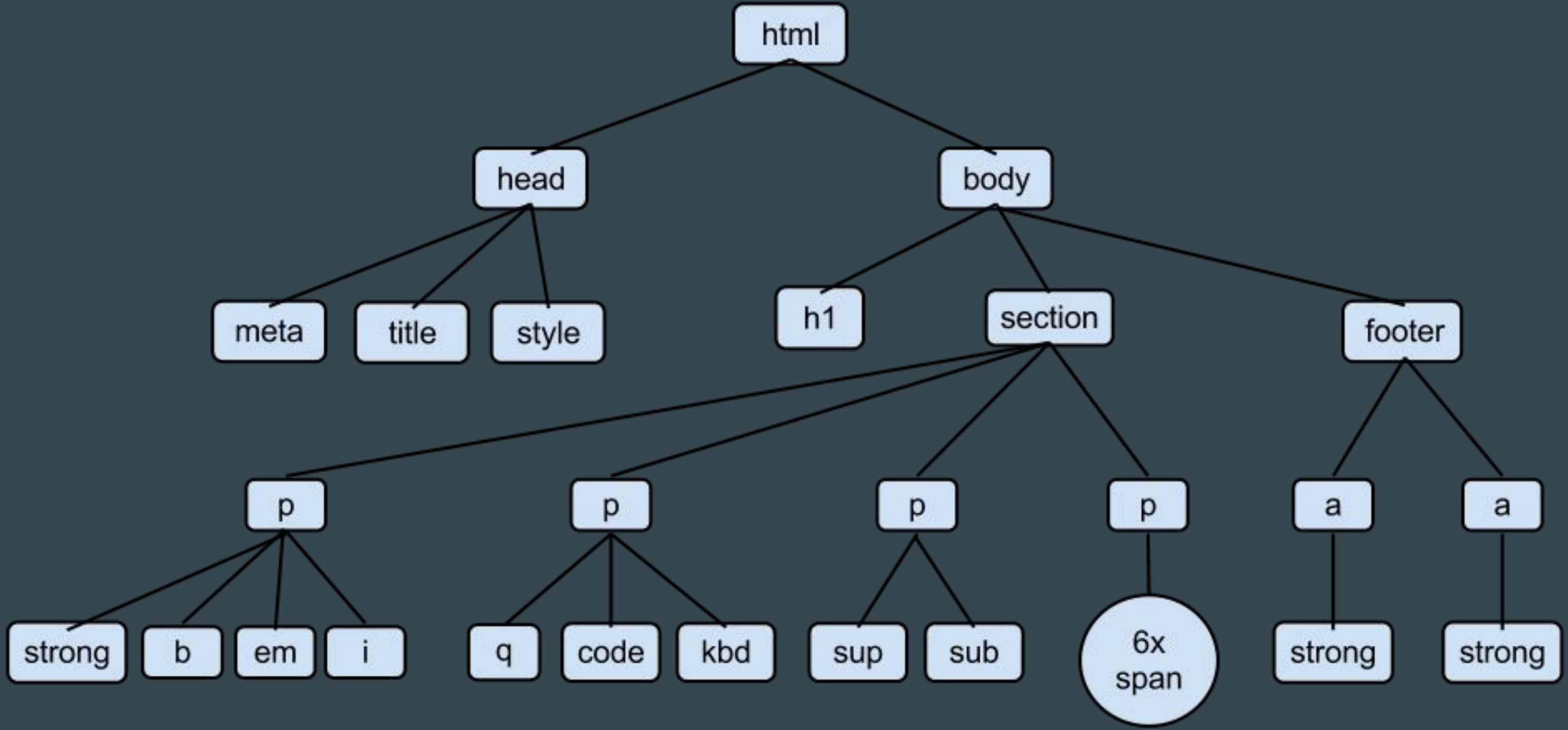
ugly graphics
low energy
costs many $$$
drop and flop
no ambitions
1.5GB storage required
SPSSCUCKS
point and click
no deeper understanding
no learning curve
ugly output
inactive community
unintuitive language
SPSS
maps
spatial data
no limits
R
community support
high energy
popularity
beautiful graphics
shiny web applications
> 10.000 packages
machine learning
R-VOLUTIONARIES
doesn't require much disk space
open source
tidyverse
Rmarkdown
gdocs integration
data mining
object-oriented programming
LaTeX compatibility
interactive visualization
code completion

Let’s Mining!

Let’s Mining!

R package you need

- rvest for web-scraping
- tidyverse for simplifying our workflow
- tidytext for text preprocessing
- hunspell for stemming
- wordcloud and wordcloud2 for wordcloud visualization



Source: http://www.openbookproject.net/tutorials/getdown/css/images/lesson4/HTMLDOMTree.png

Structure of Web Page

CSS Selector

Selector	Example	Example Description
.class	.title	Pilih semua elemen dengan class = ”title”
#id	#artikel	Pilih semua elemen dengan id = “artikel”
*	*	Pilih semua elemen
element	p	Pilih semua elemen <p>
element, element	p, div	Pilih semua elemen p dan div
[attribute]	[href=”https”]	Pilih semua elemen dengan attribute href=”https”

More about CSS Selector: https://www.w3schools.com/cssref/css_selectors.asp

HANDS-ON TIME

Go here for cheatsheet:

http://bit.ly/perbanas-textanalysis-23-03-2019

Go here for hunspell dictionary:

https://drive.google.com/drive/folders/1s637xmy__qV_ldli6kbgPGdbtUGTH-KN?usp=sharing

Thanks!

Meet me at:

- Instagram: @math_adventurer
- Linkedin: bit.ly/radenmhadi
- Github: hadimaster65555
- Personal Web: hadimaster65555.github.io

Join Indonesia useR! Community

Link:

https://t.me/GNURIndonesia