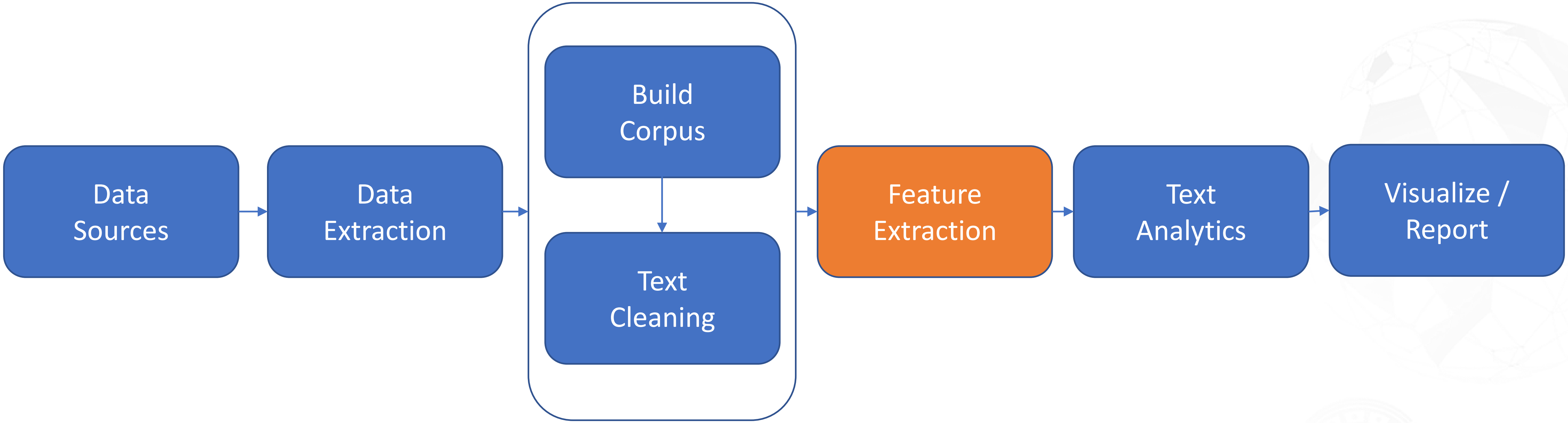
Unstructured Data Analysis

Feature Extraction

By

Kathirmani Sukumar

Text Mining Process



Data Sources

Data Extraction

Build Corpus

Text Cleaning

Feature Extraction

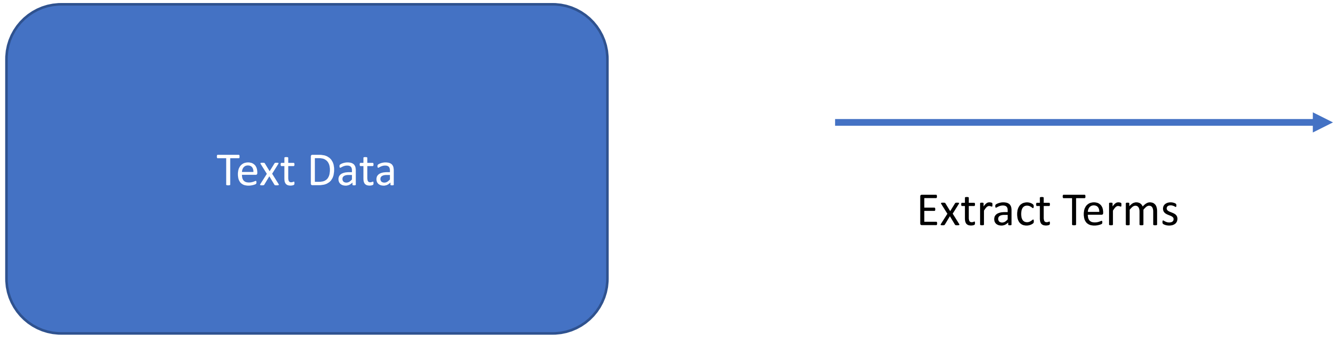
Text Analytics

Visualize / Report

Feature Extraction

- Text data is usually unstructured
- Algorithms can understand only understand numbers, vectors or matrixes
- Text data should be converted to structured data to perform useful analysis
- The process of conversion is termed as feature extraction
- Features are used to uniquely define properties of a document
- Single or group of words (which are also called as terms) are referred to as features of text document
- Theoretically even sentences or paragraphs are features

Unique terms



Terms or Tokens	Term ID
amazon	1
book	2
of	3
products	4
reviews	5
tablet	6
…..	…
…..	…
zeon	10000

Vector Representation

- Each document is represented as vector of numbers
- The i^{th} document in the corpus can be mathematically represented as vector

$$D_i = \{tw_{i1}, tw_{i2}, tw_{i3}, tw_{i4}, tw_{i5}, \ldots\ldots\ tw_{in}\}$$

- Where tw_{i1} is the frequency of the first term in the i^{th} document
- Vector dimension is equal to number of unique terms in corpus

Document Term Matrix

$$
\begin{aligned}
D_1 &= \{tw_{11}, tw_{12}, tw_{13}, tw_{14}, \ldots\ldots\ tw_{1n}\} \\
D_2 &= \{tw_{21}, tw_{22}, tw_{23}, tw_{24}, \ldots\ldots\ tw_{2n}\} \\
D_3 &= \{tw_{31}, tw_{32}, tw_{33}, tw_{34}, \ldots\ldots\ tw_{3n}\} \\
D_4 &= \{tw_{41}, tw_{42}, tw_{43}, tw_{44}, \ldots\ldots\ tw_{4n}\} \\
\ldots & \\
D_m &= \{tw_{m1}, tw_{m2}, tw_{m3}, tw_{m4}, ..\, tw_{mn}\}
\end{aligned}
$$

→

	T_1	T_2	T_3	T_4	…	T_n
D_1	5	0	…	…	…	0
D_2	0	1	…	…	…	0
D_3	0	1	…	…	…	0
D_4	0	1	…	…	…	1
…			…	…	…	1
D_m	1	2	…	…	…	10

Vector Model

Document Term Matrix (DTM)

Vector Representation

- Algorithms uses DTM to represent text documents
- Each column in DTM is a vector representation of a term
- Each row in DTM is a vector representation of a document

Document Term Matrix

	T_1	T_2	T_3	T_4	…	T_n
D_1	5	0	…	…	…	0
D_2	0	1	…	…	…	0
D_3	0	1	…	…	…	0
D_4	0	1	…	…	…	1
…			…	…	…	1
D_m	1	2	…	…	…	10

Vector representation of n^{th} term

Vector representation of m^{th} document

Characteristics of DTM

- It is usually a high dimensional matrix (m x n)
  - Ex: Amazon reviews data set with 1000 reviews when represented using DTM, the dimension was 1000 x 16542 (no. of documents x no. of unique terms)
- Most of the values in DTM will be zero, leading to a very sparse matrix
- Sparse matrix -> Mostly zeros

Term Document Matrix

- Term Document Matrix (TDM) is transpose of DTM
- In TDM, rows represents terms and columns represents documents
- Few algorithms uses TDM instead of DTM

Term Document Matrix

	D_1	D_2	D_3	D_4	…	D_m
T_1	5	0	…	…	…	1
T_2	0	1	…	…	…	2
T_3	…	…	…	…	…	…
T_4	…	…	…	…	…	…
…			…	…	…	…
T_n	0	1	…	…	…	10

DTM vs TDM

Entity	Document Term Matrix	Term Document Matrix
Row-wise	Every row represents a document	Every row represents a term
Column-wise	Every column represents a term	Every column represents a document
Dimensions	No. of docs vs No. of unique terms (mxn)	No. of unique terms vs No. of docs (nxm)
Applications	Text classification Document clustering, Document similarity, Topic Modelling	Word clustering, word similarity

Terms Frequency

Document Term Matrix

	T_1	T_2	T_3	T_4	…	T_n
D_1	5	0	…	…	…	0
D_2	0	1	…	…	…	0
D_3	0	1	…	…	…	0
D_4	0	1	…	…	…	1
…			…	…	…	0
D_m	1	2	…	…	…	10

Column sum = Term Frequency

Ex: Frequency of Tn = 11

Document Length

Document Term Matrix

	T_1	T_2	T_3	T_4	…	T_n
D_1	5	0	…	…	…	0
D_2	0	1	…	…	…	0
D_3	0	1	…	…	…	0
D_4	0	1	…	…	…	1
…			…	…	…	0
D_m	1	2	…	…	…	10

Row sum = Document Length

Rank Table

	T_1	T_2	T_3	T_4	…	T_n
D_1	5	0	…	…	…	0
D_2	0	1	…	…	…	0
D_3	0	1	…	…	…	0
D_4	0	1	…	…	…	1
…			…	…	…	0
D_m	1	2	…	…	…	10

Document Term Matrix

Column-wise sum & order Terms

Terms	Freq.	Rank
wall	2000	1
unit	1000	2
products	580	3
nook	400	4
reviews	300	5
…	…	…
Camera	10	10000

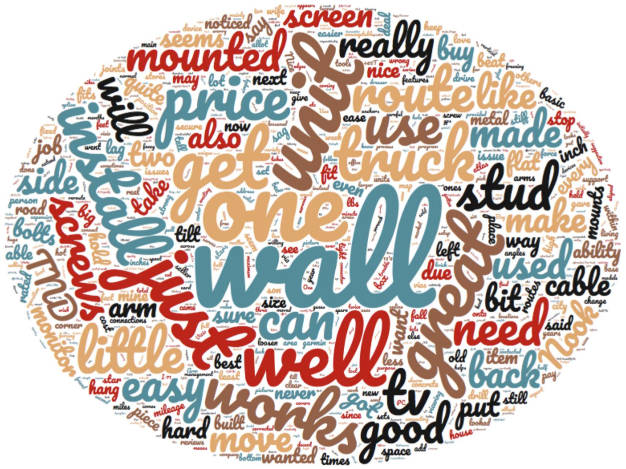
Rank/Frequency Table

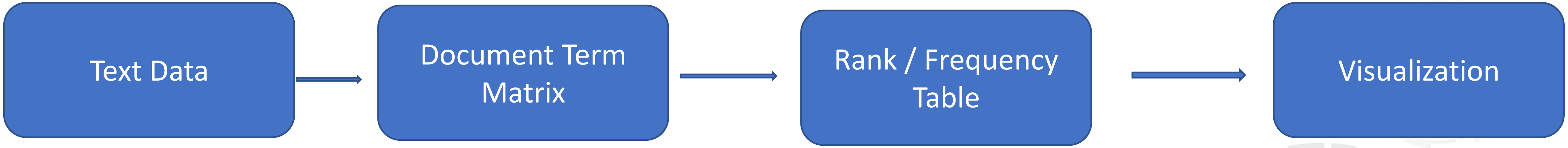
Bag of word analysis

Text Data

	T_1	T_2	T_3	T_4	…	T_n
D_1	5	0	…	…	…	0
D_2	0	1	…	…	…	0
D_3	0	1	…	…	…	0
D_4	0	1	…	…	…	1

Terms	Freq.
wall	200
unit	100
products	58
nook	40
reviews	30





Text Data → Document Term Matrix → Rank / Frequency Table → Visualization

Wordcloud

wall
one
get
price
install
unit
route
like
truck
use
made
stud
great
just
well
can
works
good
tv
need
back
little
easy
screws
mounted
screen
really
buy
will
also
two
side
bolts
able
monitor
move
hard
hang
piece
built
wanted
times
space
put
still
item
cable
bit
used
way
left
due
make
mounts
ability
flat
inch
every
stop
metal
nice
next
lot
may
quite
seems
noticed
say
fits
job
take
road
person
big
hold
tilt
see
size
sure
arm
best
never
got
since
less
want
fall
old
said
Nook
routes
place
ease
now
sag
fit
even
lag
mine
star
mileage
bottom
house
drill
pay
deal
keep
love
beat
drive
others
basic
features
wrong
tools
support
held
gave
living
within
base
change
city
years
half
new
loosen
area
garmin
connections
corner
came
inches
done
feet
fixed
months
give
hole
screw
units
map
ones
arms
angles
issue
stiff
force
lbs
free
box
clear
least
wife
main
rated

- **Font size**: Based on frequency
- **Colors**: Generally random. Can be used to represent some category
- **Layout:** Generally random. Can be defined to represent any object

Problems with Unigrams

- Unigram – One word per term
- Contexts of the words might get lost
- Case 1:
  - “Good” – 100 , “Not” – 80
  - Actually in corpus it was ”Not Good”
  - Meaning changes. Its actually more negative
- Case 2:
  - “Wall” – 200, “mounted” – 170, “issues” - 180
  - It is actually ”Wall mounting issues”
  - Redundant features

Sample Frequency Table

Term	Freq.
Good	100
not	80
Wall	200
Mounting	170
issues	180

N - Grams

- **Unigrams** – One word per term
  - Example: amazon, nook, tablet, tocuch, screen etc
- **Bigrams** – Two words per term
  - Not good, nook tablet, touch screen etc
- **Trigrams** – Three words per term
  - Wall mounting issues, bad touch screen, very good service
- **N-Grams** – N words per term

Example: N-Grams

"The customer service is really good. But the waiting is not that good. I like reading books in nook tablet. Touch screen is really good"

N - Grams	Unique Terms
Unigrams	The, customer, service, is, really, good, but, waiting, not, that, I, like, reading, books, in, nook, tablet, touch, screen
Bigrams	The customer, customer service, service is, is really, really good, good but, but the, the waiting, waiting is, is not, not that, that good, …..
Trigrams	The customer service, customer service is, service is really, is really good, really good but, good but the, but the waiting, …