16 February 2018 14:01

Text mining / Text Analytics

Data in the form of Numbers, Characters, Special characters, white space etc

→ Text mining is a process of deriving high quality information from txt.
Used in Analytics.

→ Converting data in to Tidy data

→ Each Variable is a column

→ Each observation is a Row

→ Each type of observation unit is a table

India is my Country!!! → Unstructured

↓ Transform

india | is | my | country | !!! → india | my | country → Structured

Chapt	page	chr
1	1	india
1	1	my
1	2	Country
1	2	

and, or, not, is, the, a → filtered

Tidy text is a table with One token per row

Tokenization is a process of splitting text in to tokens.

Tokens can be → Words → | India | is | my | country | india | is | great |

→ Phrases / sentence → | india is my country | india is great |

Text is often stored in below 3 types

1) Strings → Text can be a character vector. → Read in to Memory

with Meta data (Description about data)

2) Corpus → Raw strings with Metadata (Description about data)

3) DTM → Document Text Matrix → Sparse matrix → Rows x Columns.

Steps → Unnest_tokens:

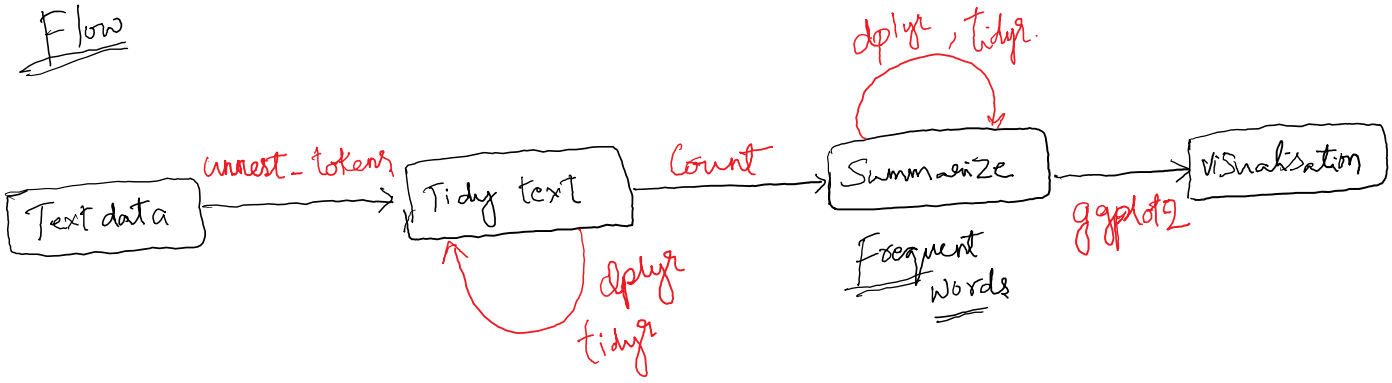
→ Convert the data in to a data-frame → tibble dataframe

→ Split each sentence in to Tokens → Assign each to a row.

I am Good! → i/am/good → i, am, good

- → Punctuations are removed
- → Converted to lower case

Flow



Tokens → It is a meaningful unit of text, most often a word, that we are interested in using for further analysis.

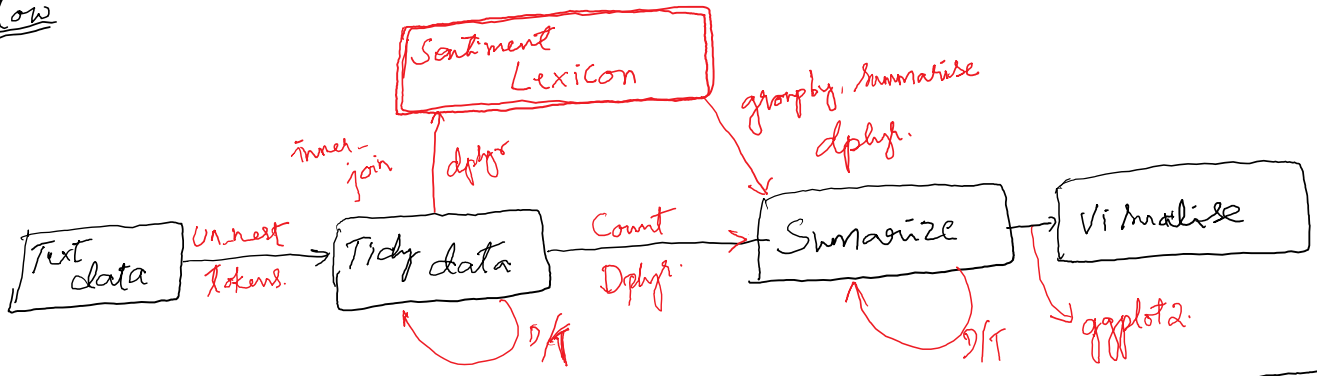
Word frequency → Most Common task in Text Analysis

→ Which word occurs the most in a given text.

Sentimental Analysis → Process of Computationally identifying & Categorizing opinions expressed in a piece of text, to determine the authors attitude towards particular topic, product etc is Positive/Negative.

Flow

Flow



Lexicon → like dictionaries. → Words are associated with Subjects

In R

get_sentiment()

↳ Sentiment of Lexicon

inner joint

in sentimental analysis

Explained in code

	Dictionary	Lexicon
Happy	Meaning	positive
Sad	—	Negative
Angry	—	Negative
Cherish	—	Positive
good	—	positive
Word	English Meaning	Sentiment Associated.

Lexicons → AFINN from Arup Nielsen

Bing from Bing Lin

NRC from Saif Mohammed & Peter Turney

Compared in R Code

Visualisation → Frequent words. → Most frequently occurring words

→ Sentiment score → Word variation along the trajectory of the story

→ Word cloud → Frequent words appear big in the cloud

→ Sentiment word cloud → Combines sentiment scale & word frequency.

We can easily identify the most important +ve/-ve words & use

Item in any of the further Analysis