# MODULE 18

1. **Amazon & Snapdeal**

**Package used in Python** :

**Package used in R**

Rvest : used to Scrape data from webpage.

XML : Generate and access the XML from the webpage

Magrittr: used for Pipeline

tm: Text Cleaning

Wordcloud : used to plot the words as a image

# **Amazon - Kent Water Purifier**

* Extracted 1500 Reviews from the Amazon
* Converted the data into Corpus and format to UTF8
* Cleaned the reviews by removing the numbers, Punctuation, stop words , white space and unwanted words from the reviews.
* Converted the list into Term Document Matrix
* Selected words repeated above 5
* Selecting only the positive words from the matrix and plotting the Positive words from the reviews
* Selecting only the Negative words from the matrix and plotting the Negative words from the reviews

## Positive Words



Negative Words

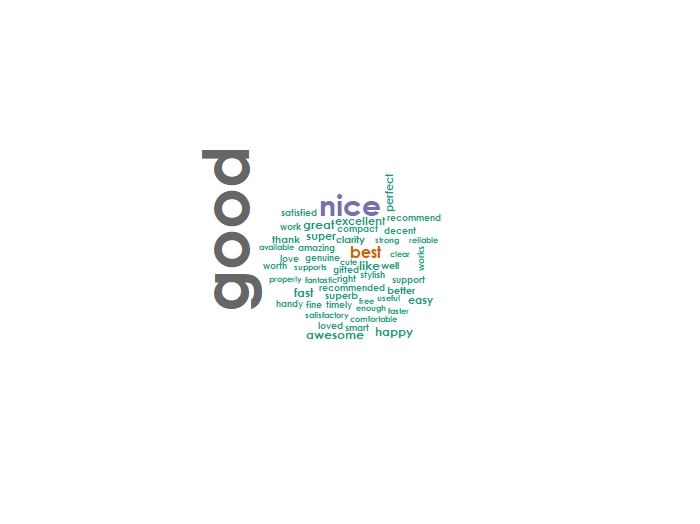


# **Snapdeal - Samsung Guru Music 2**



* Extracted 900 Reviews from the Snapdeal
* Converted the data into Corpus and format to UTF8
* Cleaned the reviews by removing the numbers, Punctuation, stop words , white space and unwanted words from the reviews.
* Converted the list into Term Document Matrix
* Selected words repeated above 2
* Selecting only the positive words from the matrix and plotting the Positive words from the reviews
* Selecting only the Negative words from the matrix and plotting the Negative words from the reviews

Positive Words



Negative Words

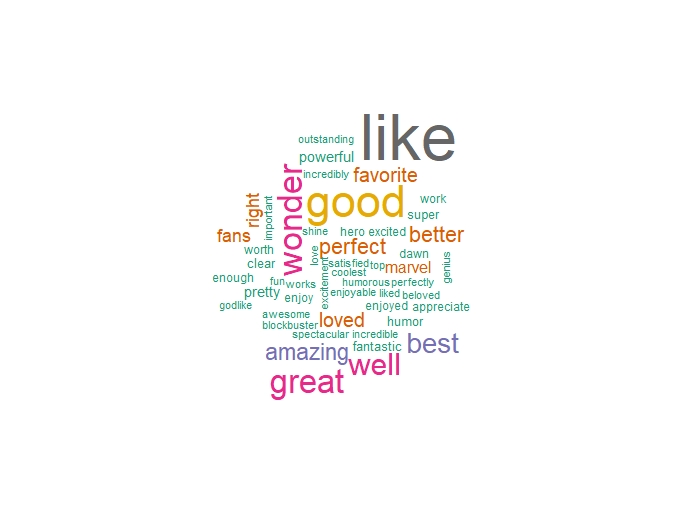


## IMDB - Batman v Superman: Dawn of Justice

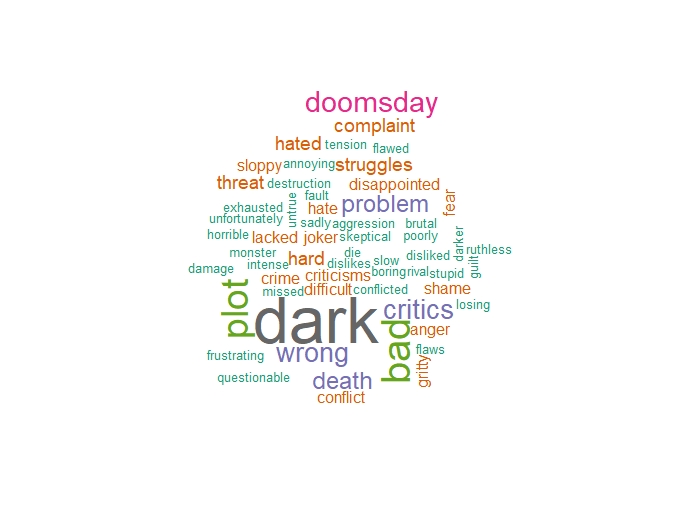


* Extracted 3500 Reviews from the IMDB
* Converted the data into Corpus and format to UTF8
* Cleaned the reviews by removing the numbers, Punctuation, stop words , white space and unwanted words from the reviews.
* Converted the list into Term Document Matrix
* Selected words repeated above 200
* Selecting only the positive words from the matrix and plotting the Positive words from the reviews
* Selecting only the Negative words from the matrix and plotting the Negative words from the reviews

Positive Words from the Reviews



Negative Words from the Reviews



**US News Headlines from Fox News and NY Times (Politics Section)**

**Package Used in R**

Rvest : used to Scrape data from webpage.

XML : Generate and access the XML from the webpage

Magrittr: used for Pipeline

Dplyr : Data Manipulation

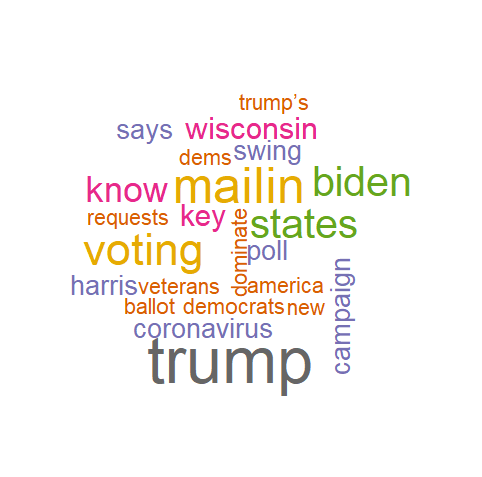
tm: Text Cleaning

Wordcloud : used to plot the words as a image

**US Presidential Election 2020**

* Extracted 61 Headlines from the Fox News and NY Times
* Converted the data into Corpus and format to UTF8
* Cleaned the reviews by removing the numbers, Punctuation, stop words , white space and unwanted words from the reviews.
* Converted the list into Term Document Matrix
* Selected words repeated above 0 times ( Used all the words due to less content)
* Selecting all the words from the matrix and plotting the words from the Headlines
* Selecting only the Negative words from the matrix and plotting the Negative words from the reviews

Important Words From the Headlines



Negative Words from the Headlines

