EVALUESERVE
POWERED BY MIND+MACHINE

भारतीय प्रौद्योगिकी संस्थान कानपुर
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
MBA | IIT KANPUR

Identification of regional distribution of topics and key influencers using Twitter data on smart cities

Capstone Project Presentation

Kinsuk Ghatak | 15125020 | MBA ,2017

Arpit Jain | 15125009 | MBA ,2017

Company details and problem statement

EVALUESERVE

POWERED BY MIND+MACHINE

- A leading analytics consulting firm
- Works in marketing ,sales, operations ,pharma ,manufacturing and financial domains
- Powered by ***"Mind + Machine"***

Problem statement & Deliverables

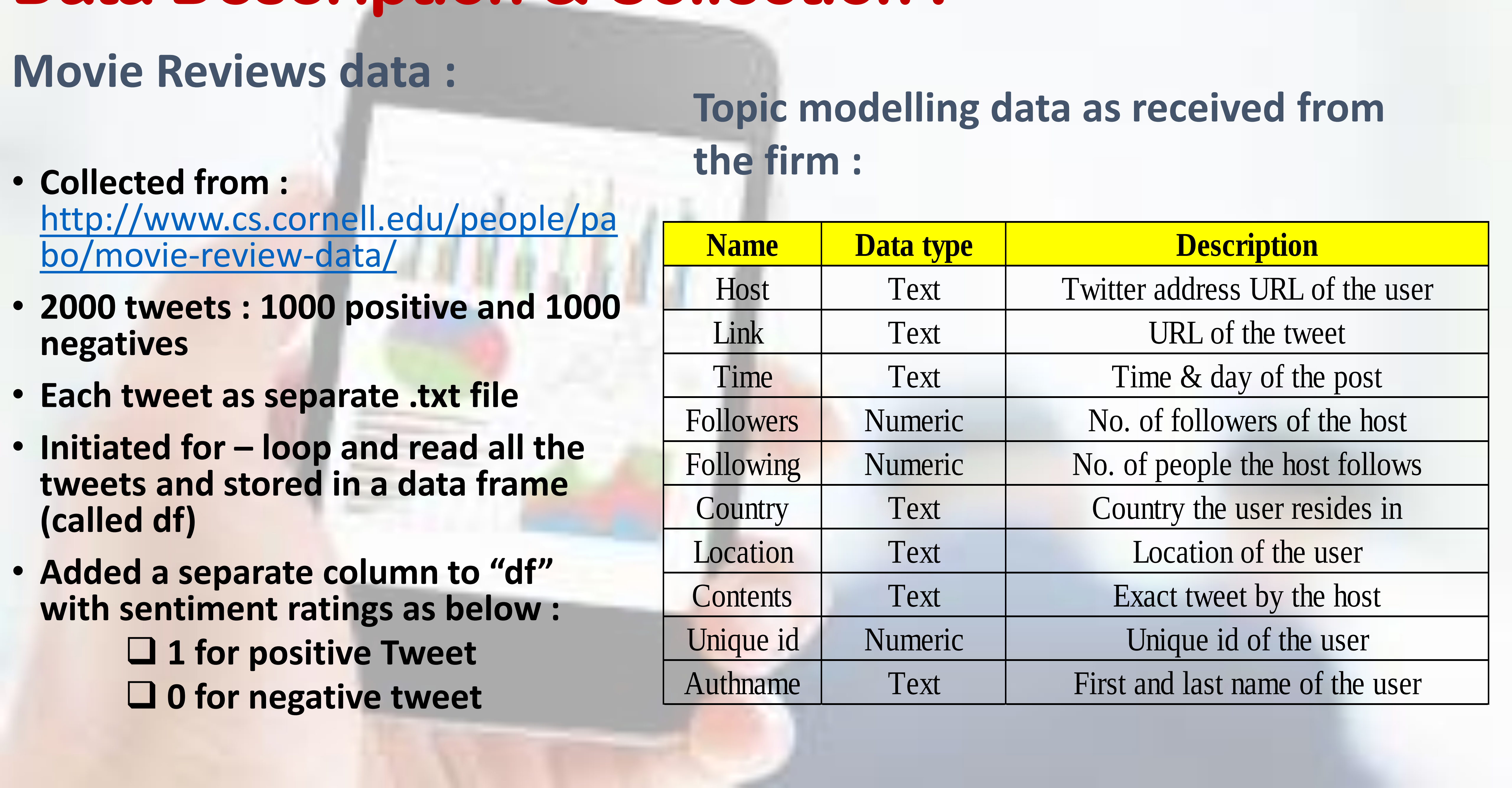
Sentiment analysis and Topic modelling of Twitter data on movie reviews and smart cities

- Pre-processing of the data
- Topic modelling of the twitter database on smart cities to
- Application of unsupervised algorithm to find out the topics. Finding the geographical distribution of the topics
- Making business sense of the topics and labelling of the topics
- Identification of key influencers on twitter considering the given data
- Development of codes ,algorithms ,visualization and reporting

Content :

- Description about Data
- Project Methodology
- Data cleaning & Pre Processing
- Description of algorithm used
- Application of algorithms & Output
- Geographic distribution of topics
- Identification of key influencers
- Business economics and benefits
- Conclusion and future scopes

Data Description & Collection :



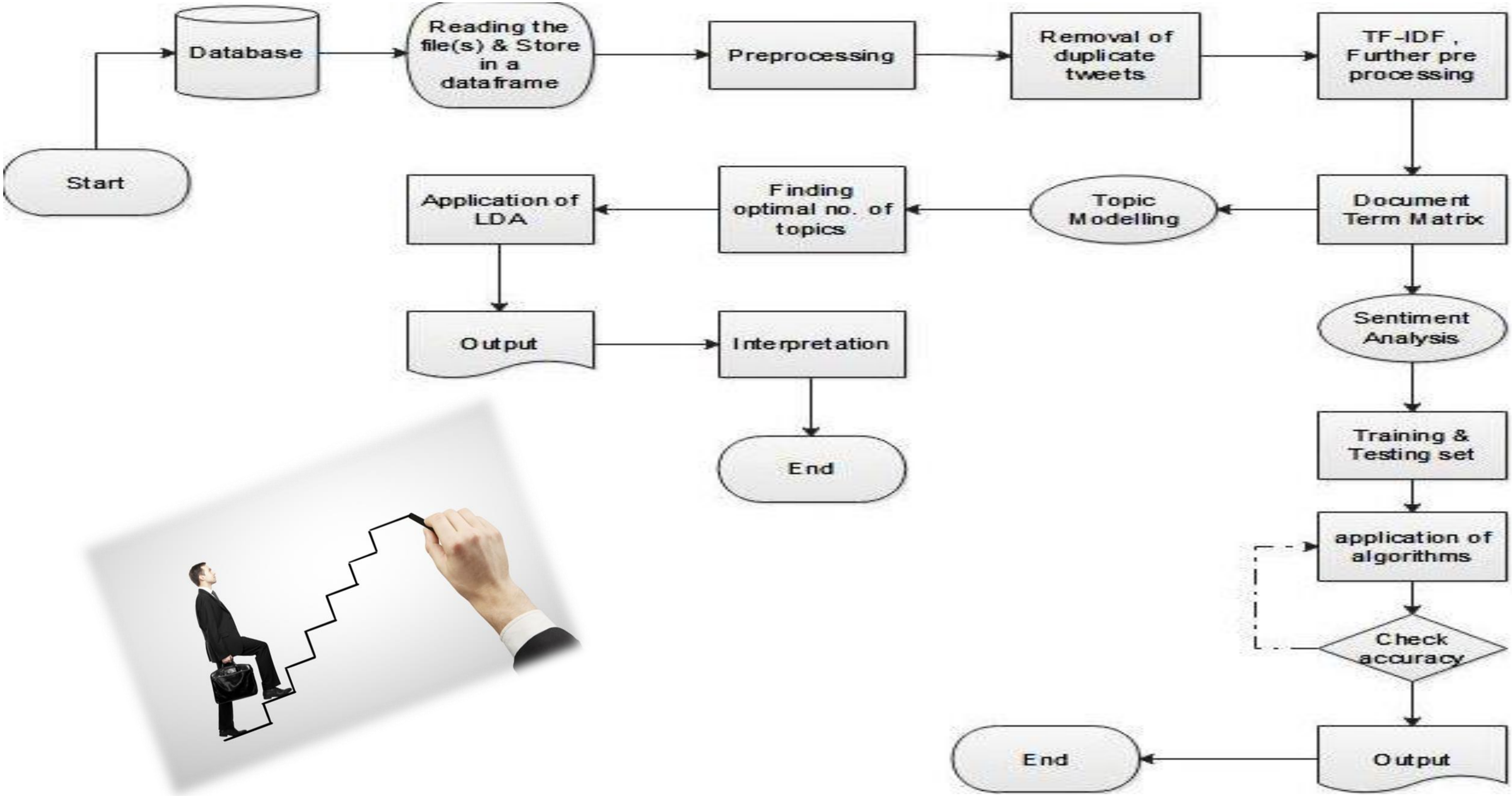
Movie Reviews data :

- **Collected from :** http://www.cs.cornell.edu/people/pabo/movie-review-data/
- **2000 tweets : 1000 positive and 1000 negatives**
- **Each tweet as separate .txt file**
- **Initiated for – loop and read all the tweets and stored in a data frame (called df)**
- **Added a separate column to "df" with sentiment ratings as below :**
  - **□ 1 for positive Tweet**
  - **□ 0 for negative tweet**

Topic modelling data as received from the firm :

Name	Data type	Description
Host	Text	Twitter address URL of the user
Link	Text	URL of the tweet
Time	Text	Time & day of the post
Followers	Numeric	No. of followers of the host
Following	Numeric	No. of people the host follows
Country	Text	Country the user resides in
Location	Text	Location of the user
Contents	Text	Exact tweet by the host
Unique id	Numeric	Unique id of the user
Authname	Text	First and last name of the user

Project Methodology followed :



Start

Database

Reading the file(s) & Store in a dataframe

Preprocessing

Removal of duplicate tweets

TF-IDF , Further pre processing

Document Term Matrix

Topic Modelling

Finding optimal no. of topics

Application of LDA

Output

Interpretation

End

Sentiment Analysis

Training & Testing set

application of algorithms

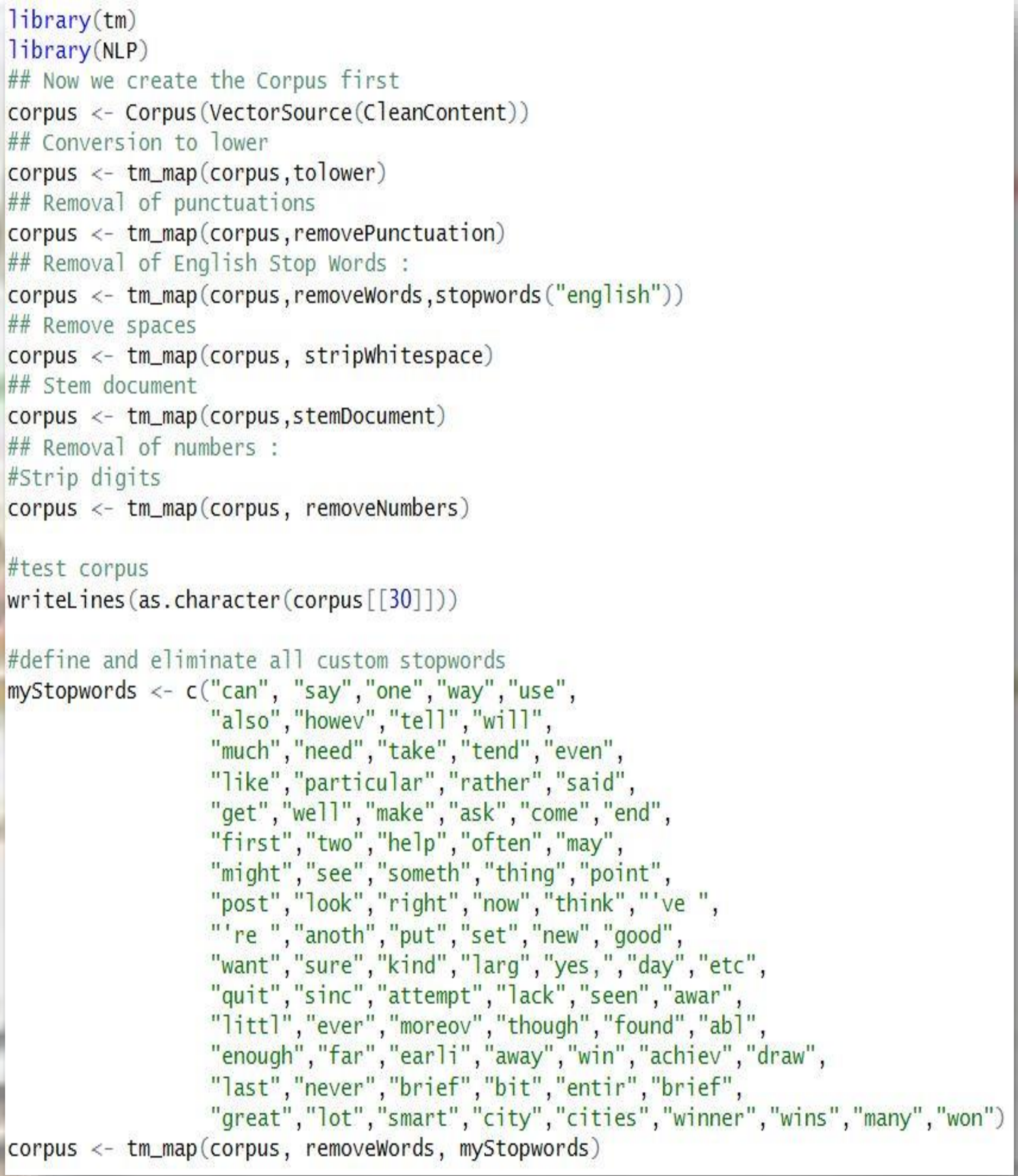
Check accuracy

Output

End

Cleaning & Pre processing :

- ❑ Package used : tm
- ❑ User defined function
- ❑ Removed repetitive tweets but kept re tweets (considering same string from different users)
- ❑ This reduced the dataset significantly and made the computation faster
- ❑ Defined and redefined stop words list using our understanding of output and senses.



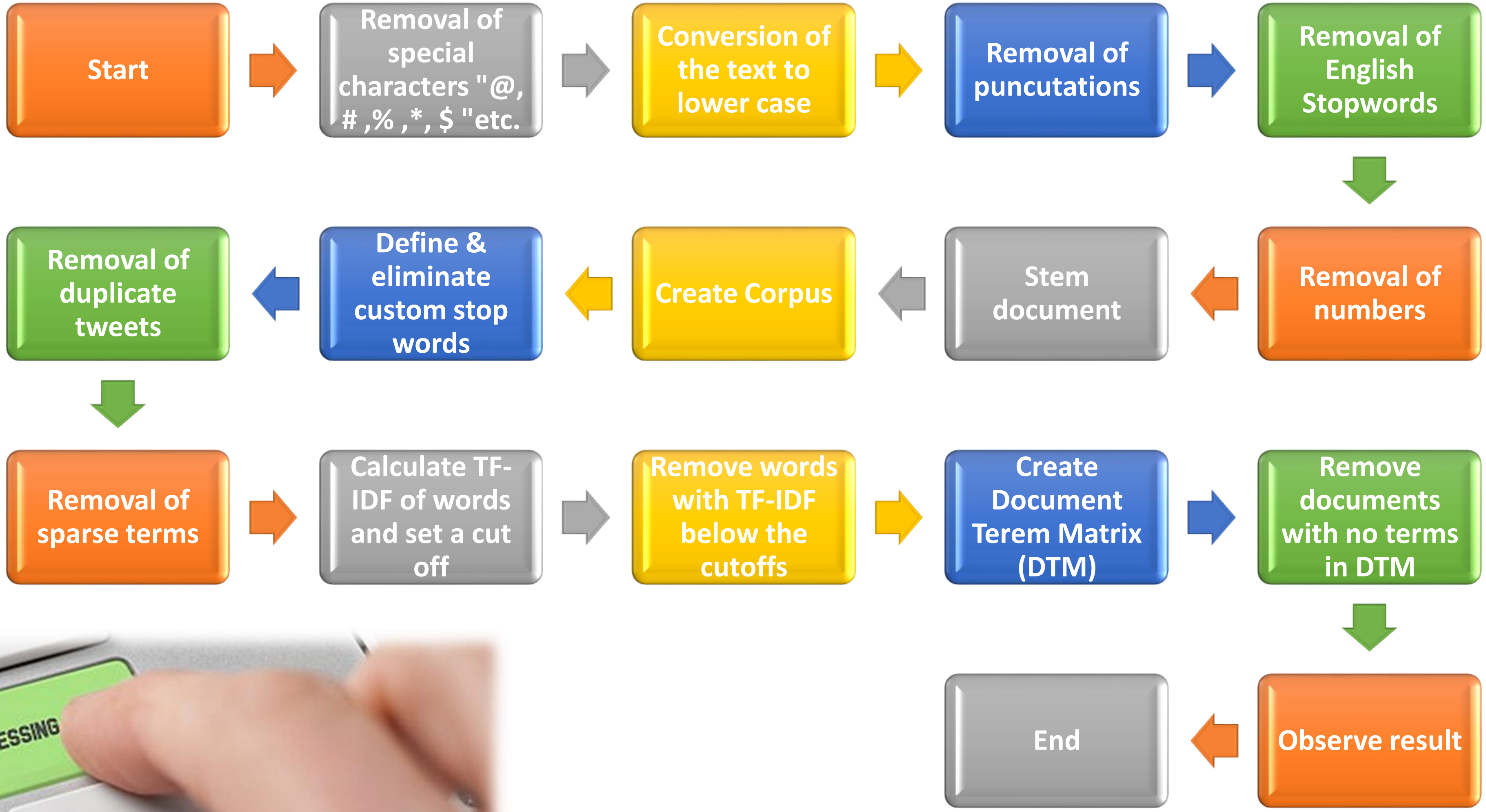
```r
library(tm)
library(NLP)
## Now we create the Corpus first
corpus <- Corpus(VectorSource(CleanContent))
## Conversion to lower
corpus <- tm_map(corpus,tolower)
## Removal of punctuations
corpus <- tm_map(corpus,removePunctuation)
## Removal of English Stop Words :
corpus <- tm_map(corpus,removeWords,stopwords("english"))
## Remove spaces
corpus <- tm_map(corpus, stripWhitespace)
## Stem document
corpus <- tm_map(corpus,stemDocument)
## Removal of numbers :
#Strip digits
corpus <- tm_map(corpus, removeNumbers)

#test corpus
writeLines(as.character(corpus[[30]]))

#define and eliminate all custom stopwords
myStopwords <- c("can", "say","one","way","use",
                 "also","howev","tell","will",
                 "much","need","take","tend","even",
                 "like","particular","rather","said",
                 "get","well","make","ask","come","end",
                 "first","two","help","often","may",
                 "might","see","someth","thing","point",
                 "post","look","right","now","think","'ve ",
                 "'re ","anoth","put","set","new","good",
                 "want","sure","kind","larg","yes,","day","etc",
                 "quit","sinc","attempt","lack","seen","awar",
                 "littl","ever","moreov","though","found","abl",
                 "enough","far","earli","away","win","achiev","draw",
                 "last","never","brief","bit","entir","brief",
                 "great","lot","smart","city","cities","winner","wins","many","won")
corpus <- tm_map(corpus, removeWords, myStopwords)
```

Cleaning & Pre processing steps

- Defined stop words apart from the regular ones
- Deleted all the stop words
- Prepared corpus
- Created Document Term Matrix by omitting words occurring less than 0.5%



Start

Removal of special characters "@, # ,% ,*, $ "etc.

Conversion of the text to lower case

Removal of puncutations

Removal of English Stopwords

Removal of numbers

Stem document

Create Corpus

Define & eliminate custom stop words

Removal of duplicate tweets

Removal of sparse terms

Calculate TF-IDF of words and set a cut off

Remove words with TF-IDF below the cutoffs

Create Document Terem Matrix (DTM)

Remove documents with no terms in DTM

Observe result

End

Pre Processing : Before & After

Before :

- > df <- read.csv("SmartCities.csv")
- > df$contents[1]

- ***[1] "#Cbus will soon burst w. electric vehicles autonomous shuttles platooning trucks bus rapid transit & smart traffic lights." ~@WIRED***

After :

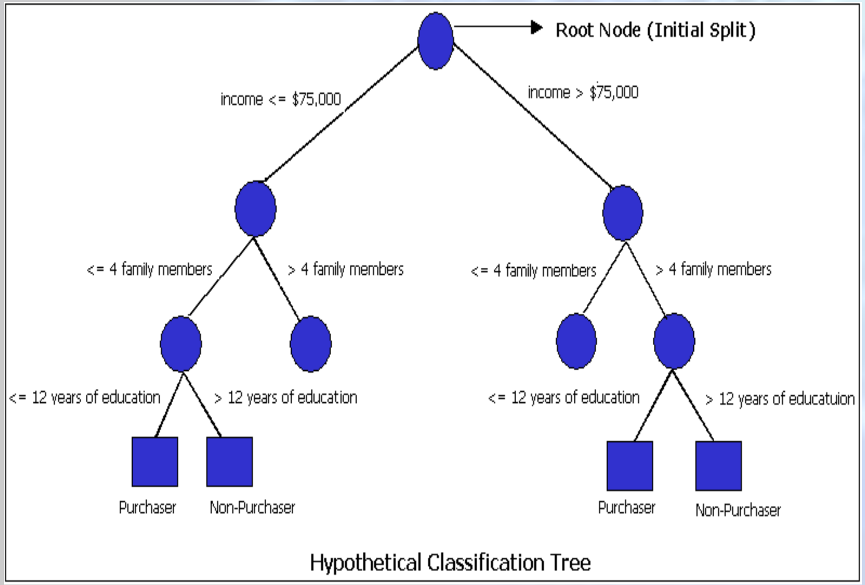
- > writeLines(as.character(corpus[[1]]))

- ***will soon burst w electric vehicles autonomous shuttles platooning trucks bus rapid transit smart traffic light***

Description of algorithms used in Sentiment Analysis

Classification Tree :

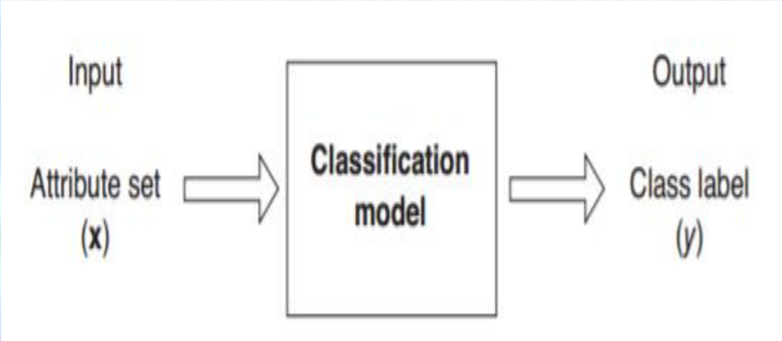
- Undirected graph, which is essentially used for classification
- Supervised learning algorithm
- The tree is created using a method called binary recursive partitioning process



Hypothetical Classification Tree

Random Forest

- Ensembling method which comprises of many tree
- combines the outputs from each and every tree and gives the final result
- Divide , rule & predict better



Input

Attribute set (x)

Classification model

Output

Class label (y)

Naïve Bayes

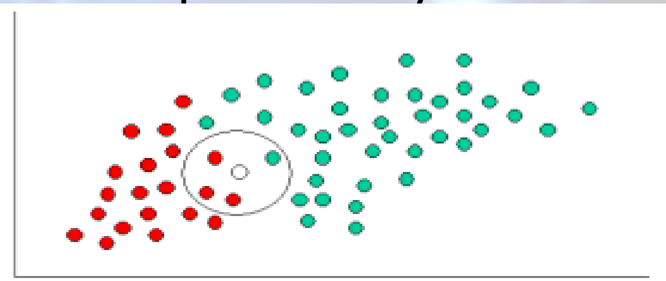
- $P(A_i/B) = (P(B/A_i)*P(A_i))/P(B)$

A_i: Mutually Exclusive events

B: unique set of attribute

B is B_1 Π B_2 Π .... B_m
j=[1,m] ; m= no. of unique words in a document ; B_m = frequency of m^{th} word in a document

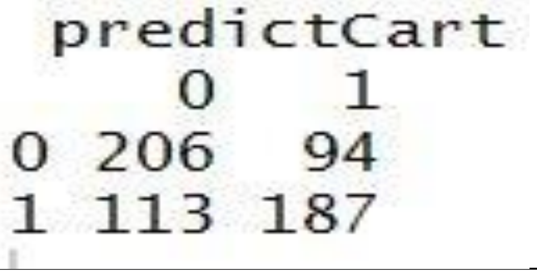
- Any two attributes in the input set are independent of each other
- Concept of posterior and current probability



Output of sentiment analysis of the movie database :

Classification Tree :

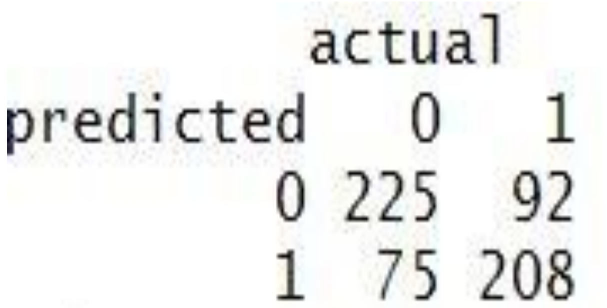
- Used CART, rpart & rpart.plot packages in R
- Accuracy of prediction : 62%



```
  predictCart
      0   1
  0 206  94
  1 113 187
```

Naïve Bayes

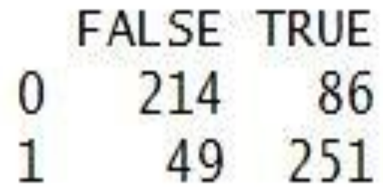
- Package : e1071
- Accuracy came : 72.2 %



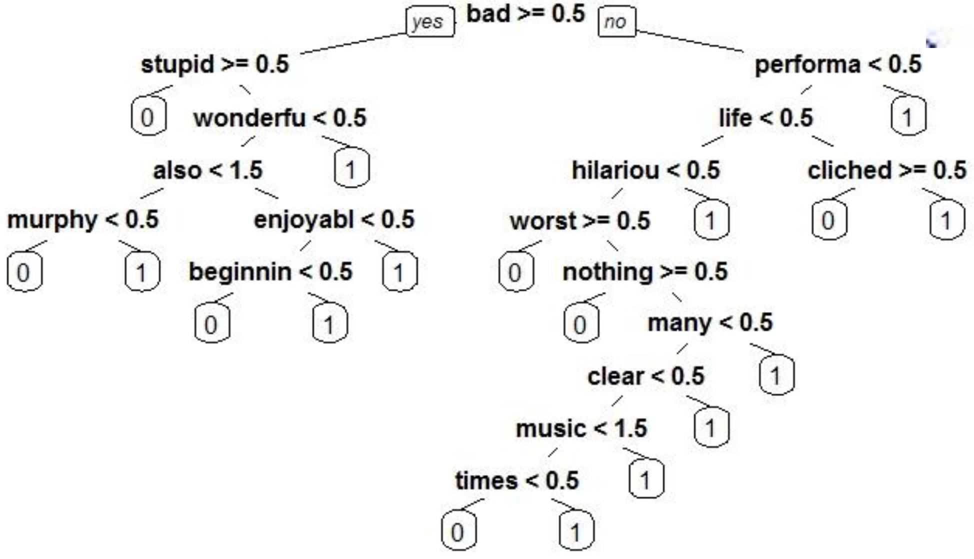
```
          actual
predicted   0   1
        0 225  92
        1  75 208
```

Random Forest

- Package : randomForest
- Accuracy : 77.5 %
- If probability predicted > 0.5 then the same is tagged as a true positive.



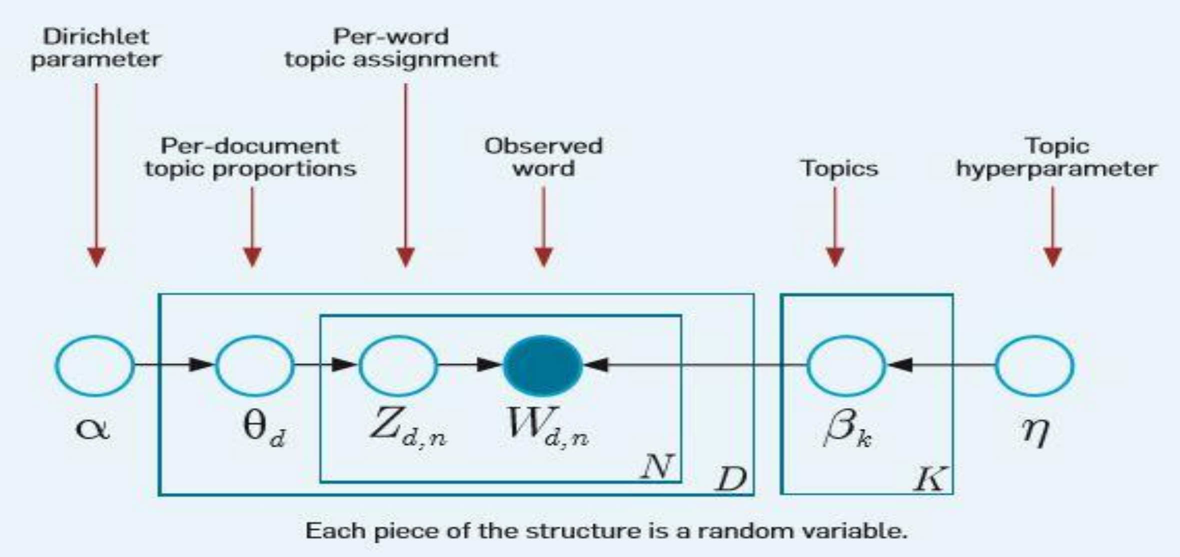
```
  FALSE TRUE
0   214   86
1    49  251
```



Algorithms used for Topic Modelling :

LDA (Latent Dirichlet Allocation)

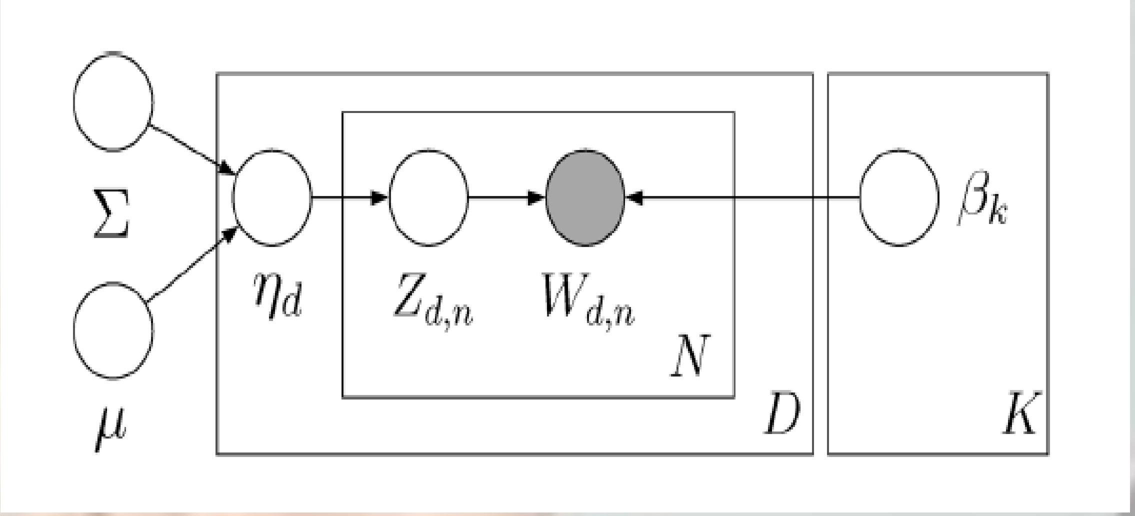
- α,β : controlling factors controlling per document topic distribution and per topic word distribution
- θ : Weight of topics (Distribution of documents over topics)
- Topic : as list of words with assigned probabilities of belonging
- Documents contain multiple topics



Source: Anthes, 2010

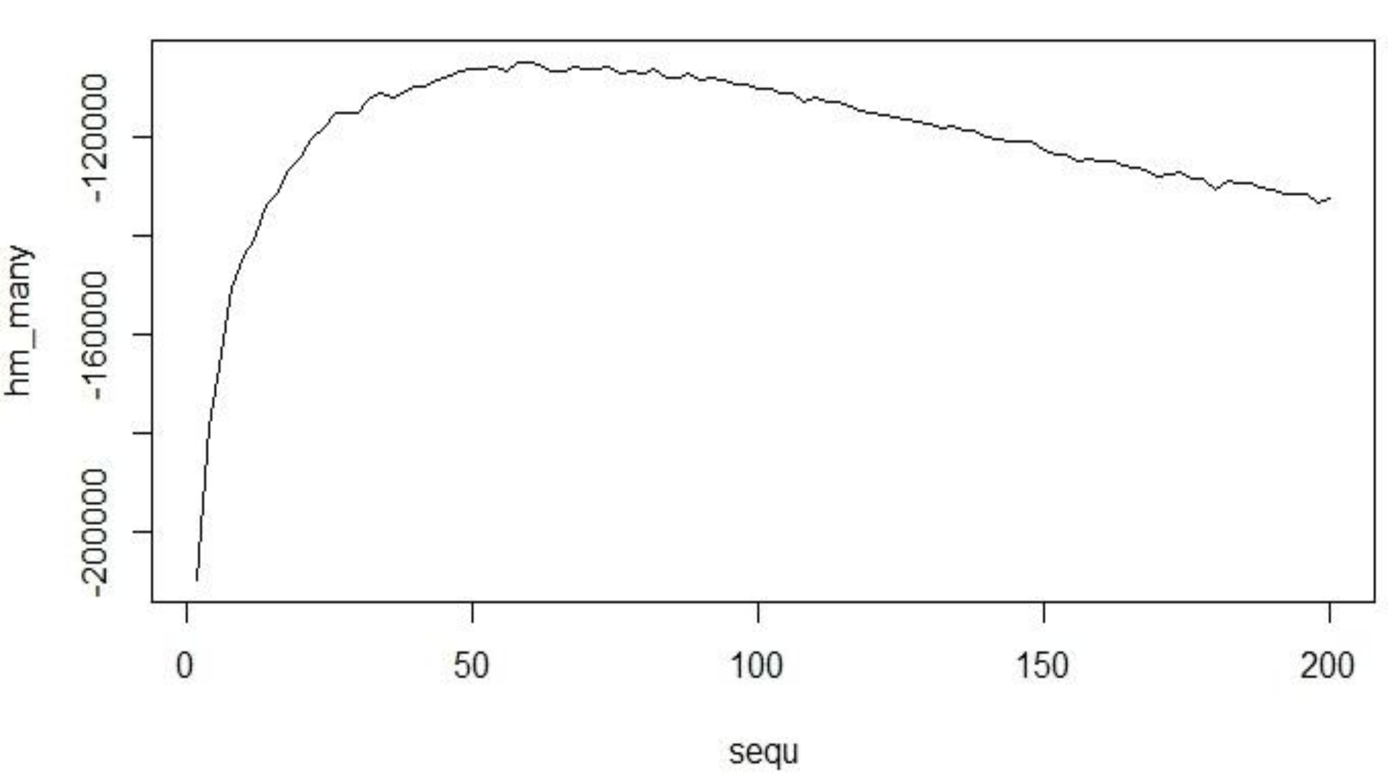
CTM (Correlated Topic Models)

- Each topic is a proportion of words
- Each document is proportion of topics
- Correlation among topics considered
- Bag of Words model
- Each documents assigned to a topic with maximum weightage



Source: Blei & Lafferty, A Correlated Topic Model of Science, 2007

Deciding optimal no. of topics :



- Plotted log likelihood corresponding to each of the topics.
- LDA algorithm applied from 2 to 200 at steps of 2 and then estimating the best fit of the LDA outcome depending on the number of topics.
- Optimum no. comes at 58
- Chose 40 for LDA and 30 for CTM using sense and considering repetition

Results of LDA :

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
tech	columbus	ohio	lighting	columbus	iot	pitch	mendix
transportation	challenge	dot	home	challenge	cisco	final	platform
partnership	million	million	light	million	startup	watch	takes
challenge	dot	officially	building	dot	open	finalists	applications
ideas	competition	challenge	just	vehicles	manchest	live	update
next	challenges	columbus	driving	innovation	incubating	usdot	creates
making	future	total	become	bid	centre	seven	faster
competition	key	check	latest	just	creating	denver	building
proposals	meters	capital	innovation	driving	innovation	become	data
become	heart	key	people	federal	grant	portland	latest

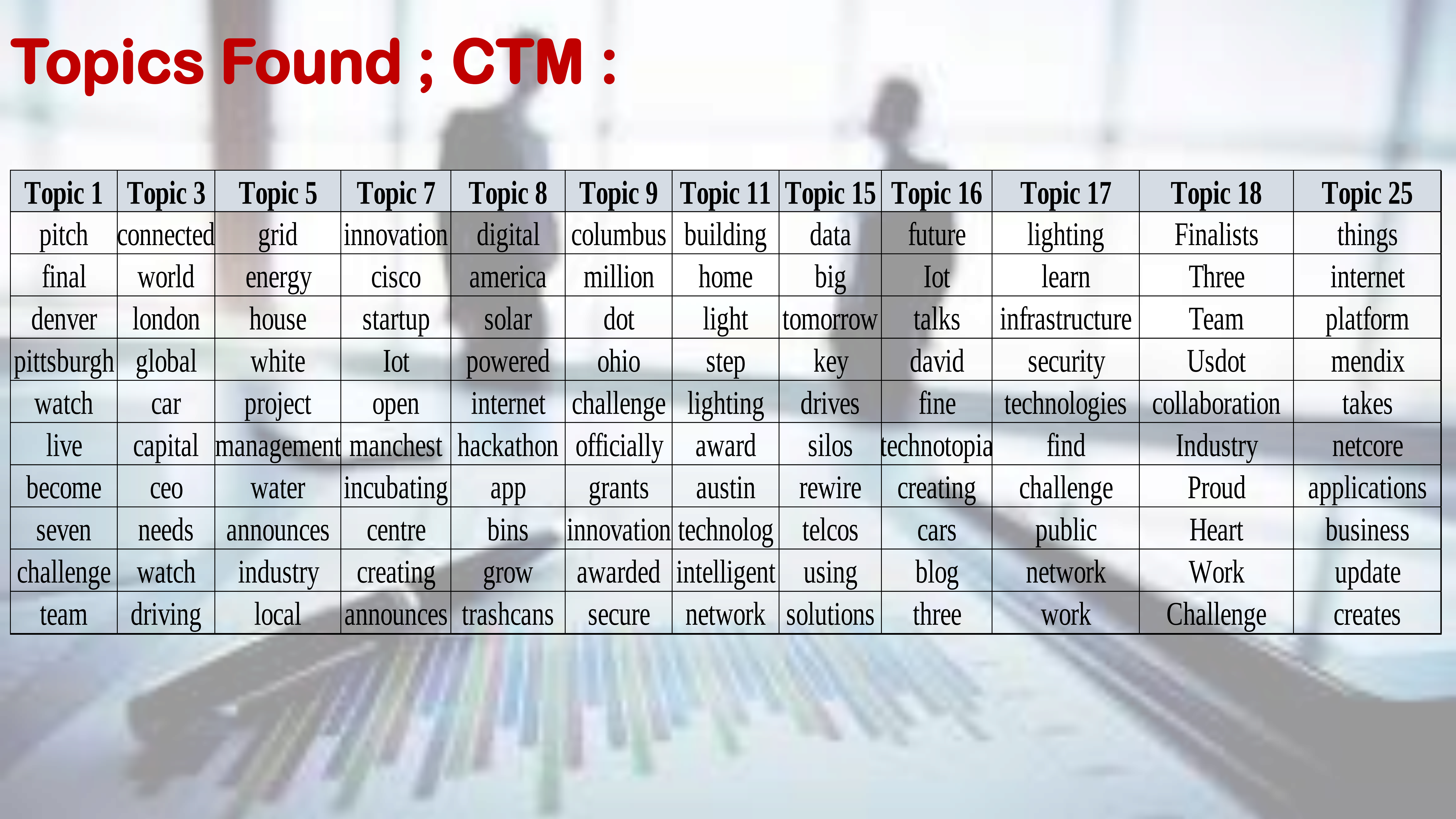
Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
challenge	columbus	public	next	future	innovation	san	data
pittsburgh	challenge	digital	data	urban	summit	francisco	big
video	three	america	futur	cars	austin	plans	solutions
makes	collaboration	lighting	tomorrow	proposals	mayor	proposal	key
proud	heart	team	silos	iot	says	announces	drives
proposal	just	grow	rewire	citizens	officially	real	challenges
sustainable	ohio	infrastructure	telcos	sustainable	challenge	platform	using
usdot	challeng	market	kansas	system	work	portland	work
learn	transportation	looking	create	mobility	find	looking	people
final	awarded	energy	loses	competition	hackathon	ideas	transportation

Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24
million	beat	challeng	grants	meters	iot	connected	netcore
columbus	innov	house	secure	utility	creating	london	things
awarded	grant	columbus	rumored	power	blog	global	business
technolog	europe	white	total	grid	information	car	internet
grant	managem	project	winn	companies	traffic	capital	gaia
infrastructure	water	announces	dot	meter	future	needs	acquires
heart	local	dot	ohio	join	driving	ceo	equity
light	top	join	app	team	futur	world	deal
future	columbus	future	centre	work	home	finalists	app
congrats	competiti	latest	congrats	mobility	seven	project	become

Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30	Topic 31	Topic 32
ohio	traffic	austin	world	build	columbus	challeng	cost
innovation	solution	loses	things	technology	ohio	beats	concept
dot	cloud	bid	internet	cars	dot	breaking	standards
officially	microsoft	grant	projects	using	transportation	federal	faces
challenge	building	news	network	learn	secretary	portland	just
million	system	kansas	create	vehicles	million	six	coordination
denver	cubic	transit	biggest	technologies	foxx	transport	deep
grid	app	key	energy	people	iot	grid	hurdles
seven	technolog	panel	sustainable	top	people	news	find
growth	making	competition	check	iot	challenge	officially	people

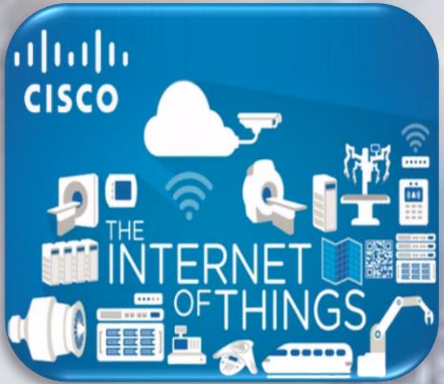
Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
building	columbus	parking	internet	talks	grid	transportation	columbus
award	million	transport	solar	david	energy	denver	challeng
growth	dot	system	powered	fine	industry	challenge	winning
intelligent	ohio	market	hackathon	technotopia	security	department	congrats
step	challenge	systems	challenge	future	solutions	program	challenge
future	check	demand	panel	iot	work	says	mobility
infrastructure	officially	looking	bins	join	proud	ohio	ohio
work	build	tech	trashcans	project	panel	become	systems
sustainable	utility	growth	work	beat	water	technologies	takes
people	business	ceo	iot	drives	powered	next	business

Topics Found ; CTM :



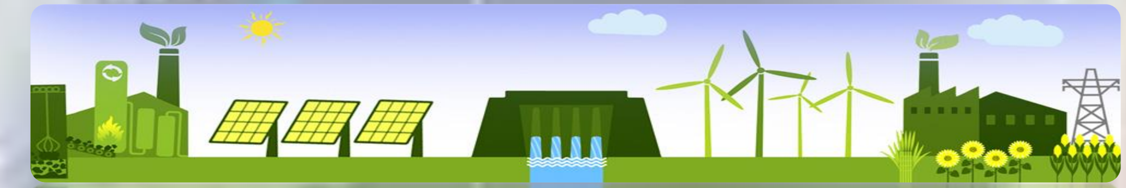
Topic 1	Topic 3	Topic 5	Topic 7	Topic 8	Topic 9	Topic 11	Topic 15	Topic 16	Topic 17	Topic 18	Topic 25
pitch	connected	grid	innovation	digital	columbus	building	data	future	lighting	Finalists	things
final	world	energy	cisco	america	million	home	big	Iot	learn	Three	internet
denver	london	house	startup	solar	dot	light	tomorrow	talks	infrastructure	Team	platform
pittsburgh	global	white	Iot	powered	ohio	step	key	david	security	Usdot	mendix
watch	car	project	open	internet	challenge	lighting	drives	fine	technologies	collaboration	takes
live	capital	management	manchest	hackathon	officially	award	silos	technotopia	find	Industry	netcore
become	ceo	water	incubating	app	grants	austin	rewire	creating	challenge	Proud	applications
seven	needs	announces	centre	bins	innovation	technolog	telcos	cars	public	Heart	business
challenge	watch	industry	creating	grow	awarded	intelligent	using	blog	network	Work	update
team	driving	local	announces	trashcans	secure	network	solutions	three	work	Challenge	creates

- IoT startup by CISCO

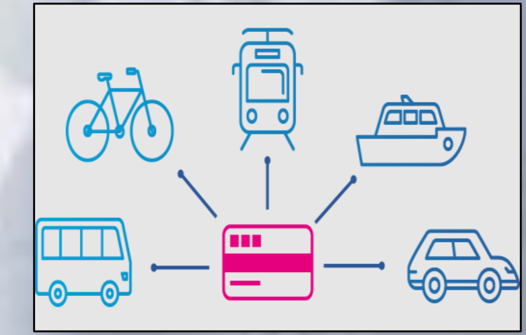


- Energy management
- Water management
- Prevent waste

- Brexit
- London: Global connected city
- Corporates against



- Million Dot Challenge
- DoT, US
- Columbus, Ohio
- Smart transportation



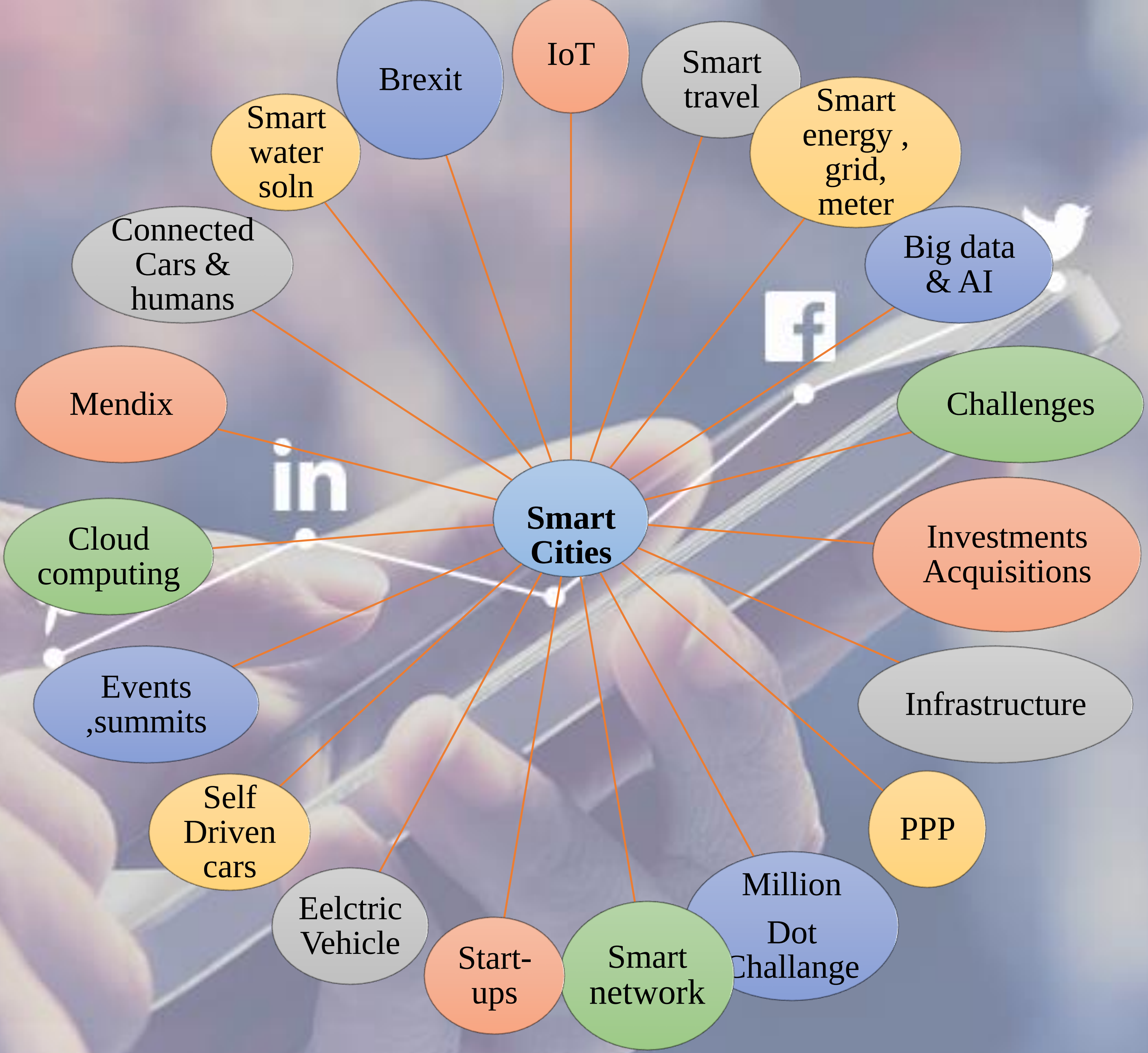
mx mendix
the app platform

- IoT
- IoT enabled cars
- Technotopia

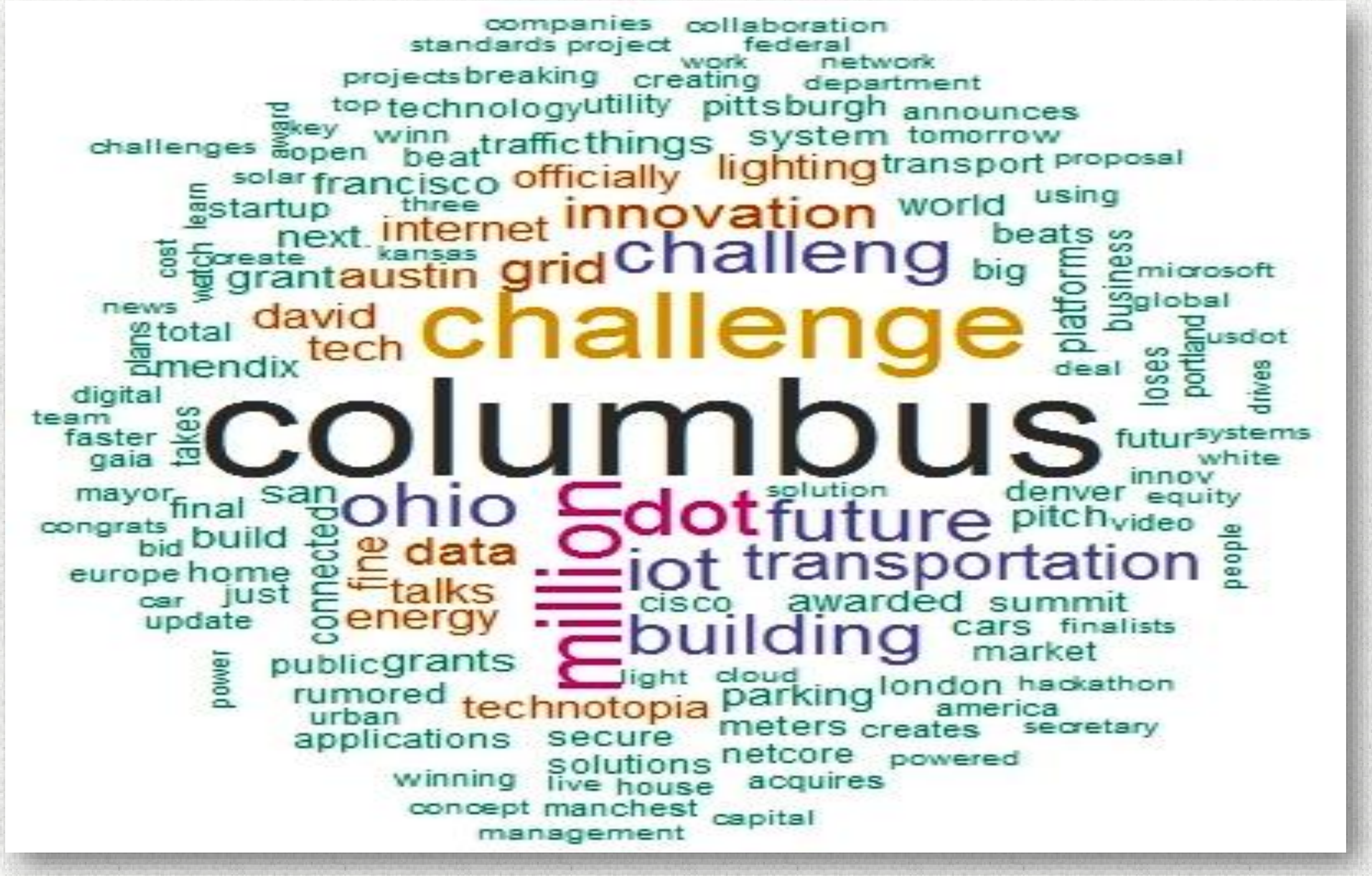
- Mendix
- Acquisition of Netcore by Gaia smart cities

- Participation of Denver & Pittsburgh is Smart City Challenge
- Top 7 finalists

Gist of topics

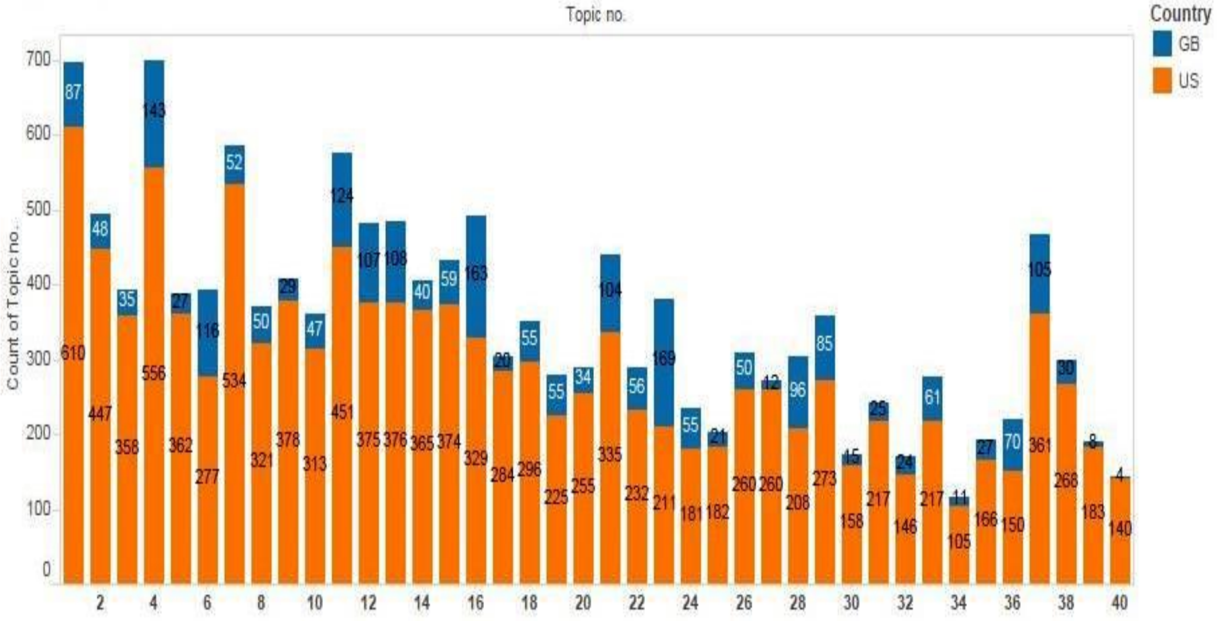


Word Cloud found :



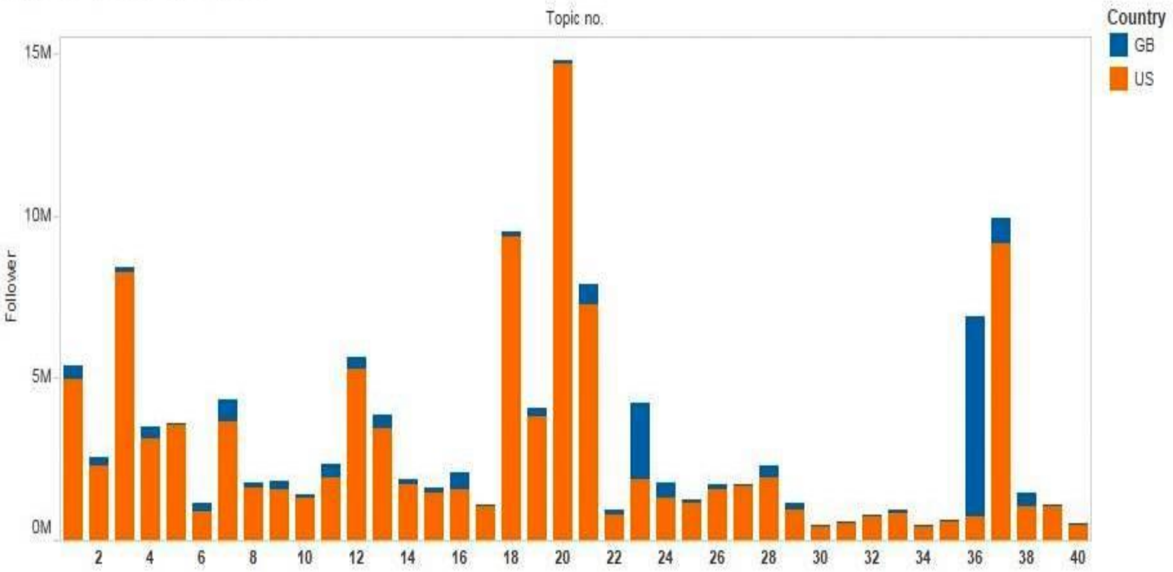
Geographic Distribution of topics across US & UK

Topic wise no. of tweets



Count of Topic no. for each Topic no.. Color shows details about Country.

Topic wise sum of followers

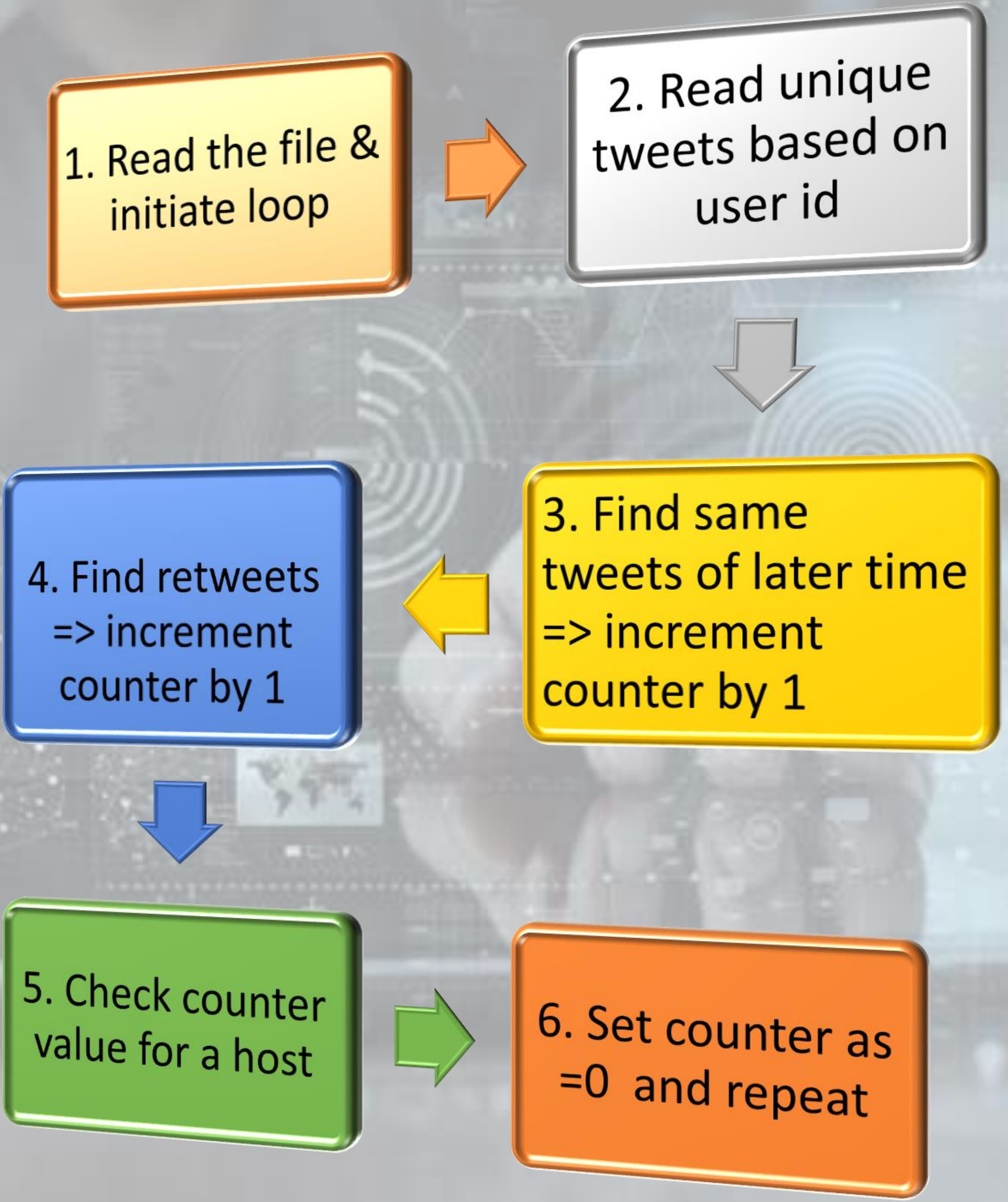


Sum of Follower for each Topic no.. Color shows details about Country.

- Prepared data with topics , region and followers
- Result is little biased because of the nature of data
- US has more influence on discussions , which is evident
- 1st graph : Topics 1, 4, 7, 11, 12, 13, 16, 21, 37 are of utmost importance and have most number of tweets belonging to.
- 2nd graph : Topic 20 has the highest number of followers and the other important topics are 1, 3, 12, 18, 21, 36 and 37.

Finding key influencers :

- Thoughts and views are liked by population.
- Effect the feelings, actions of others.
- **Influencer score** = No. of re-tweets + No. of tweets copied
- **Assumption:** No two tweets of exactly the same content be tweeted at any time without copying the original tweet.



1. Read the file & initiate loop

2. Read unique tweets based on user id

3. Find same tweets of later time => increment counter by 1

4. Find retweets => increment counter by 1

5. Check counter value for a host

6. Set counter as =0 and repeat

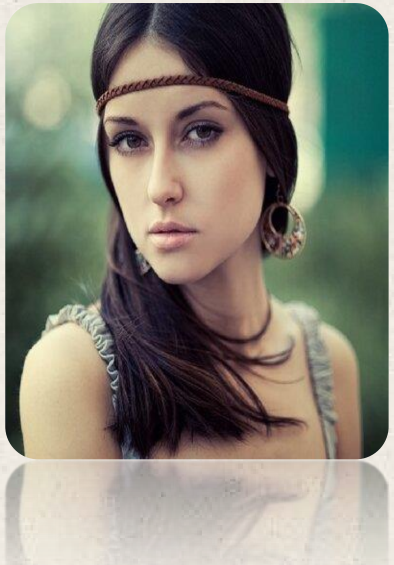
Key Influencers :

Host	Profile Information	Score	Followers
http://twitter.com/daily_paper	Daily tiding updates	569	1361
http://twitter.com/brianjoneill	CEO of HouseMaids	518	5984
http://twitter.com/bradsterhill	Stays updated in technology and web	517	106
http://twitter.com/lacomputertech	Computer repair and IT servicing	508	2207
http://twitter.com/steverobinsojr	Gadgets and technology stock enthusiast	501	1034
http://twitter.com/tweetfortechies	Interested in Smart cars, virtual machines	500	356
http://twitter.com/techtweeties	Interested in technology websites	494	332
http://twitter.com/techallo	Interested in technology news and updates	492	279
http://twitter.com/webtech_update	Updates on latest news and opinion	491	650
http://twitter.com/rosettemak	Techie, Digital marketer	488	44

Brandon Hill
 : Stays updated latest in tech and web.

LA Computer Tech :

- Onsite computer repair and IT servicing
- Tweets regarding minute tech details, cool computers, iPhone.



Techallo

- Technology news and updates
- Not restricted to any specific technology

Daily Papers :
Tweets regarding daily tidings

Brian o' Neil :

- Co-Founder and CEO of House Maids.
- Residential cleaning services throughout Sarasota, Florida.
- Master in Service Industry.

Steve Robinson Jr
- Tech enthusiast
- Tech gadgets, Tech stocks.

Tweets
for
Techies

Tweets for Techies
- Tweets about latest technology
- Smart cars, Virtual machines, E-Commerce

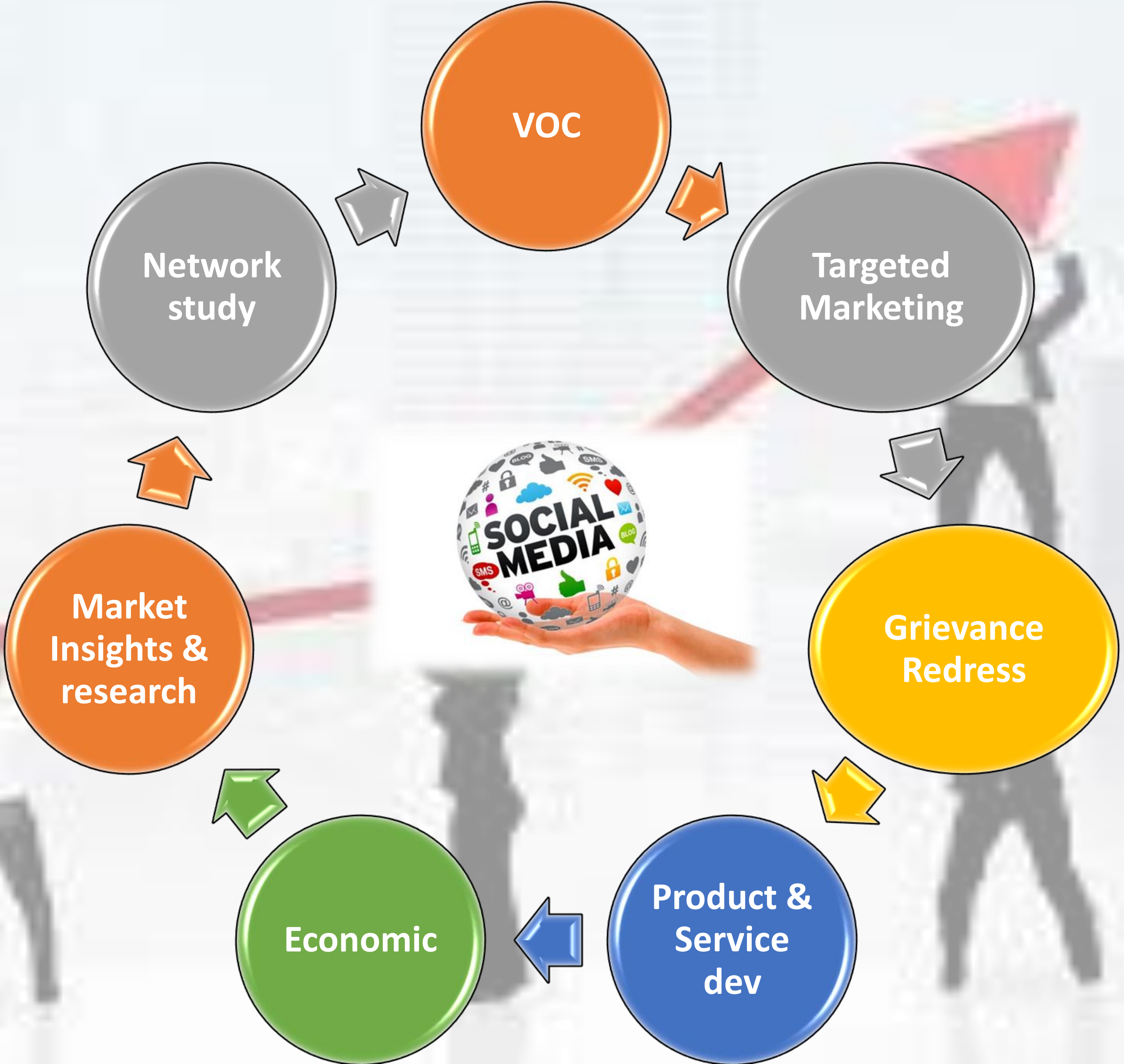
Web
Tech
Update

Web Tech Update
- Latest news and opinion on everything.

Tech Tweeties
- Tweets from Technology websites

Mak Rosette
- Techie, Digital Marketer
- Interest areas: Internet security and Privacy, Startups

Business Economics and benefits


VOC

Targeted Marketing

Grievance Redress

Product & Service dev

Economic

Market Insights & research

Network study

Conclusion and scopes of improvements

- Cost effective solution
- Topics in line with our business understanding
- The change in classic framework of marketing : Knowing customer is a different ball game now
- Identification of key influencers can help in branding & positioning
- Targeted marketing approach
- Future scope : Application of Hidden Markov Modelling *(Ref.: . Gruber, Zivi, & Weiss, 2007)* and Tensor Factorization *(Ref. :Arora, 2015)*

Thank You!