DATA SCIENCE THOUGHT LEADERS ON TWITTER

HOW THE THOUGHT LEADERS INFLUENCING PEOPLE AROUND THE WORLD ON TWITTER

A. INTRODUCTION

In today's connected world it is relatively easy to share individual thoughts and moments with the people. The evolution of social media is only making it easier. I was always intrigued by the questions that how the thought leaders in different domains are influencing people through social media and is there any common characteristics that can be found in their shared content. When initially the influencers put more domain or competency focused contents in professional networks like LinkedIn, but in last few years they are equally giving importance to other social medias like Twitter, Facebook etc. In that context, I considered a list of all influencers in Data Science domain, published by Onalytica. In this report, I will try to provide my findings on those two questions mentioned.

B. DATASET

Onalytica published a list of 100 most influential people in social media, who are considered thought leaders in Data Science domain. I used the list of their twitter handles to extract all tweets within a maximum allowable limit (~3200 tweets per twitter handle). Along with that, I got few other relevant metrics as shown below:

Twitter Handle | Profile Photo Link | Location | Tweets | Date of Tweets | Retweets Count | Likes Count | Followers Count

I used Tweepy API in python to collect and store all tweets from twitter to my local storage. However, for this analysis I used last one year (01/12/2017 till 01/12/2018) tweets from all personnel. The end dataset was consisting of 2,34,401 tweets.

C. PROCESS

Data was collected in different forms including text, numeric, geolocation etc. Then data was stored in multiple files in csv format. Using glob and pandas library in python, data was read and converted to dataframe. However, data preprocessing was done according to the need of data visualization.

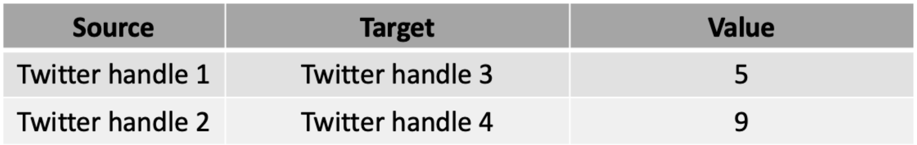
C.1 Data Pre-processing

Data were divided into two parts: first part containing all tweets (text data) along with screen names and second part containing other numeric metrics data like followers count, tweets count, retweets count, total likes along with location data and profile image urls. Appropriate aggregate functions were used to get total tweet counts, retweet counts and likes. Date strings were also converted to pandas date-time format.

To visualize meaningful text data, separate data cleaning and processing steps were taken into consideration.

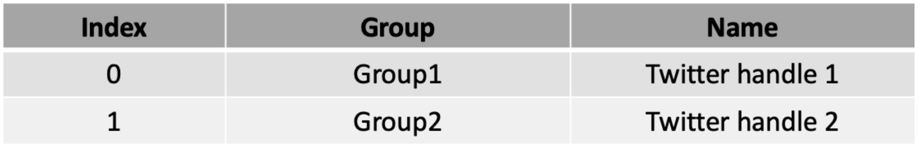
- Step 1: All mentions (starting with @) of people and communities are separated using regular expression in python

- Step 2: A dataframe was created with three features: source, target and value. Source denotes the twitter handle that mentions about any other twitter handles whereas the mentioned twitter handles are defined as targets. Value describes the total frequencies the source twitter handle mention about target.



Source	Target	Value
Twitter handle 1	Twitter handle 3	5
Twitter handle 2	Twitter handle 4	9

- Step 3: Another dataframe (node) was built with three attributes: Index, Group, and twitter handle name. Groups were created based on the types of content any twitter handle publishes like AI, Blockchain, Cyber Security etc.



Index	Group	Name
0	Group1	Twitter handle 1
1	Group2	Twitter handle 2

- Step 4: All text data were tokenized, and irrelevant words (stop words) and special characters were removed. In addition, words were lemmatized. In this way a clean data corpus was built. A preprocessing function in python was written to do this job.

C.2 Interactive Map

I developed an interactive map showing current locations of thought leaders along with few profile details of individuals like Name, Followers count, total tweets etc.

- Chart Type: Pinpoint Map

- Design Principles Followed:

- *Visual Contrast:* Map features and page elements contrast with each other.
- *Legibility:* Selecting right symbols and its size to make map more understandable.
- *Balance:* It involves organization of the map and different elements within map.

- Libraries/Functions Used: folium [Map (location, zoom_start, tiles), Marker (location, icon, tooltip, popup)]

C.3 Chord Diagram

An Interactive Chord Diagram was made to show the inter-relationship between different twitter handles. In chord diagram, value defines the frequency of interactions between two twitter handles. To make it more readable, strength threshold was kept high (200).

- Chart Type: Radial Chord Diagram
- Design Principles Followed:
  - *Grouping in sequence:* Nodes and Links data to be sorted before drawing diagram
  - *Color Relevance:* Selecting right color palette for nodes and links
  - *Legend:* Legend is very relevant when multiple groups exist
- Libraries/Functions Used: holoviews, bokeh, pandas, numpy

C.4 Topic Modeling

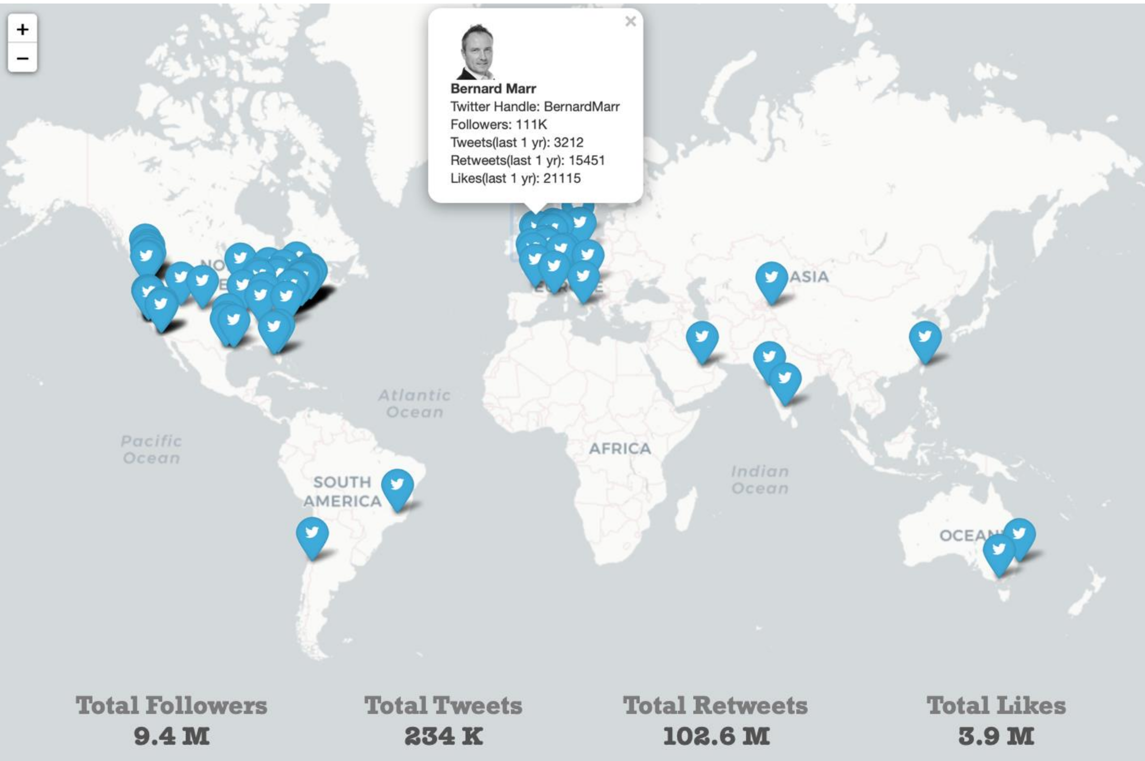
Latent Dirichlet Allocation (LDA) is frequently used to classify text to abstract topics and I used this technique to find different topics discussed by the influencers on social media.

- Chart Type: Bubble chart and Bar Chart
- Design Principles Followed: As described by pyLDAvis library
- Libraries/Functions Used: pyLDAvis, nltk, gensim, numpy

D. RESULT

D.1 Distribution

In the first spatial graph I tried to find distribution of these thought leaders across globe.



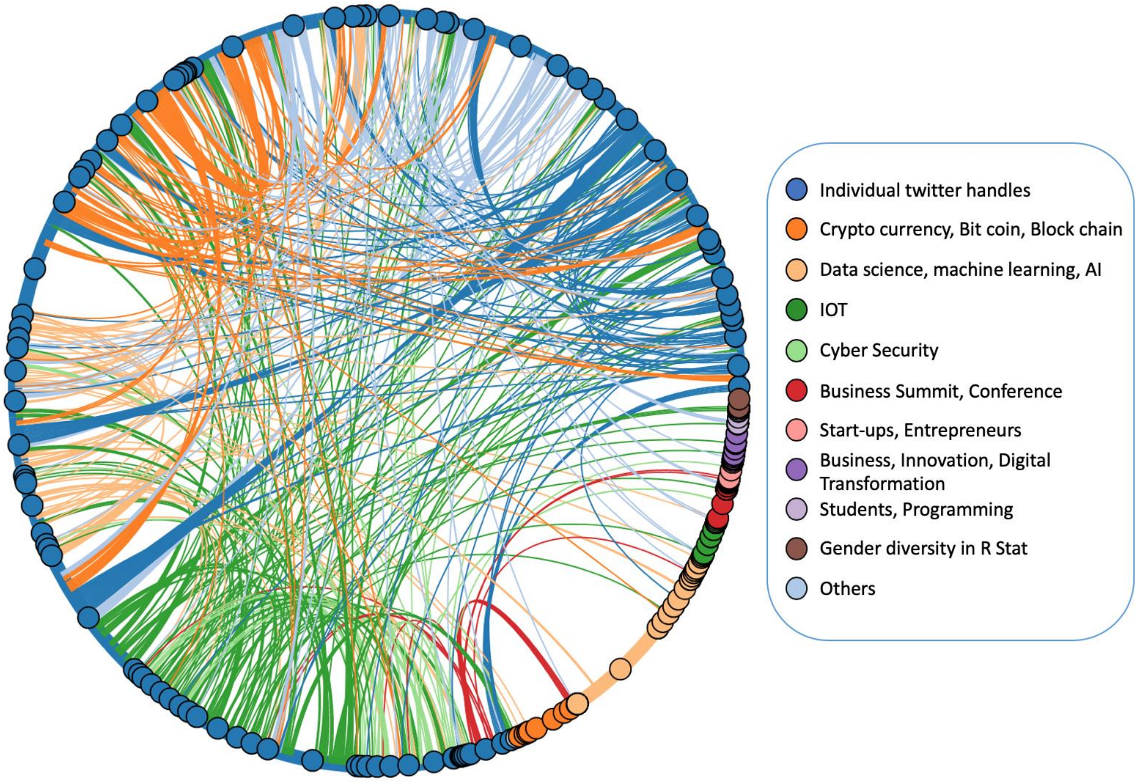
Viz 1: Distribution of Influencers around the world

Data Analysis:

- At first view, I can see that top influencers are concentrated mostly in two continents: North America and Europe. Top two cities where they are mostly based on are Seattle and New York City.
- The influence level can be easily understood from few basic metrics as collected or aggregated. They have together 9.4 million followers on Twitter.
- In last one year, they have tweeted more than 2,34,000 times on their twitter handle. Their messages got retweeted more than 102 million times.
- All those tweets got 3.9 million likes in total.

D.1 Mention

In the second visualization, I tried to find the inter-relationship between different twitter handles they mentioned from time to time in last 1 year. Then I grouped together these twitter handles into eleven different groups based on the type of content they used to share. I found few really interesting insights after analyzing the groups from this interactive chord chart.


Individual twitter handles
Crypto currency, Bit coin, Block chain
Data science, machine learning, AI
IOT
Cyber Security
Business Summit, Conference
Start-ups, Entrepreneurs
Business, Innovation, Digital Transformation
Students, Programming
Gender diversity in R Stat
Others

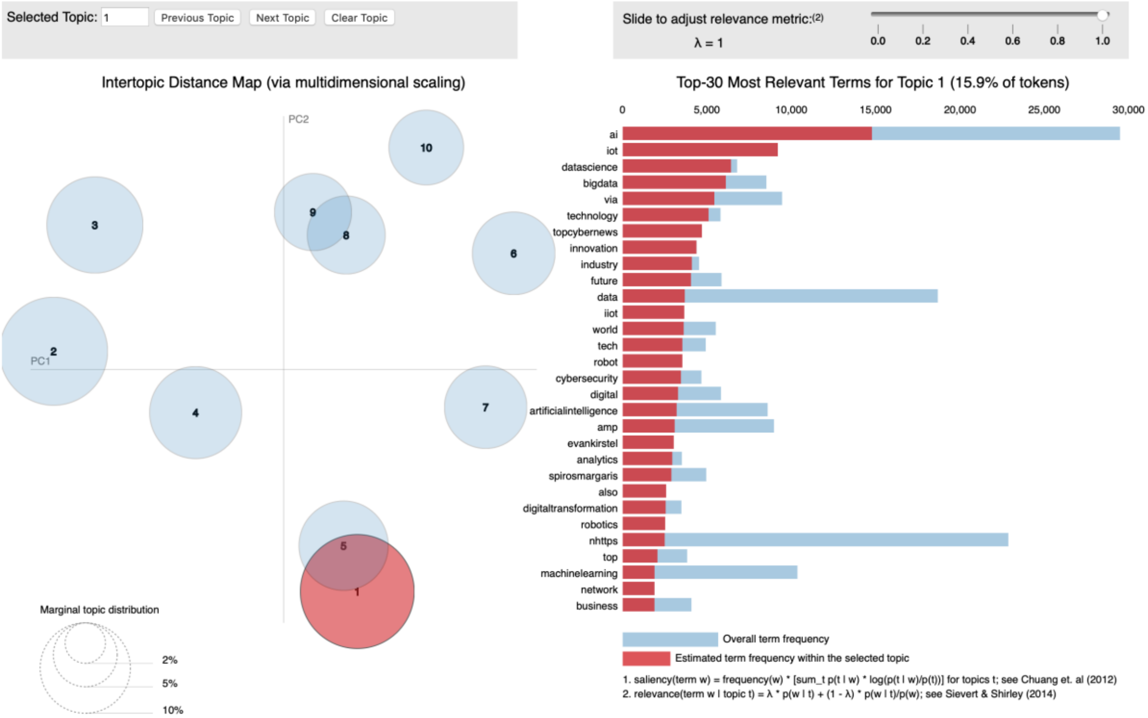
Viz 2: Interrelationship between twitter handles

Data Analysis:

- The edge color index is based on source of the tweets. It can be inferred that, the thought leaders used to mention other leaders very frequently in their tweets. They could have done that by either mentioning others' work published in different media or by re-sharing their contents from timeline.
- They also mention about other relevant community twitter handles across domains. In this way they are actually sharing more valuable contents over social media. AI, Blockchain, IOT, Cyber Security are on the top mentions' category in their tweets.
- Few Red lines at bottom are showing their mentions on Business Summit and Conferences in Data Science, Blockchain etc. In this way their followers would be more informed about upcoming conferences and meet-ups.

D.1 Discussion

In the final visualization, I have tried to analyze all tweets to find the most discussed topics in their contents. I used LDA for topic modeling and pyLDAvis for visualization. Top ten topics are indicated by circles on left side and associated most frequent words are shown on the right side of Viz 3. I found quite few interesting topics to discuss on.



Selected Topic: 1

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t l w) * log(p(t l w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w l topic t) = λ * p(w l t) + (1 - λ) * p(w l t)/p(w); see Sievert & Shirley (2014)

Viz 3: Top ten topics discussed over last one year

Data Analysis:

- The very first topic is all about different technologies like AI, IOT, Data Science, Bigdata etc. The fifth topic is quite intersected with topic 1 and it mainly focus on machine learning, deep learning, nlp, programming.
- Topic three is mostly discussed on statistics, R, R Studio, ggplot etc.
- Topic four is defined by shared content on blog post and other writings shared by these influencers.
- Topic seven is about banking and financial content. Top pics are 'fintech', 'banking', 'insurtech', 'finserv' (with lambda 0.33 on top sliding contron)
- Apart from that they also talk about other different relevant topics like smart city, cyber security, digital transformations etc.

E. CONCLUSION

In this report, I tried to depict few basic properties of influence and how these thought leaders in Data Science are continuously putting their best effort to let us take the informed decisions in businesses or even in our daily lives. I also tried to follow best design principles while creating different interactive charts and defining markers and their attributes. I have also included my code workbook for reference.