CLUSTERING TWEETS AND WEBPAGES

SAKET VISHWASRAO
SWAPNA THORVE

SOCIAL NETWORK

LIJIE TANG

May 3, 2016

Spring 2016
CS 5604 Information storage and retrieval
Instructor : Dr. Edward Fox

Virginia Polytechnic Institute and State University,
Blacksburg, VA 24061

CLUSTERING

CLUSTERING OVERVIEW

- Feature Extraction
  - TF-IDF
  - word2vec
- K-means
- WSSE evaluation
- HBase Schema
- Clustering and LDA result evaluation
- Future Work

FEATURE EXTRACTION

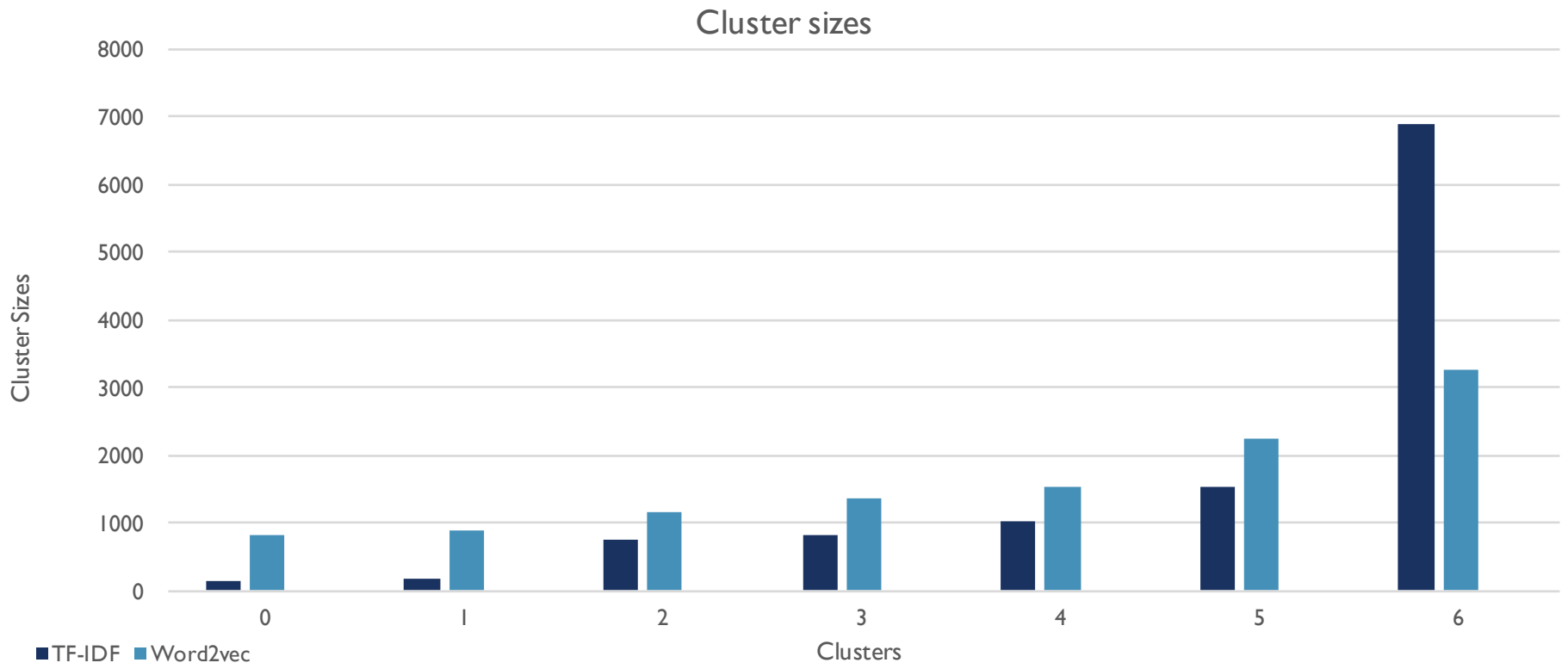
- TF-IDF
  - Represent documents as a bag of words model
  - Vector dimension = Vocabulary size
  - Word score = TF-IDF

- Word2vec
  - Combine all tweets to a single document
  - Train a neural network and extract vector representation of each word
  - Document vector = Sum all vectors (for each word) in a document

CLUSTERING WITH K-MEANS

- Normalize data using $||L_2||$ norm
- Write to HDFS as part files.
- Compute
  - Within Set Sum of Squares (WSSE) scores
  - Tweet distribution

COMPARISON OF TF-IDF VS. WORD2VEC



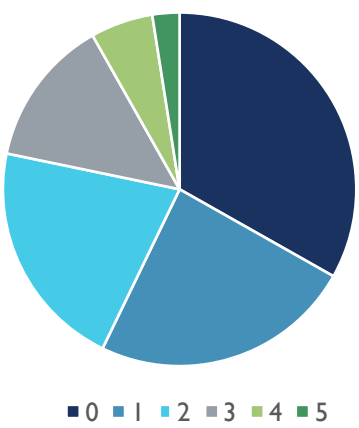
WSSE BASED EVALUATION

RUNNING OVER DIFFERENT TWEET COLLECTIONS

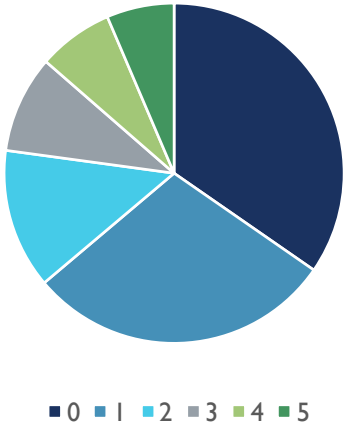
Collection Clusters	541 #NAACPBombing	602 #Germanwings	668 #houstonflood	686 #Obamacare	694 #4thofjuly
0	1374	31189	283	5333	6627
1	7504	26297	899	36070	5529
2	18375	5785	2605	16899	762
3	3166	11979	3672	69555	13663
4	1009	8328	1488	36383	1427
5	2437	6459	5884	19302	3961

RUNNING OVER DIFFERENT TWEET COLLECTIONS

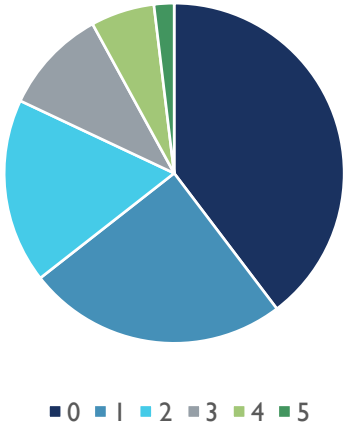
541
#NAACPBombing



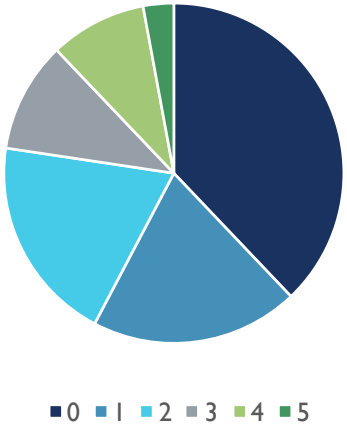
602
#Germanwings



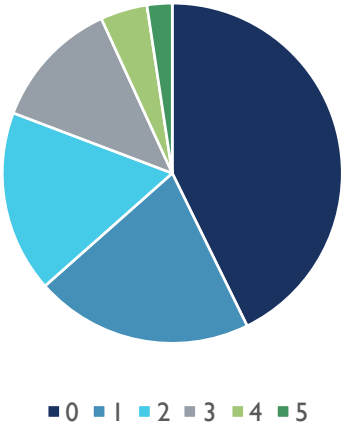
668
#houstonflood



686
#Obamacare



694
#4thofjuly



HBASE SCHEMA DESIGN

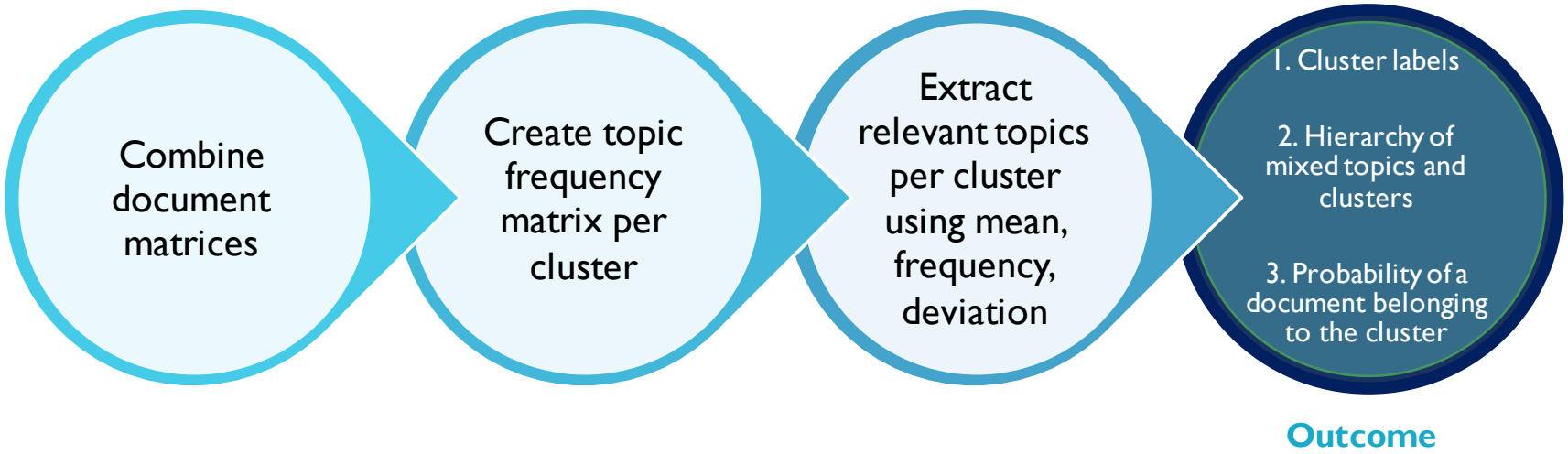
Document ID	Cluster no.	Cluster label	Probability
String (Tweet-id/URL)	Integer	String	Float

COLLECTIONS EVALUATED

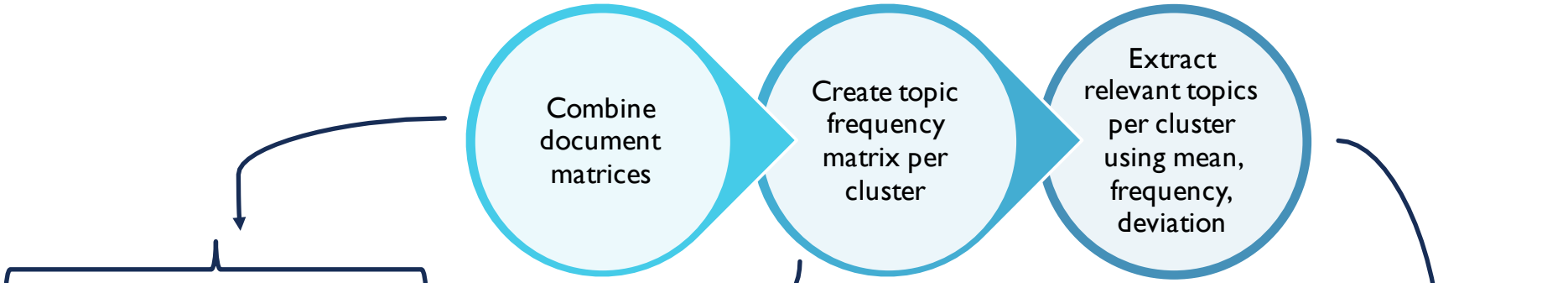
- Tweets
  - 602,668,694,541

- Webpages
  - 686 (TODO)

CLUSTERING EVALUATION USING TOPIC ANALYSIS DATA



EXPLANATION



Doc id	Cluster no

+

Doc id	Topics **Probability**

How many times a topic occurs in a cluster ?

Cluster no	Topic frequency
1	T1=2, T5=2, T2=1, T3=1
2	T1=3, T3=3, T5=1, T2=5

Cluster no	Topics per cluster
1	T3, T2
2	T1, T2

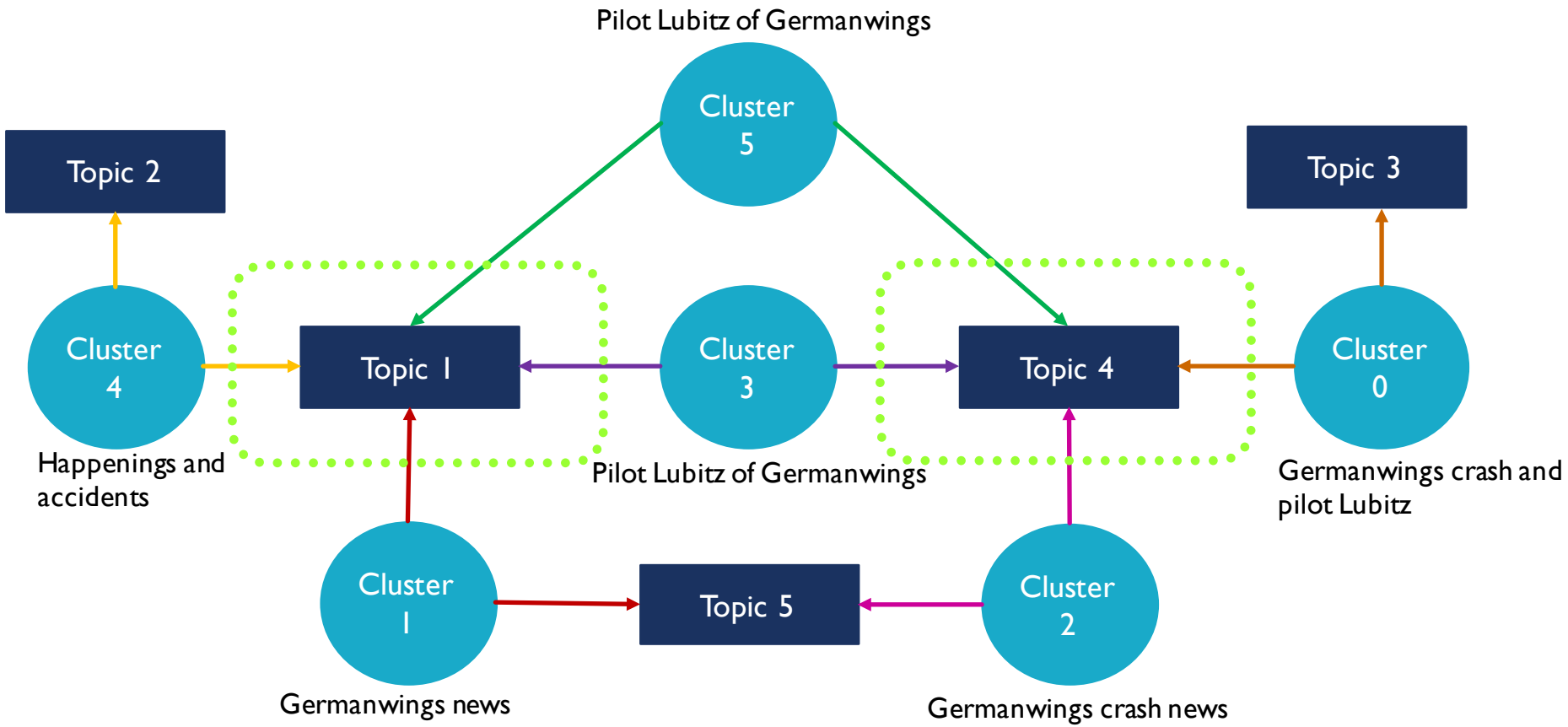
RESULTS

Collection 602 - tweets

Cluster no (wordToVec)	Topics	Topic words	Cluster labels
0	Topic 3, Topic 4	crash,germanwings,pilot,victims,plane,Lufthansa germanwings,lubitz,copilot,flight,playing	Germanwings crash and pilot Lubitz
1	Topic 1, Topic 5	germanwings,ripley,believe,click,bigareveal germanwings,world,charliehebdo,lives,garissa	Germanwings news
2	Topic 4, Topic 5	germanwings,lubitz,copilot,flight,playing germanwings,world,charliehebdo,lives,garissa	Germanwings crash news
3	Topic 1, Topic 4	germanwings,ripley,believe,click,bigareveal germanwings,lubitz,copilot,flight,playing	Pilot Lubitz of Germanwings
4	Topic 1, Topic 2	germanwings,ripley,believe,click,bigareveal germanwings,copiloto,vctimas,accidente,vuelo	Happenings and accidents
5	Topic 1, Topic 4	germanwings,ripley,believe,click,bigareveal germanwings,lubitz,copilot,flight,playing	Pilot Lubitz of Germanwings

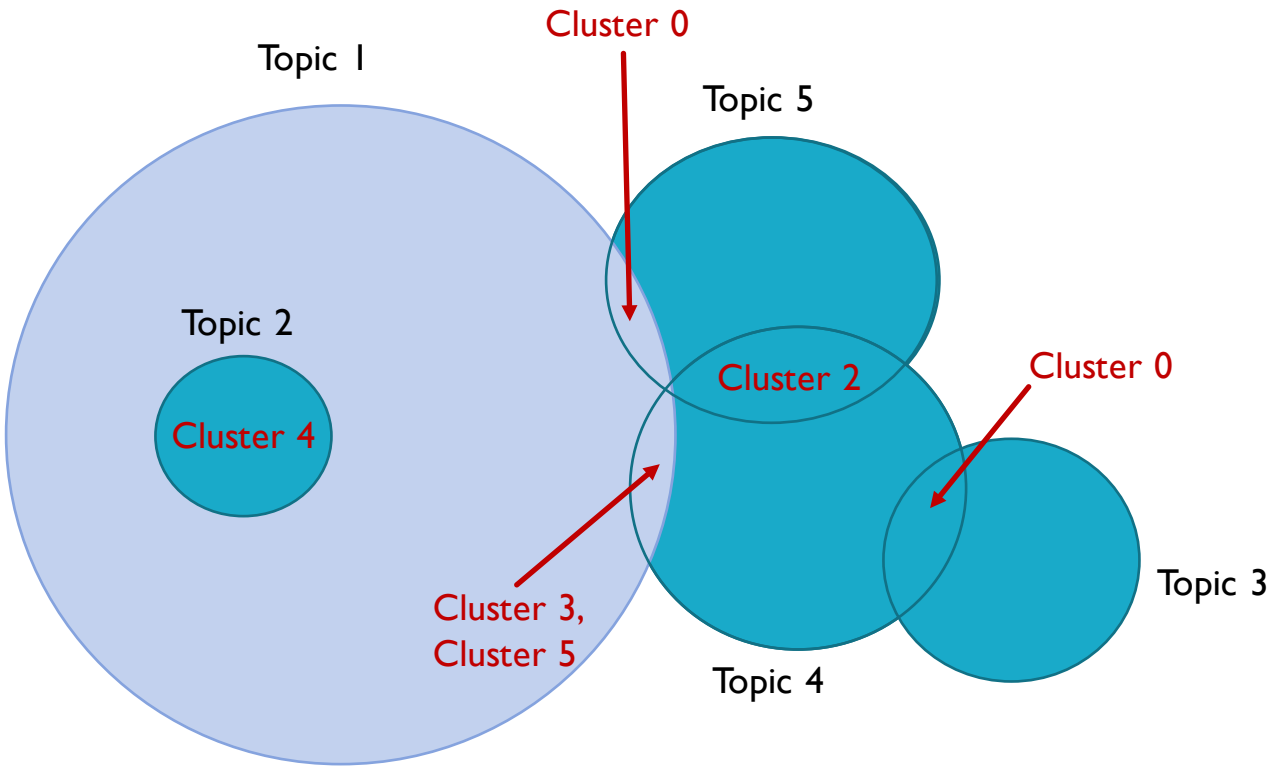
HIERARCHY OF MIXED TOPICS AND CLUSTERS

Collection 602 - tweets



HIERARCHY OF MIXED TOPICS AND CLUSTERS

Collection 602 - tweets



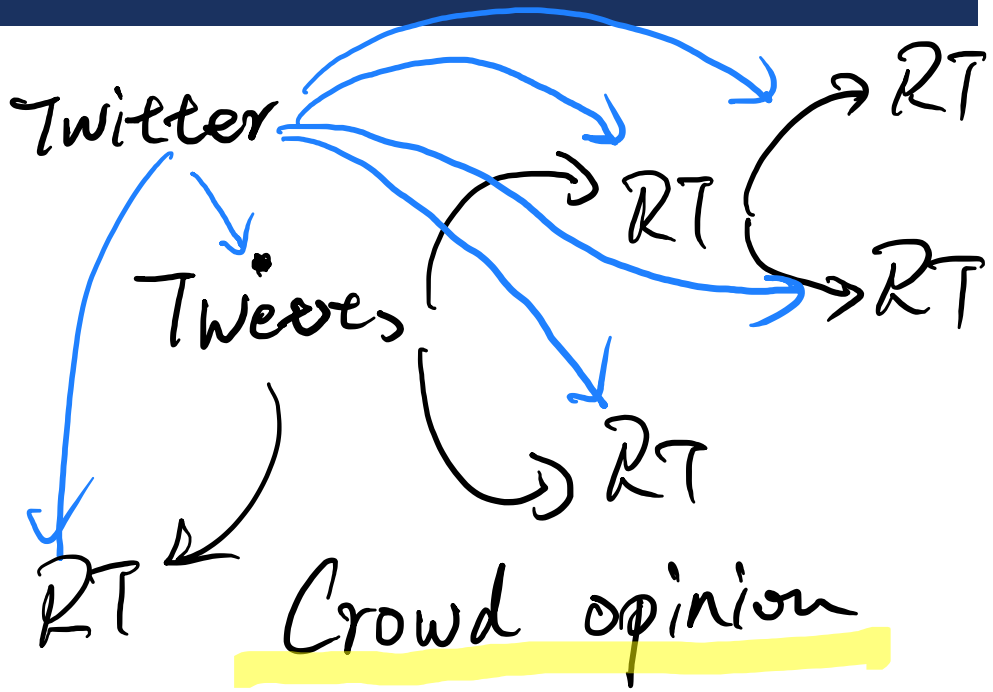
FUTURE WORK

- Use probabilistic models for clustering
- Clustering evaluation
  - Evaluate clusters with internal and external criteria
  - Implement Silhouette scoring in Spark-Scala
  - Establish ground truth for comparison of evaluation results (probabilities)
- Labeling of clusters
  - Label extraction from clusters words
  - Compare cluster labeling methods
- Clustering as a start point
  - Feed clustering results to LDA/classification/collaborative filtering for more accurate results.

SOCIAL NETWORK

OBJECTIVE

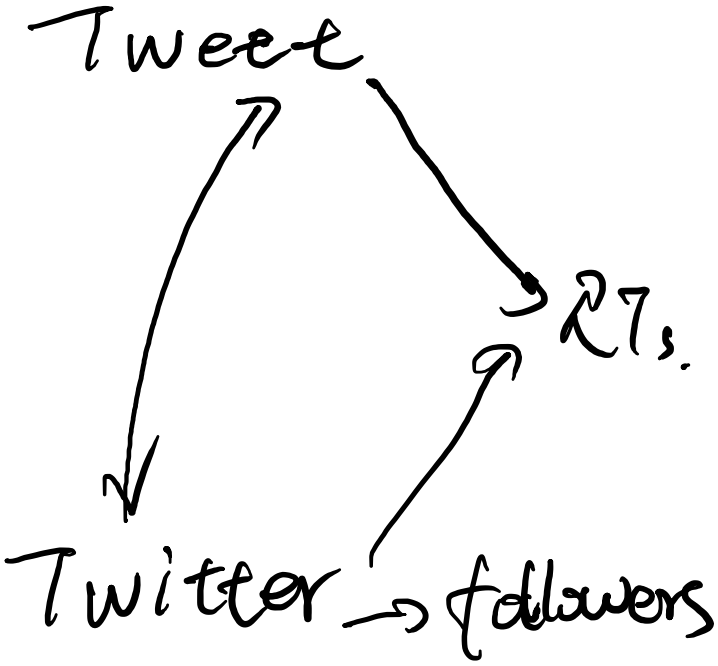
- To find the most relevant content given an User Query
  - Clusters/Classification/Topic Modeling are content driven
  - Social Network are User driven
    - Query is user generated



ASSUMPTIONS

- 1. More RT = important tweets
- 2. More RT = important accounts
- 3. More follower = important accounts
- 4. Generated/Distributed by important accounts = important tweets
- 5. Higher RT ratio inside cluster = important tweets/accounts in the cluster
- 6. More follower inside cluster = important accounts in cluster

- Clusters could be content cluster or SN cluster



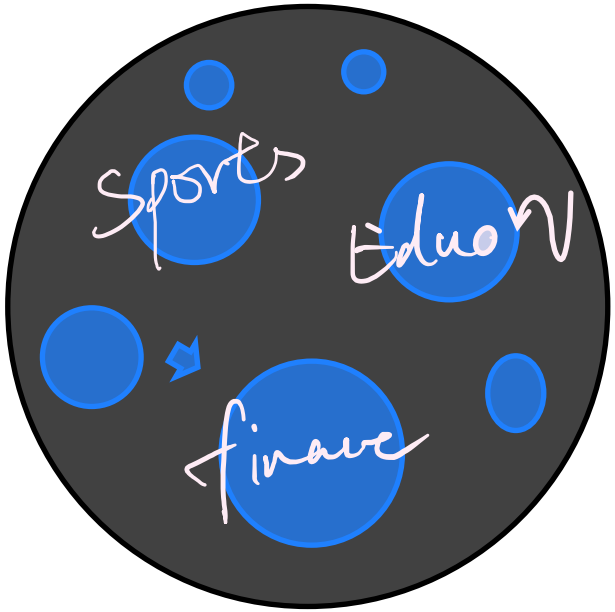
IMPORTANCE FACTORS (GENERAL)

- 1. Start from follower counts:
  - $IF_{account}$ = Number of followers/Maximum follower count in SN
- 2. Tweet IF:
  - IF_{tweet} = SUM (RT * $IF_{account}$) for each RT
- 3. Account IF:
  - $IF_{account}$ = SUM (IF_{tweet})/Tweet count
- Repeat 2, 3 until converge

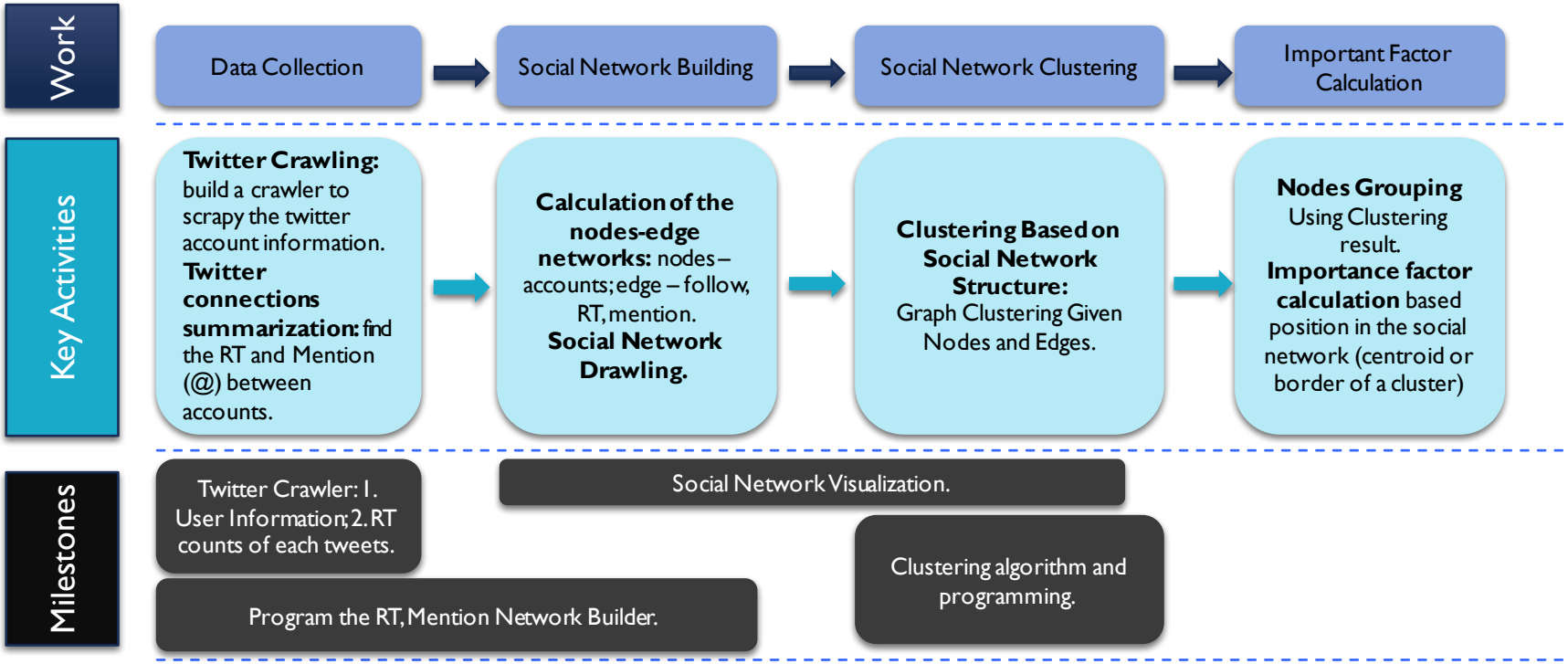
IMPORTANCE FACTORS (CLUSTER)

- ? Should we consider outside influence ?
  - Two tweet with same RT network inside a cluster
    - Are they the same important?

- In our approach
  - Inside IF is calculated as the general approach
  - Inside IF is more important than general IF
    - Inside IF first
    - General IF second

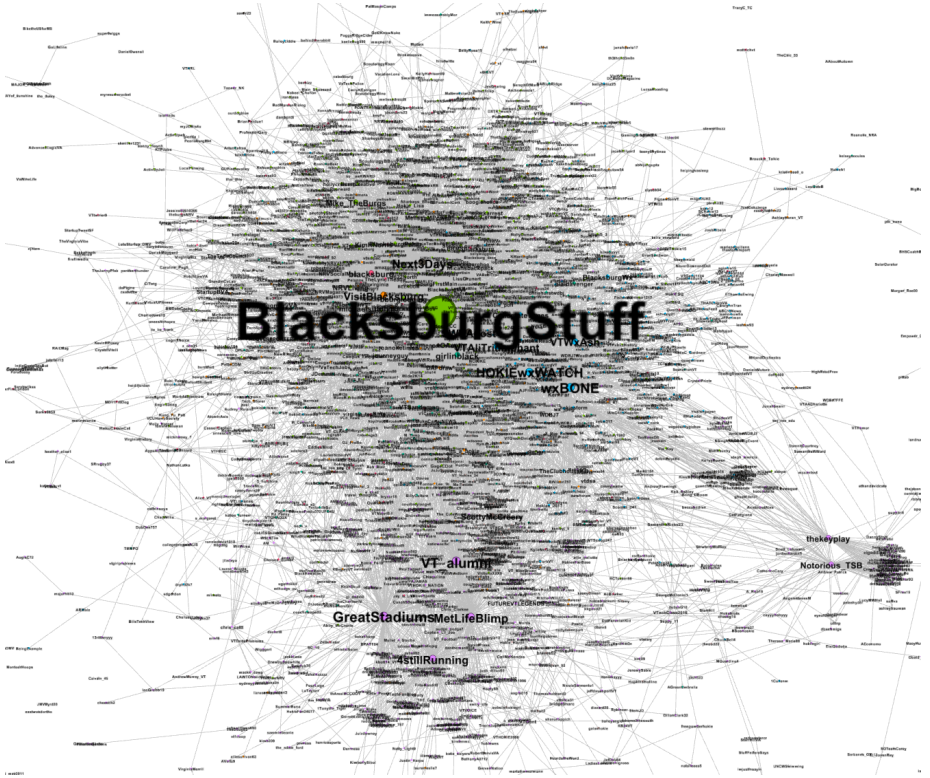


WORK FLOW


Work
Data Collection
Social Network Building
Social Network Clustering
Important Factor Calculation
Key Activities
Twitter Crawling: build a crawler to scrapy the twitter account information. Twitter connections summarization: find the RT and Mention (@) between accounts.
Calculation of the nodes-edge networks: nodes – accounts; edge – follow, RT, mention. Social Network Drawling.
Clustering Based on Social Network Structure: Graph Clustering Given Nodes and Edges.
Nodes Grouping Using Clustering result. Importance factor calculation based position in the social network (centroid or border of a cluster)
Milestones
Twitter Crawler: 1. User Information; 2. RT counts of each tweets.
Social Network Visualization.
Clustering algorithm and programming.
Program the RT, Mention Network Builder.

TOOL USED

- Crawl: Python -> tweepy
  - Streaming tweets in real time
- SN build: Python
- Simple Visualization: Gephi.
  - Could expand to D3.



ACKNOWLEDGMENTS

- Dr. Fox
- NSF for grant IIS – 1319578
- IDEAL project
- GRAs – Sunshin Lee and Mohamed Magdy Farag
- Teams – Collection management, Solr, Front end, Topic Analysis
- Class – Discussions, suggestions

QUESTIONS?

THANK YOU