Clustering and Topic Analysis

Presenters: Abigail Bartolome, MD
Islam, Soumya Vundekode

December 6, 2016

CS 5604 Information Storage and Retrieval, Fall 2016
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
Professor Edward A. Fox

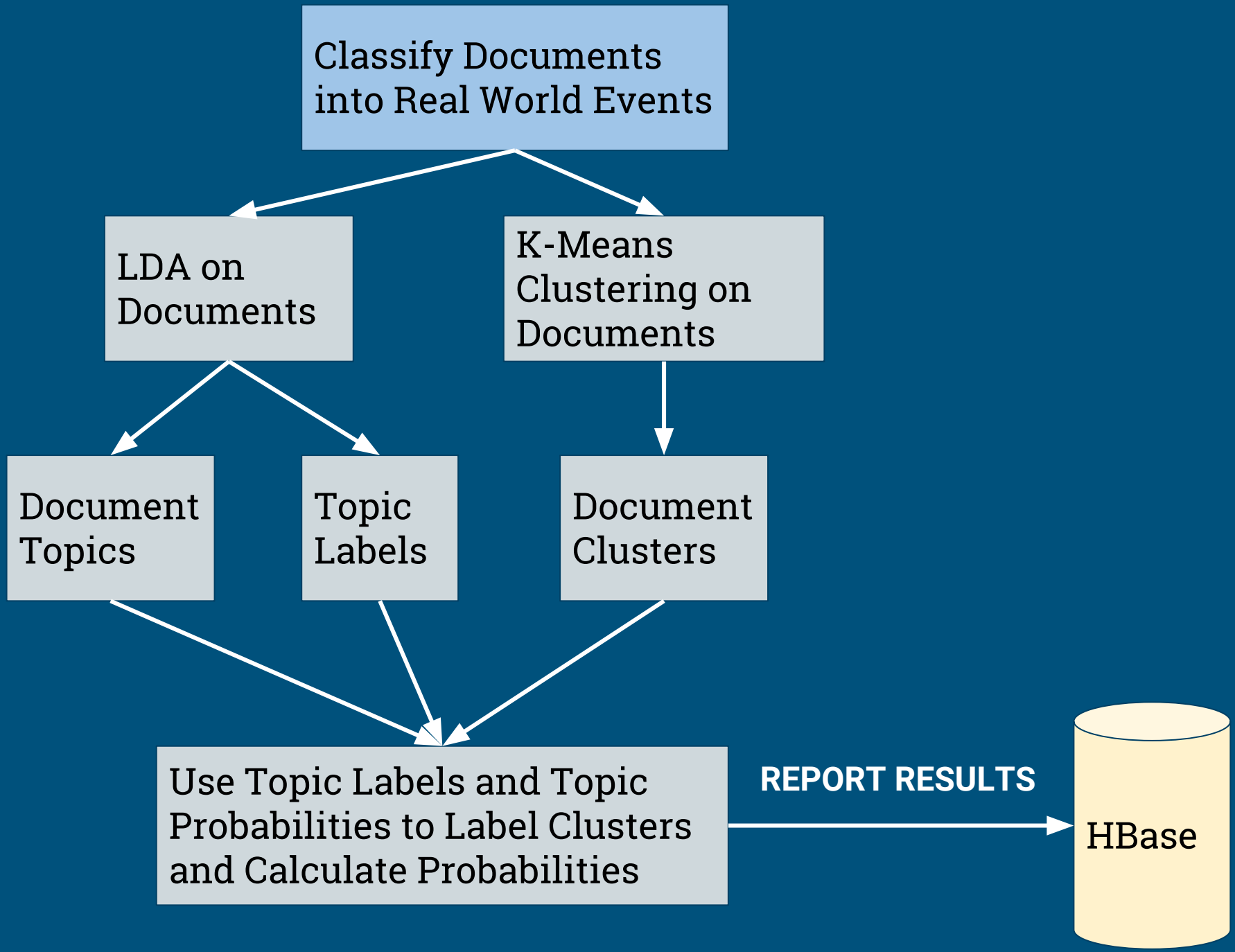
Acknowledgements

- Dr. Edward Fox and GRA Sunshin Lee
- Digital Libraries Research Laboratory (DLRL)
- Integrated Digital Event Archiving and Library (IDEAL) Grant:
  - IIS-1319578
- Global Event and Trend Archive Research (GETAR) Grant:
  - IIS-1619028
- All of the teams in CS 5604 Fall 2016
- Topic Analysis Team of Spring 2016
- Clustering Teams in Fall 2015 and Spring 2016

Goal

To use topic analysis and clustering algorithms on documents about real world events to extract topics discussed regarding the real world events and to find groupings of similar documents

- Pull and Clean Documents, then Store in HDFS
- Topic Analysis
- Clustering
- Evaluation


Classify Documents into Real World Events
LDA on Documents
K-Means Clustering on Documents
Document Topics
Topic Labels
Document Clusters
Use Topic Labels and Topic Probabilities to Label Clusters and Calculate Probabilities
REPORT RESULTS
HBase

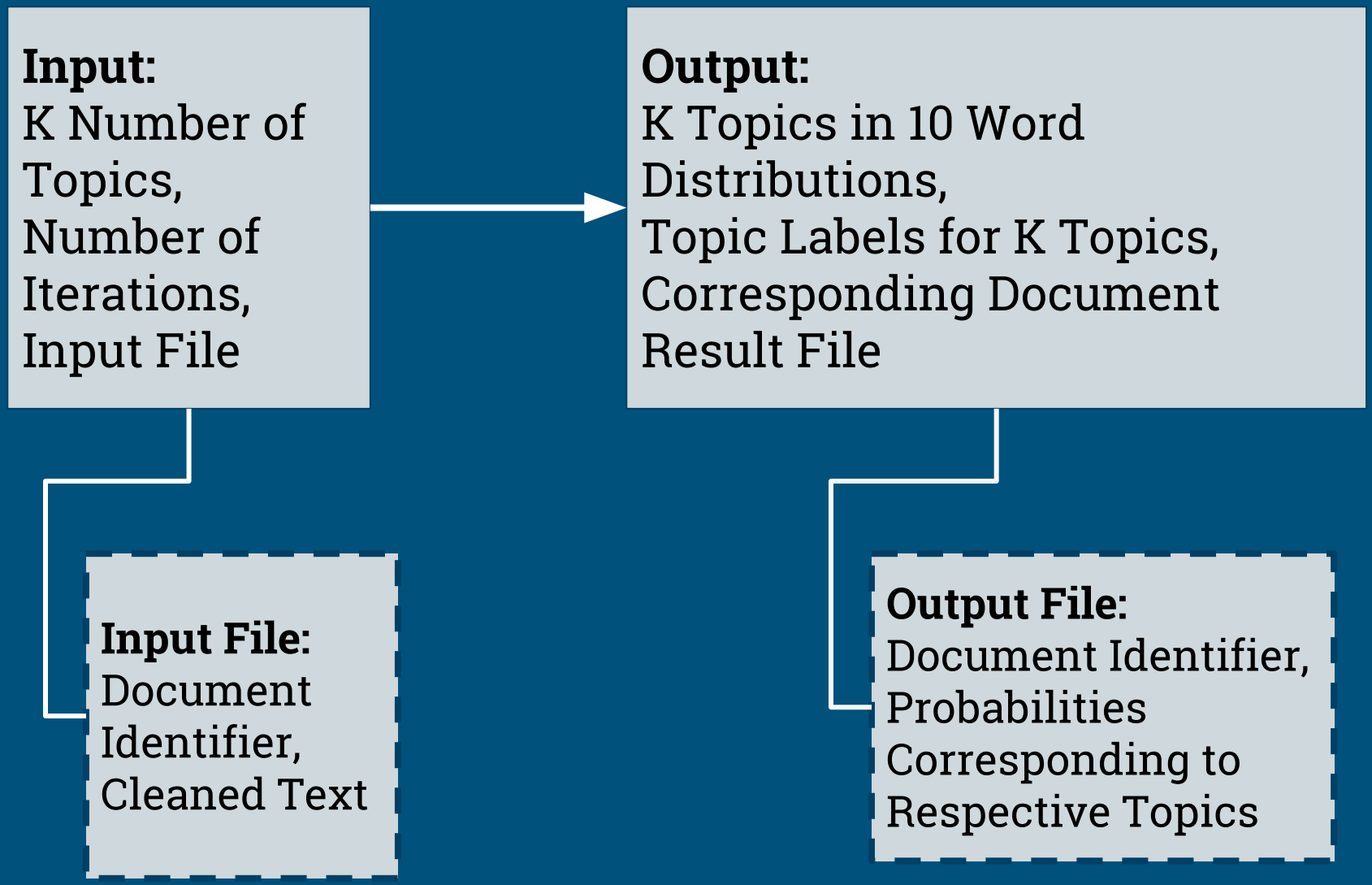
What is a Topic?

Distribution of words that describes a topic

Likelihood that a word belongs to a topic

video,shooting,miami,firefighter,obama,release,accidentally,official,morning,davis	0.0579,0.0420,0.0366,0.0357,0.0290,0.0245,0.0191,0.0183,0.0168,0.0168
shoot,kentucky,rifle,sister,parent,first,group,promptly,despite,control	0.1724,0.1687,0.0601,0.0534,0.0455,0.0394,0.0281,0.0203,0.0183,0.0176
firefighter,ambush,wound,fatally,monday,gunman,standoff,maryland,notify,describe	0.3285,0.0430,0.0393,0.0376,0.0325,0.0175,0.0171,0.0158,0.0145,0.0131
firefighter,shooting,leave,suspect,volunteer,breaking,follow,shooter,baltimore,child	0.0610,0.0372,0.0324,0.0211,0.0210,0.0182,0.0143,0.0119,0.0115,0.0114
shoot,firefighter,three,today,street,harlem,blaze,avenue,battle,think	0.1437,0.0484,0.0262,0.0244,0.0182,0.0174,0.0152,0.0145,0.0140,0.0129
shooting,firefighter,injure,jacksonville,state,target,accuse,station,brother,break	0.1347,0.1097,0.0433,0.0313,0.0272,0.0223,0.0200,0.0178,0.0162,0.0149
police,people,officer,tragedy,worst,recent,memory,suspect,stike,shoot	0.1208,0.0855,0.0534,0.0527,0.0519,0.0504,0.0503,0.0462,0.0452,0.0361
shooting,firefighter,death,woman,charge,arrest,fatal,investigation,facebook,connection	0.1684,0.0523,0.0484,0.0427,0.0386,0.0296,0.0247,0.0217,0.0163,0.0143

Latent Dirichlet Allocation (LDA)



Topic Extraction

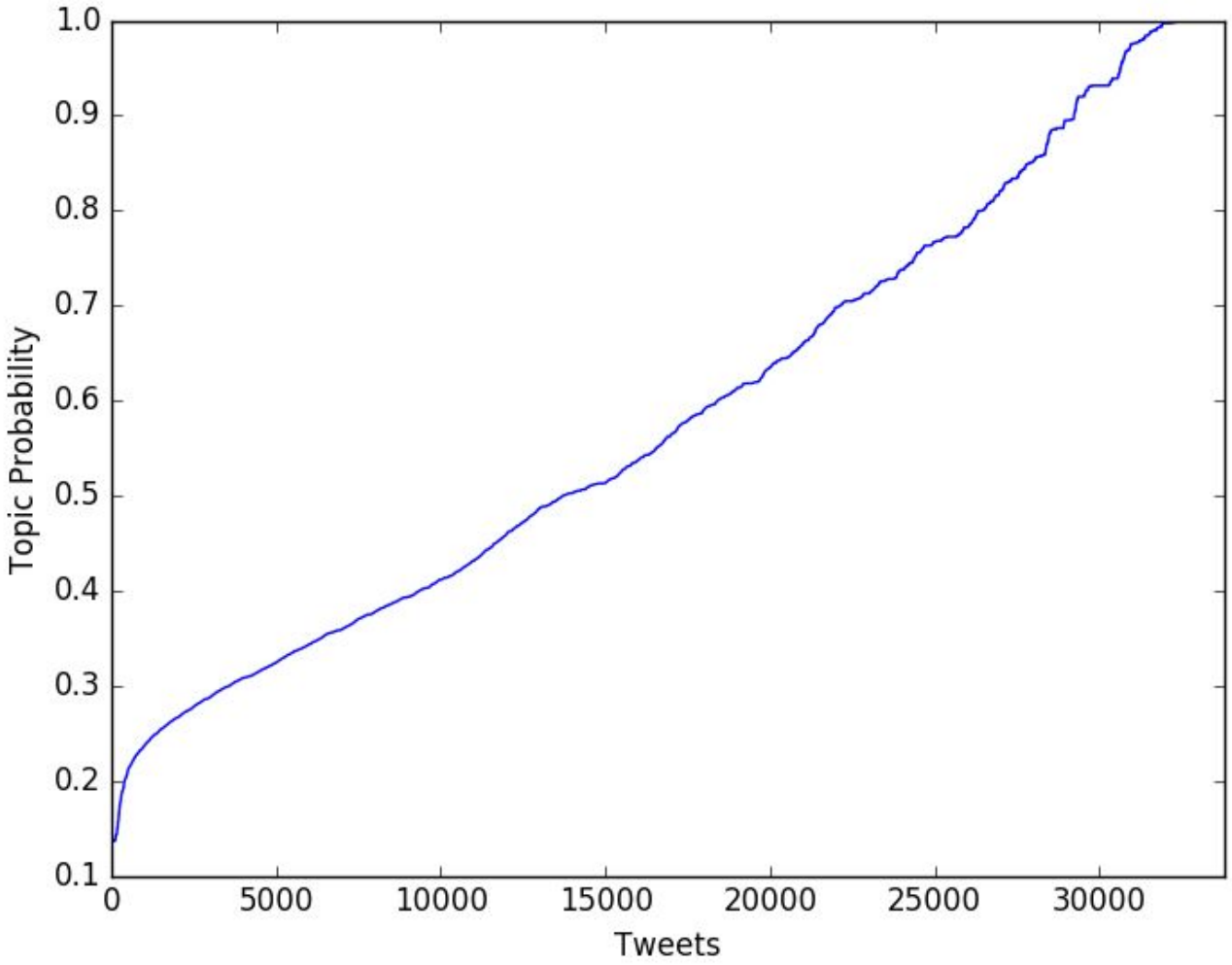
NewYorkFirefighterShooting-- 8 topics	
video	video,shooting,miami,firefighter,obama,release,accidentally,official,morning,davis
shoot	shoot,kentucky,rifle,sister,parent,first,group,promptly,despite,control
firefighter	firefighter,ambush,wound,fatally,monday,gunman,standoff,maryland,notify,describe
shooting	firefighter,shooting,leave,suspect,volunteer,breaking,follow,shooter,baltimore,child
three	shoot,firefighter,three,today,street,harlem,blaze,avenue,battle,think
injure	shooting,firefighter,injure,jacksonville,state,target,accuse,station,brother,break
police	police,people,officer,tragedy,worst,recent,memory,suspect,stike,shoot
death	shooting,firefighter,death,woman,charge,arrest,fatal,investigation,facebook,connection

Manhattan Building Explosion- 4 topics	
harlem	harlem,newyork,newyorkcity,video,building,bigapple,photo,midtown,travel,health
centralpark	centralpark,harlem,taevision,newyork,reverbnation,fashion,music,newyorkcity,singer,design
building	harlem,building,newyork,today,newyorkcity,collapse,photo,blast,least,happy
brooklyn	brooklyn,queens,newyork,harlem,available,active,citibike,bronx,newyorkcity,building

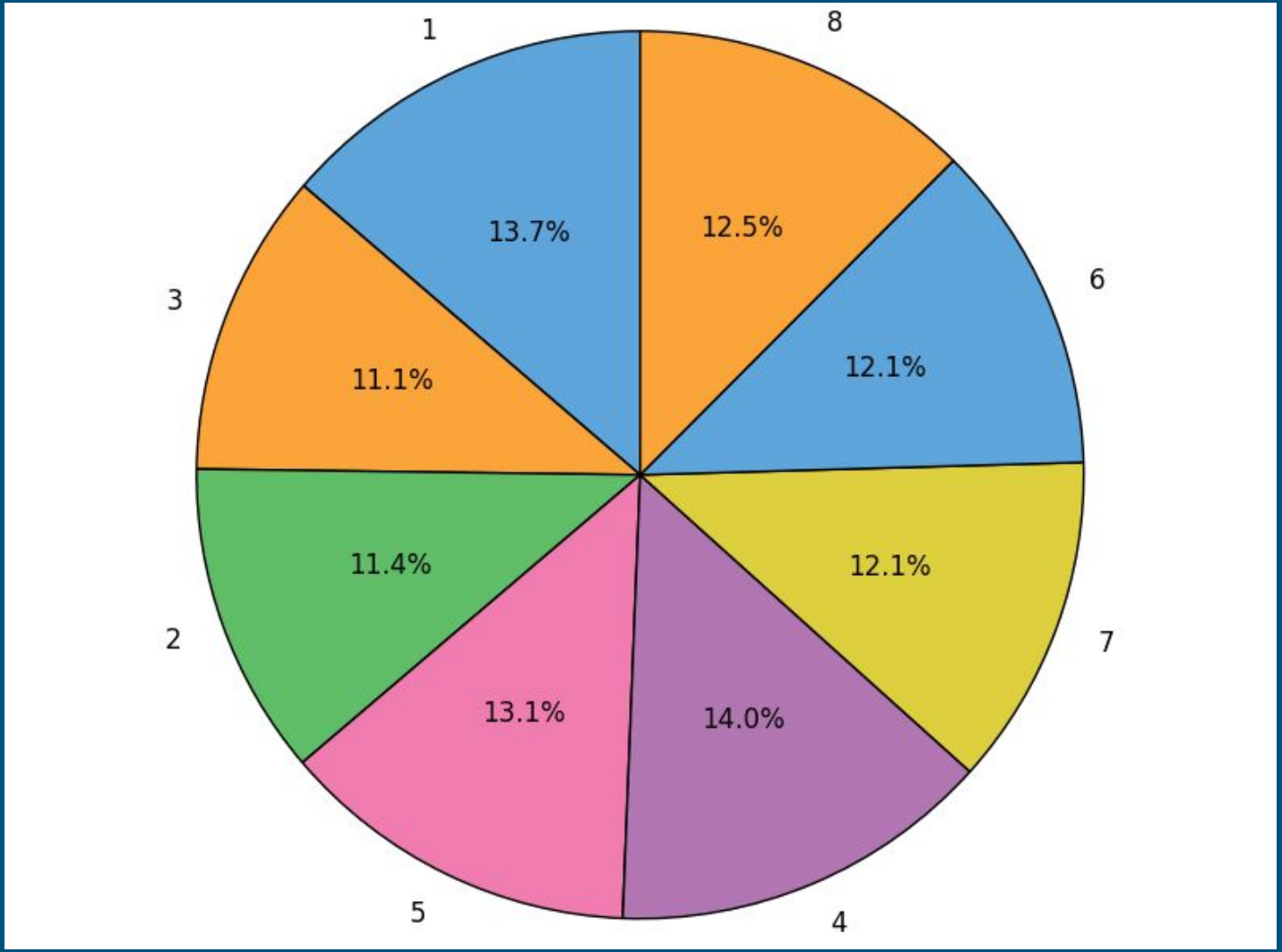
Topic Distributions for New York Firefighter Shooting

Tweet Identifier	P(Topic 1), P(Topic 2), P(Topic 3), P(Topic 4), P(Topic 5), P(Topic 6), P(Topic 7), P(Topic 8)
23-728176223734616064 RT @brendanredmond: @kevburkeie @CitizenGain @frankmcdonald60 @Giulia_Vallone agreeed - New York is taking it as an opportunity	0.2815, 0.0775, 0.0806, 0.1441, 0.1652, 0.0795, 0.1003, 0.0714
43-284087471258615808 'Brother hang tight' wounded New York firefighter told as two colleagues lay ... - CNN	0.0089, 0.0011, 0.2462, 0.0144, 0.0118, 0.7006, 0.0049, 0.0123
43-466357519523532801 RT @BillBishopKHOU: Houston firefighter arrested for allegedly telling co-workers he was "going to start shooting people." #KHOU	0.0224, 0.0013, 0.2190, 0.0625, 0.2721, 0.0675, 0.1503, 0.2049

Likelihood that a Tweet Belongs to its Most Probable Topic for New York Firefighter Shooting

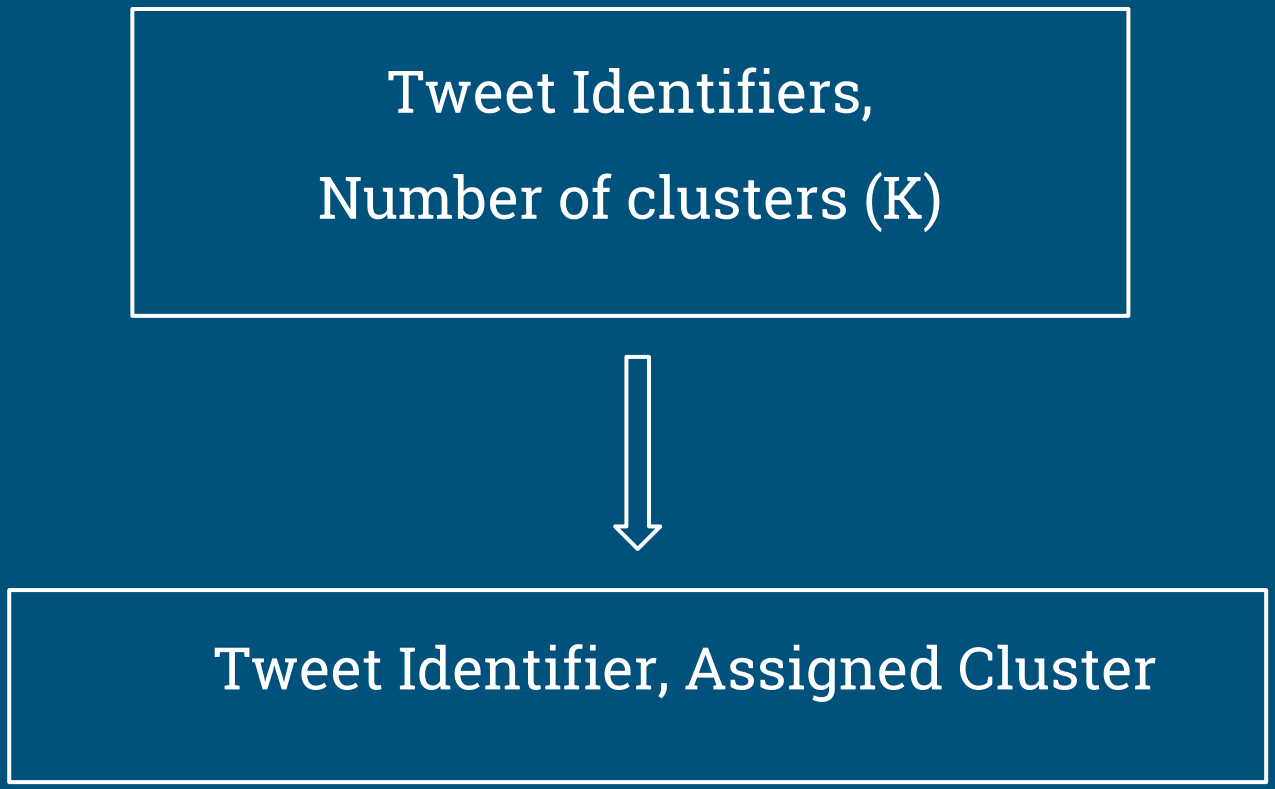


Distributions of Topics for New York Firefighter Shooting



K-Means Clustering

Feature extraction : Word2Vec



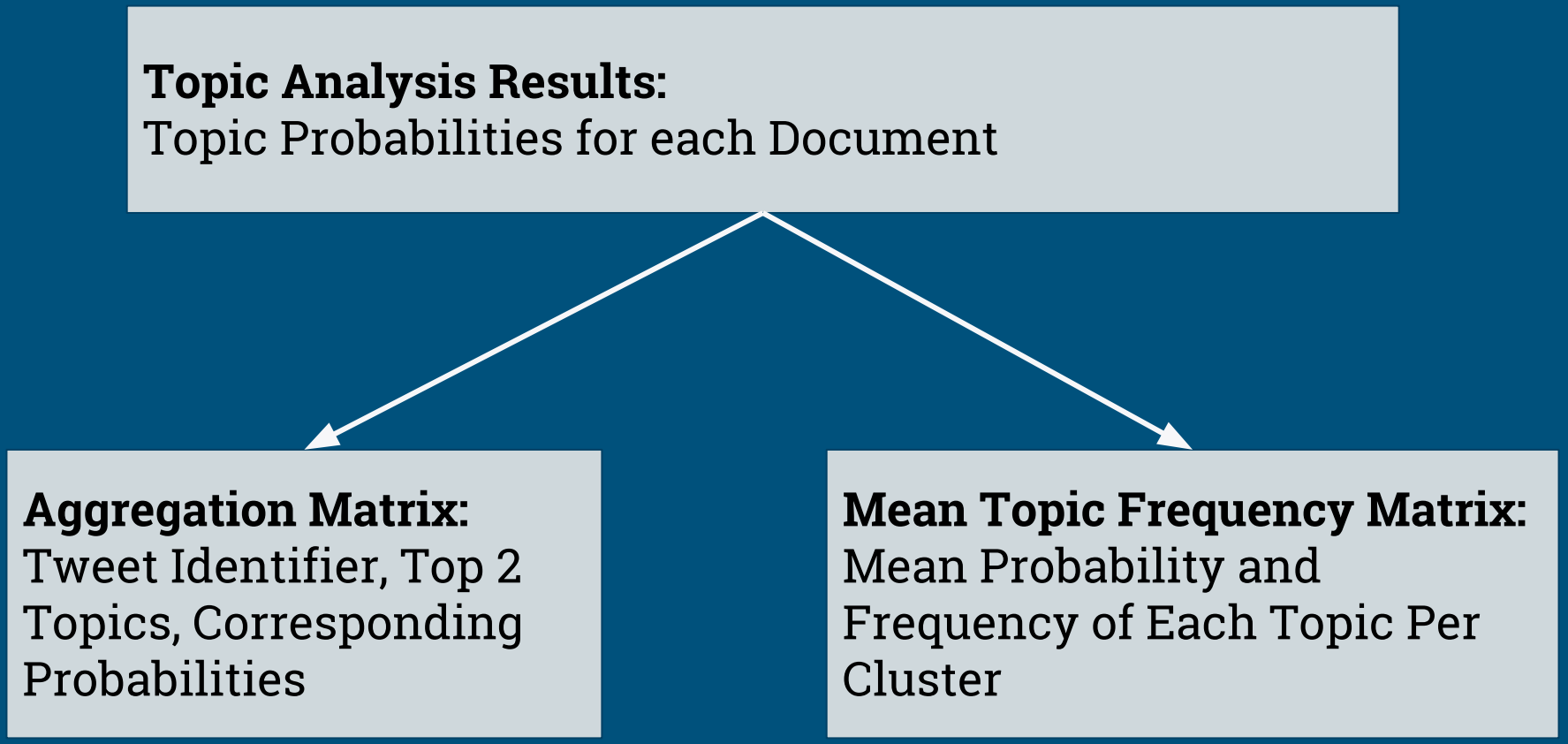
K-Means Clustering - Input

Tweet ID	Tweet
42-279768965641797632	ctshoot suspect brother take custody general question official say
42-280759867420069890	live video new york city mayor michael bloomberg make announcement gun control
42-2815592837856665563	autofollow newtown one one one one fresh heartbreak hearse cri
42-2821687982058536696	shoot way door
42-2876178323159121920	gun show debate organizers tone down displays amid scrutiny chicago lead nation gun violence despite tough
42-291064820155969536	senate take first step tighten gun controls

K-Means Clustering - Output

K = 6	
Tweet ID	**Cluster**
42-2815592837856665 63	0
42-2807598674200698 90	2
42-2876178323159121 92	1
42-2797689656417976 32	3
42-2910648201559695 36	5
42-2821687982058536 96	4

Improving Clustering with Topic Analysis



Cluster	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8
'0'	1304, 0.0123	16, 0.2365	12, 0.2597	76, 0.3677	67, 0.3330	15, 0.2789	1293, 0.9820	21, 0.3886
'1'	2118, 0.4291	1917, 0.4732	1020, 0.3329	3757, 0.4035	2218, 0.3263	869, 0.5131	787, 0.3231	502, 0.2842
'2'	1737, 0.3432	644, 0.5752	925, 0.4522	1307, 0.2970	1284, 0.4866	449, 0.5321	211, 0.4113	219, 0.3853
'3'	2966, 0.3934	544, 0.1530	7167, 0.2850	3211, 0.3421	1646, 0.3275	4738, 0.4051	3181, 0.2583	4939, 0.448
'4'	1175, 0.3495	4728, 0.4354	1049, 0.3444	969, 0.2935	3194, 0.2907	829, 0.3041	2119, 0.4327	1165, 0.3759
'5'	62, 0.3107	327, 0.1498	36, 0.2762	68, 0.240	561, 0.6632	23, 0.3062	98, 0.2303	13, 0.2706

Cluster Probabilities

Cluster - Most frequent topics	
Cluster	**Topics**
'0'	Topic 1, Topic 7
'1'	Topic 4, Topic 5
'2'	Topic 1, Topic 4
'3'	Topic 3, Topic 8
'4'	Topic 2, Topic 5
'5'	Topic 5, Topic 2

Cluster Probabilities

Cluster probability for each tweet = prob(Ta) + prob(Tb)

Where T_k : Probability of tweet belonging to Topic K

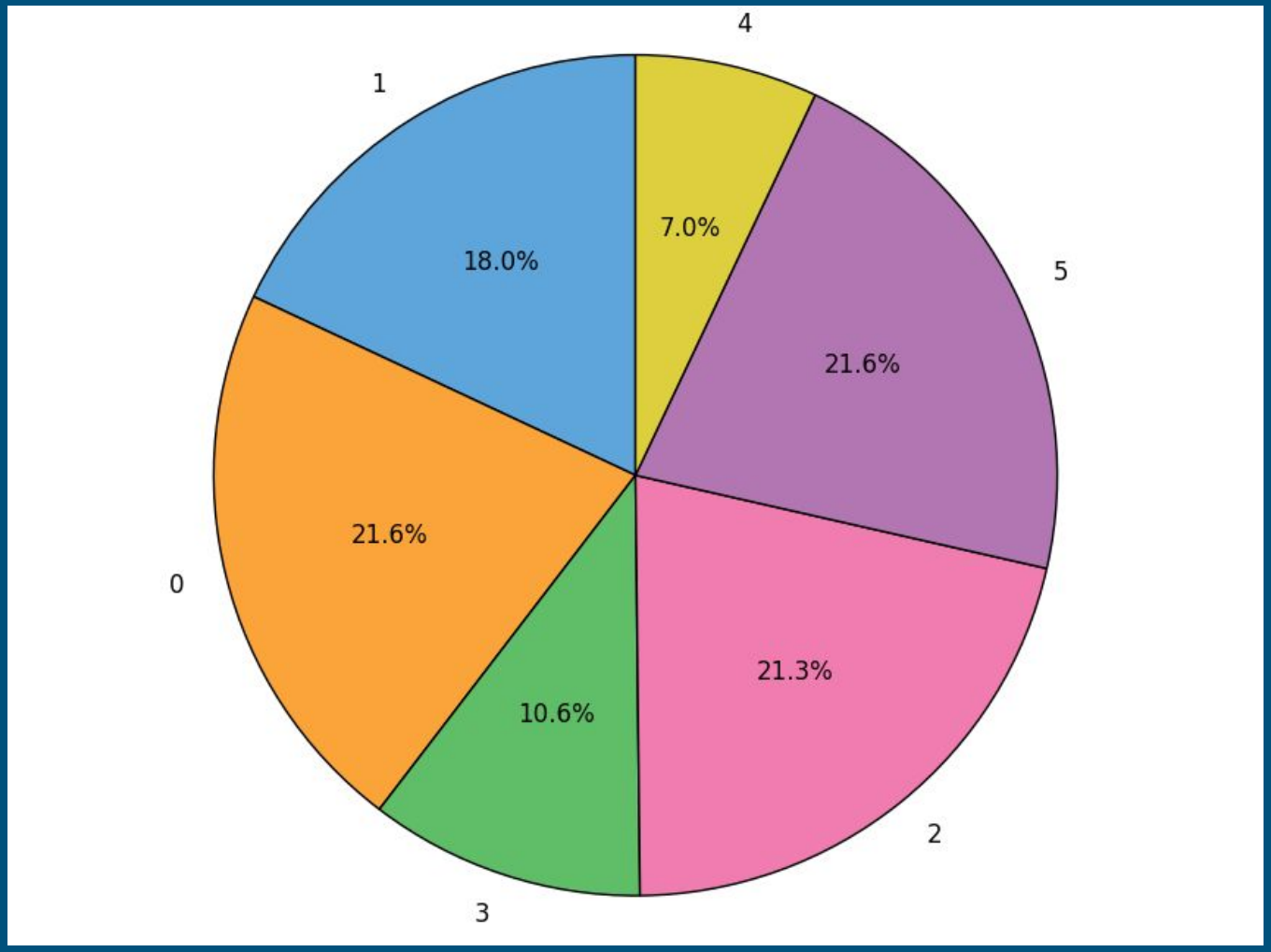
Tweet ID	Cluster	Cluster Probability
'45-68963932753867571 2'	2	0.962096525408969 3
'43-54770953526863462 5'	5	0.27724837702230876

Automated Cluster Labeling

NewYorkFirefightershooting	
Cluster 0	video,shooting,police,people
Cluster 1	shooting,firefighter,three,shoot
Cluster 2	video,shooting,firefighter
Cluster 3	firefighter,ambush,death,shooting
Cluster 4	shoot,kentucky,three
Cluster 5	three,shoot,kentucky

Use labels of the most frequent topics of the cluster for cluster labeling

K-Means Clustering - Kentucky Accidental Child Shooting



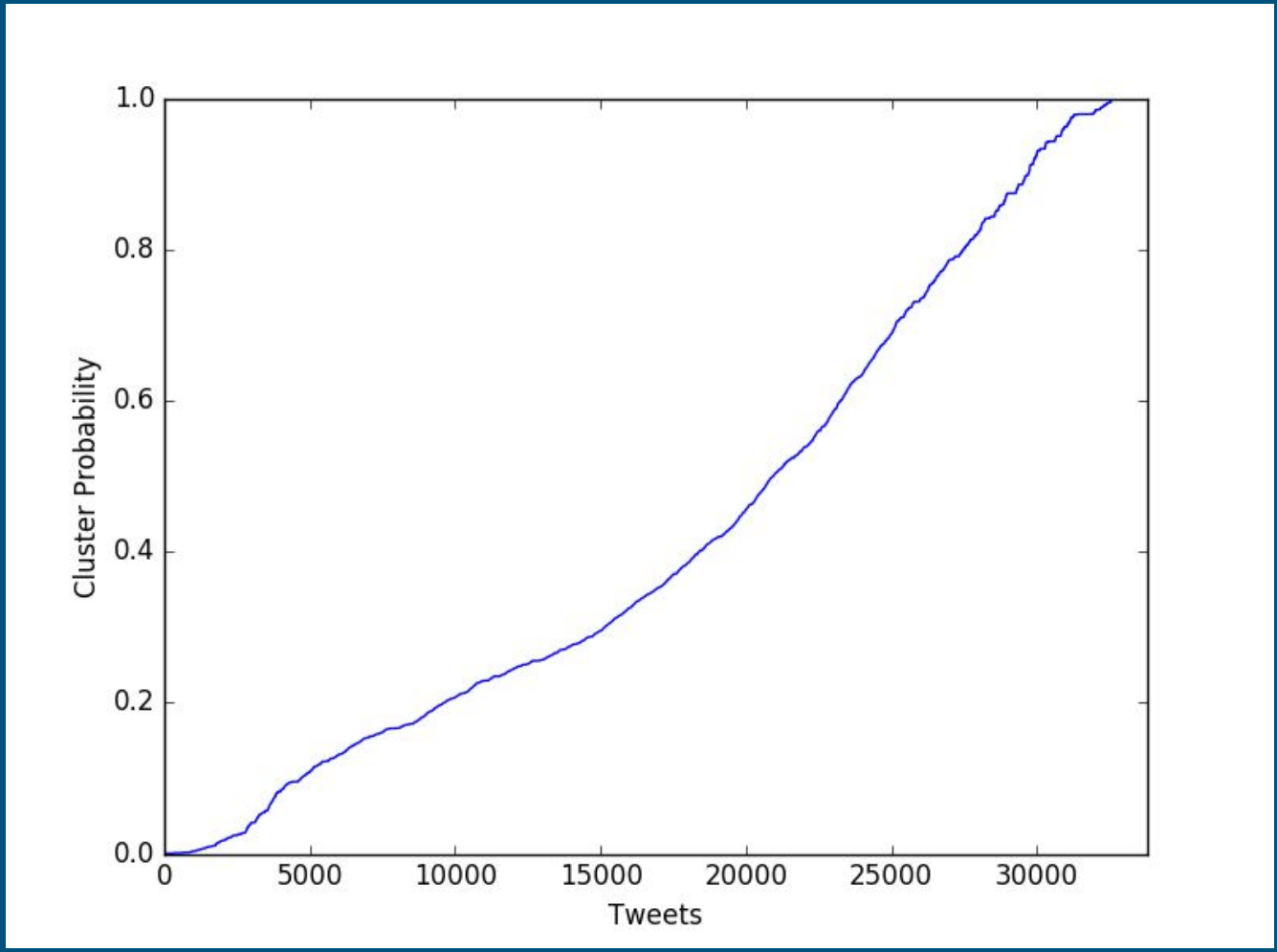
Cluster Evaluation

- Used cluster probabilities of tweets to pick K
- Best case : Highest mean probability of tweets belonging to their assigned clusters.

Experiments :

- K = 4, 5, 6
- Picked most efficient value of K for each collection

Cluster Probabilities

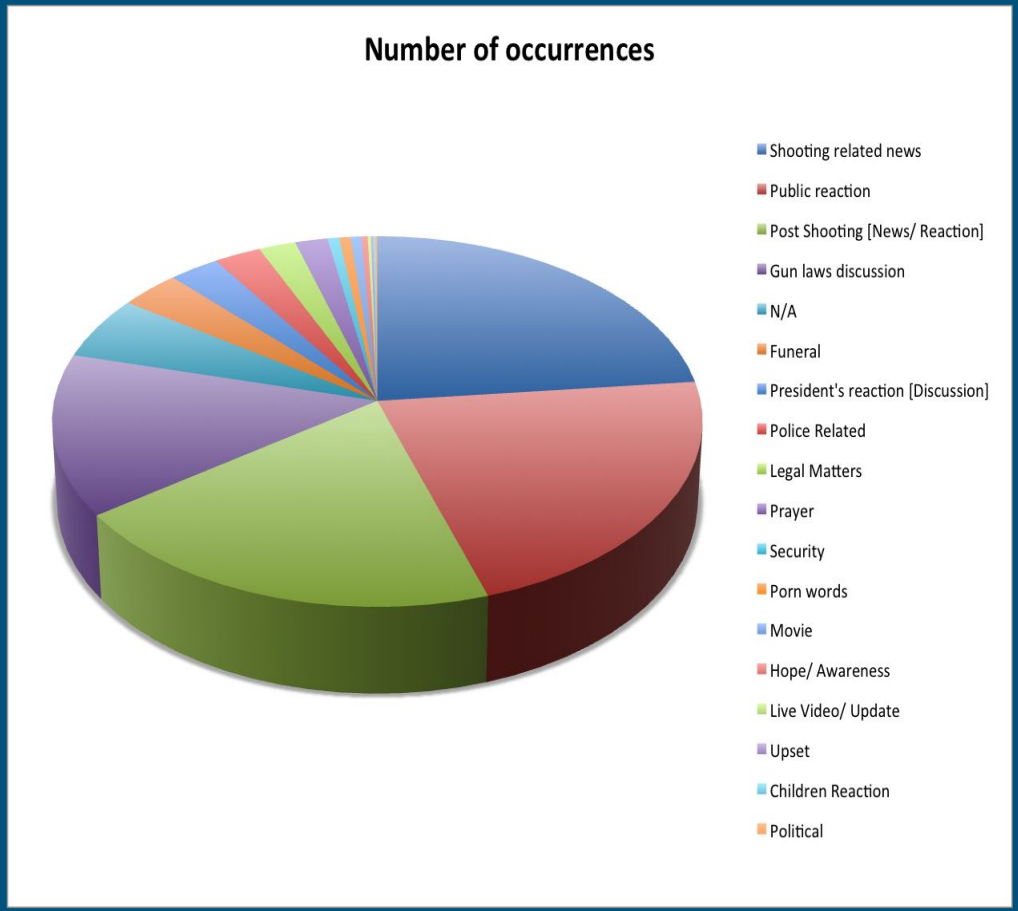


Cluster Labeling (Manually)

- Manually group documents into subsets by similarity
- Logically create clusters that are logically similar
- Internal documents are as similar as possible
- Internal documents are dissimilar from documents in other clusters

Manual Cluster Labeling Results (Sandy Hook Elementary School shooting)

Cluster Name	Number of occurrences	Collections (%)
Shooting related news	7641	23.32
Public reaction	7120	21.73
Post Shooting [News/ Reaction]	6556	20.01
Gun laws discussion	4615	14.08
N/A	1812	5.53
Funeral	1090	3.33
President's reaction [Discussion]	916	2.80
Police Related	856	2.61
Legal Matters	673	2.05
Prayer	590	1.80
Security	211	0.64
Porn words	206	0.63
Movie	195	0.60
Hope/ Awareness	114	0.35
Live Video/ Update	59	0.18
Upset	40	0.12
Children Reaction	37	0.11
Political	37	0.11



Conclusions

- Extracted topics from collections of tweets about 9 real world events
- Automatically labeled the topics and mapped the topic probabilities back to each tweet
- Clustered tweets about 9 real world events and used the topic labels and probabilities to determine cluster labels and cluster probabilities
- Compared clustering results to a collection that was manually clustered

Results

Real World Event	Number of Topics	Topics
NewYorkFirefighterShooting	8	Video, Shoot, Firefighter, Shooting, Three, Injure, Police Death
KentuckyAccidentalChildShooting	8	Field, Connecticut, Police, Throw, Wisconsin, Shoot, Sister
NewtownSchoolShooting	8	School, Kentucky, Harlem, Victim, Newtown, Obama, Report, Elementary
ManhattanBuildingExplosion	4	Harlem, Centralpark, Building, Brooklyn
ChinaFactoryExplosion	8	People, Sandy, Computer, Media, Black, Kentucky, Police, Hurricane
TexasFertilizerExplosion	10	Federal, Cause, Firefighter, Report, Boston, Video, Explode, Obama, First, Blast
HurricaneSandy	8	Manhattan, Amazing, Newyork, Isaac, Skyline, Speak, Brooklyn, Latest
HurricaneArthur	8	Sandy, Power, Merlin, Texas, Still, Minha, North, Missingmerlin
HurricaneIsaac	8	Sandy, School, Please, Power, Storm, History, Since, Victim

Real World Event	Number of Clusters	Clusters
NewYorkFirefighterShooting	6	“Video,shooting,police,people”,“shooting,firefighter,three,shoot”,“video,shooting,firefighter”,“firefighter,ambush,death,shooting”,“shoot,kentucky,three”,“three,shoot,kentucky”
KentuckyAccidentalChildShooting	6	“Police,trooper,connecticut,people”,“connecticut,people,throw,shoot”,“wisconsin,shoot,throw”,“sister,shoot,guard”,“shoot,guard,sister”,“field,percent,wisconsin,shoot”
NewtownSchoolShooting	6	“Report,school,newtown”,“harlem,school,report”,“school,newtown,victim”,“obama,school,newtown”,“kentucky,police,harlem,school”,“victim,school,newtown”
ManhattanBuildingExplosion	6	“centralpark,harlem,brooklyn,queens“,“harlem,newyork,building“,“building,harlem,centralpark”,“newyork,centralpark“,“harlem,newyork,brooklyn,queens”,“building,harlem,newyork”
ChinaFactoryExplosion	5	“media,think,hurricane,sandy”,“kentucky,state,computer”,“black,white,computer,kentucky”,“people,romney,police,rifle”,“sandy,hurricane”

Real World Event	Number of Clusters	Clusters
TexasFertilizerExplosion	6	“first,responder,report,massive”,“report,massive,firefighter,investigation”,“boston,bombing,firefighter,investigation”,“federal,still,firefighter,investigation”,“explode,firefighter,report,massive”,“cause,criminal,blast,facility”
HurricaneSandy	6	“skyline,manhattan,latest”,“newyork,manhattan,amazing”,“speak,manhattan,latest”,“skyline,manhattan,amazing”,“isaac,manhattan,speak”,“manhattan,harlem,speak”
HurricaneArthur	6	“minha,lindo,still,storm”,“sandy,death,texas”,“missingmerlin,merlin”,”north,storm,still”,”texas,sandy,power,canada”,”missingmerlin,merlin,texas,sandy”
HurricaneIsaac	6	“storm,right,power,sandy”,”school,sandy,history,deadliest”,”victim,sandy,history,deadliest”,”power,sandy,school”,”storm,right,history,deadliest”,”since,sandy,school”