Topic Analysis

Presenters: Radha Krishnan, Sneha Mehta
April 28, 2016

CS5604, Information Retrieval and Storage, Spring 2016
Virginia Polytechnic Institute and State University
Blacksburg VA
Professor: Dr E. Fox

Acknowledgements

- Prof. Edward Fox and GRA's for IDEAL (Sunshin & Mohamed)
- Digital Libraries Research Laboratory (DLRL)
- NSF for grant IIS - 1319578, Integrated Digital Event Archiving and Library (IDEAL)
- All the teams in CS5604

Outline

- Goals
- Introduction
- Design and implementation
- Experiments on collections
- Results on tweet collections
- Comparison of tweet and web page results
- Evaluation
- Conclusions and future work

Goals

- Create meaningful topic models for tweet and webpage collections
- Provide intuitive labels for the topics to enable faceted search

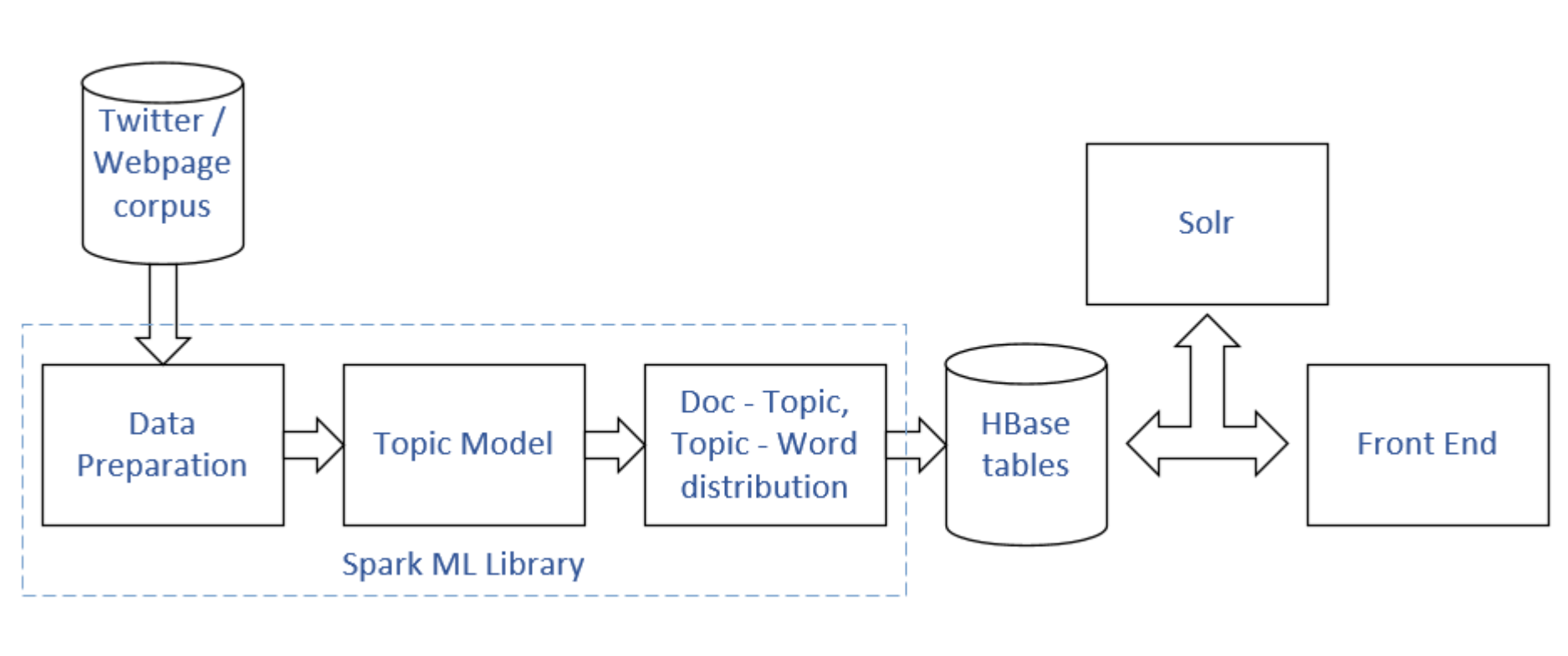
Introduction

- LDA is a generative statistical model
- Tweets/Webpages can be thought of as mixtures of topics
- Each word in a tweet/webpage is attributable to one or more topics

Output of LDA:

- Topic-word matrix – Word distribution for each topic
- Document-topic matrix - Topic distribution for each document

Design


Twitter / Webpage corpus
Data Preparation
Topic Model
Doc - Topic, Topic - Word distribution
Spark ML Library
HBase tables
Solr
Front End

Implementation

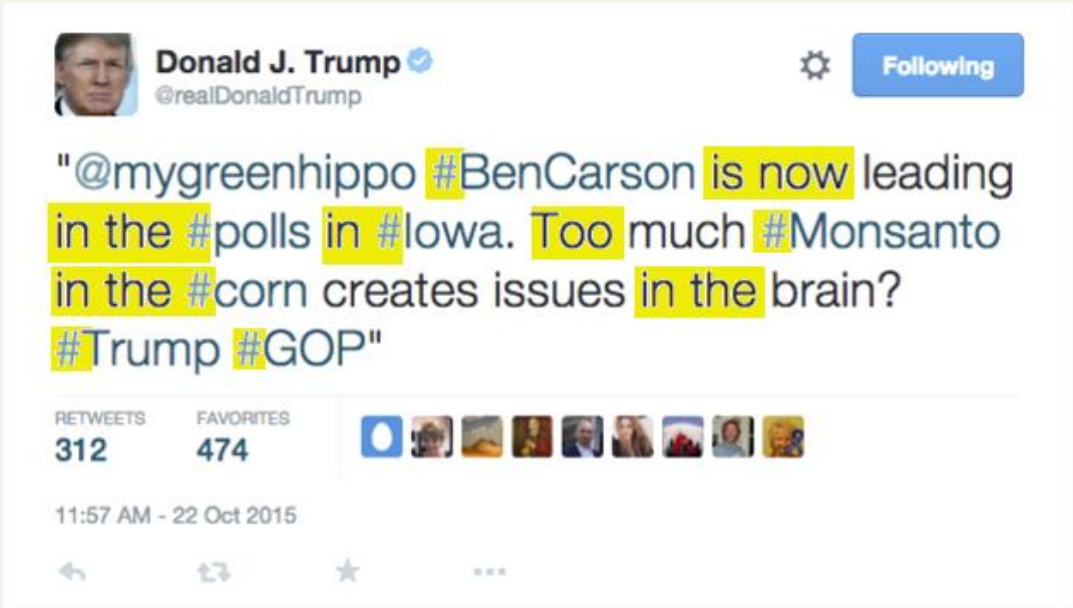
- Workflow:
  - Data preparation
  - Topic extraction
  - Topic labeling
  - Loading results into Hbase

- LDA in Scala using Spark Mllib
- SBT for packaging into JAR file
- spark-submit to deploy on cluster

Data Preparation

- Based on test runs, devised our own set of data pre-processing steps
- Removed stop words, most frequent words (25), words < 4 characters and non-alphabetic letters
- Also, created a custom stop words list to remove

means, while, because, would etc.



Topic Extraction

- Five topics extracted for each tweet collection
- Ten words with highest probabilities for each topic extracted

Collection	Topics
WDBJ shooting	Journalists, Roanoke, victims, shooting, virginia

Topic	Words
Journalists	Journalists, shooting, newscast, victims, tribute, heartbroken, condolence, watch, everytown, moment

Topic Labeling

- Word with highest probability chosen as the topic label
- Topics are given unique labels

Topics	Words	Labels
Topic 1	journalists, shooting, newscast, victims…	Journalists
Topic 2	roanoke, prayers, families, friends…	Roanoke
Topic 3	journalists, shooting, reporter, police…	Shooting

Loading data into HBase

- Topics are sorted based on probabilities

- Created a document - topic column family in HBase

Rowkey	Tweet_Topics		Webpage_Topics	
Tweet / Webpage ID	Labels	Probabilities	Labels	Probabilities

- Created topic - word table in HBase

Rowkey	Collection_ID	Words	Probabilities
Topic_Label			

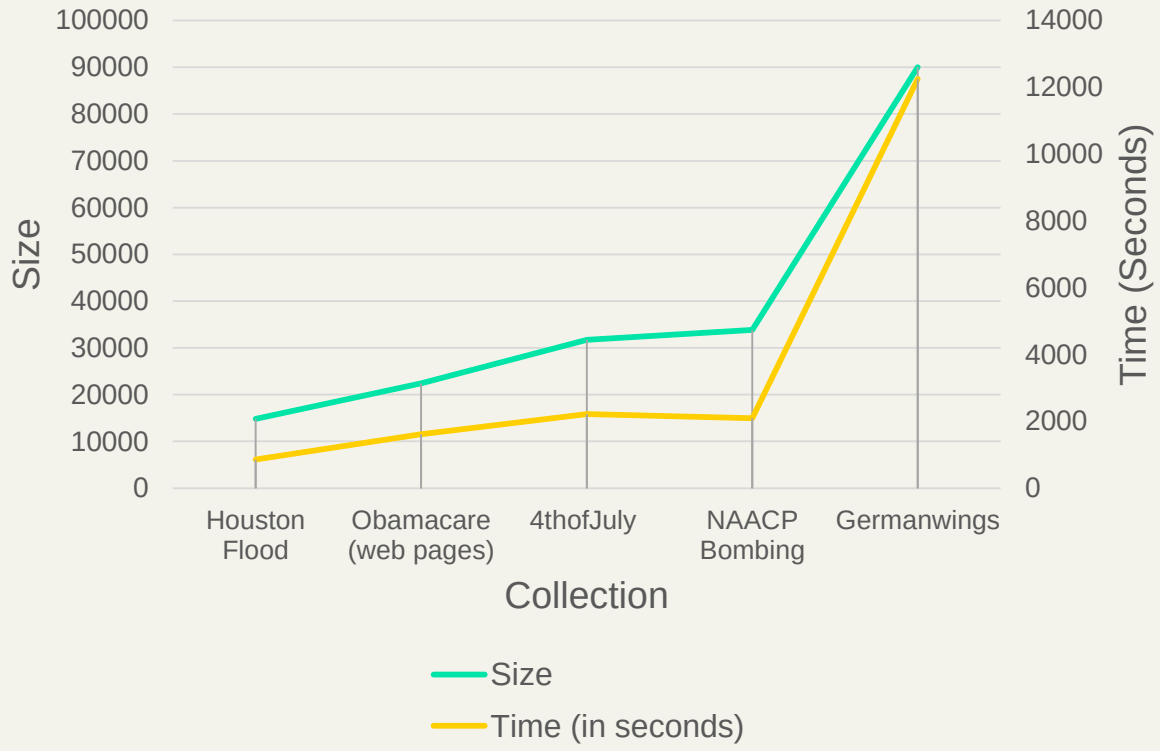
Experiments on Collections

- Ran LDA on both Spark standalone and Yarn mode
- Standalone mode supports 100 to 200 iterations in LDA

- Standalone mode results:

Tweet Collection	Size	Time (in seconds)
Houston Flood	14813	860
4thofJuly	31743	2220
NAACP Bombing	33865	2094
Germanwings	90040	12253

Webpage Collection	Size	Time (in seconds)
Obamacare	22393	1615



Experiments on Tweets

- Increase in number of iterations increases quality of topics

20 iterations		100 iterations	
Topic A	Topic B	Topic A	Topic B
reporter	heartbroken	heartbroken	reporter
heartbroken	reporter	condolence	watch
condolence	condolence	survivor	westandwithwdbj
police	survivor	community	photographer
survivor	police	terrorism	condition
suspect	watch	obama	fatal
westandwithsdbj	condition	control	moment

Results on Tweet Collections

- Ran LDA for 100 iterations
- Found 5 topics on each collection

Collection Name	Topics
WDBJ Shooting	journalists, shooting, victims, Roanoke, virginia
NAACP Bombing	naacp, naacpbombing, terrorism, michellemalkin, deray
4thofJuly	american, independenceday, fireworks, happy, america
Germanwings	germanwings, ripley, world, copilot, crash
Houston Flood	houstonflood, texas, texasflood, flood, billbishopkhou

Comparison of Tweet and Webpage results

- Topics for #Obamacare
  - Identified 4 topics in webpage collection
  - 5 topics in tweet collection

Collection Name	Webpage_Topics	Tweet_Topics
#Obamacare	insurance, plans, companies, court	obamacare, uninsured, health, arrived, pjnet

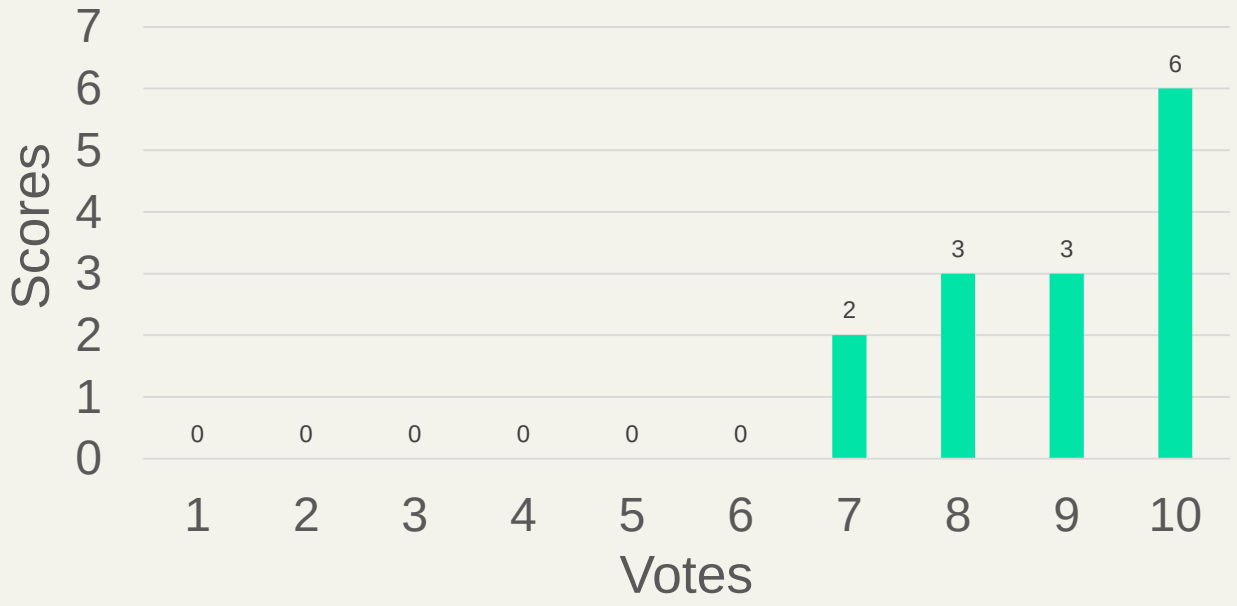
Comparison of Tweet and Webpage results

- Top words for each topic in #Obamacare collection

Topics	Tweet_Topic_Words	Webpage_Topic_Words
Topic 1	health, insurance, people, healthcare, medicaid, congress…	plans, fixed, indemnity, insurance, people, judge, health, administration…
Topic 2	uninsured, barackobama, americans, thanks, realdonaldtrump, president …	percent, increases, federal, health, insurers, affordable, state, claims ..
Topic 3	obamacare, repeal, scottwalker, prolife, little, supports, sisters…	court, religious, mandate, supreme, schools, baptist, organizations, christian…
Topic 4	arrived, abcnews, antichrist, doomsday, cometcoreinc, comet, republicans …	companies, books, playlist, cancer, medical, health, medicine, problem…
Topic 5	pjnet, cruzcrew, benghazi, teaparty, tedcruz, healthcare, uniteblue …	

Evaluation – (Part A)

- Conducted a user study with 2 questions
  - Do these words describe WDBJ shooting?
    - Out of 17 participants, 3 didn't answer



Evaluation – (Part B)

- Word intrusion - measure to identify quality of topics
- Which of these words do not belong to the group?

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
word/ votes	word /votes	word /votes	word /votes	word/votes
Virginia / 4	Alison / 6	Roanoke / 1	Shooting / 0	Victims / 0
Journalists / 1	Shooting / 0	Prayers / 0	Vicki / 7	Shooting / 0
Shooting / 0	Video / 4	Families / 1	Gardner / 4	Gunman / 0
Newscast / 3	Parker / 5	Friends / 1	Survivor / 0	Husband / 9
Heartbroken / 2	Reporter / 0	Unbelievable / 2	Condition / 4	Blame / 2
Condolence / 0	Police / 0	Ripalisonparker /4	People / 2	Foxnews /4
Tribute / 1	Watch / 5	Ripadamward / 7	Victim / 0	westandwithwdbj/ 3
Everytown / 11	Christiansburg / 2	Archct / 11	Today / 8	Brianstelter / 4
Community / 0	Moment / 9	Shooting / 0	Thoughts / 4	Tragic / 0
Support / 2	Fatal / 0	Suspect / 0	Virginia / 2	Onair / 6

Lessons Learnt

Problems	Solutions
Repeating words in topics	Increase iterations
Noisy words in topics	Add to stop word list
Topic similarity / dissimilarity	Decrease / increase topics

- More data – better quality of topics

Conclusions and Future Work

- Conclusions
  - Extracted meaningful topics from each tweet and web page collections
  - Devised an automated technique to generate intuitive labels for each topic
  - Evaluation results confirm the good quality of our topic model

- Future work
  - Create topic labels using metadata like hashtags, titles etc.
  - Using learning based approach for labeling - Word2Vec
  - Implement Hierarchical / Online LDA models

Questions ?

Thank you!!!