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
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

Cluster sizes

<table>
<thead>
<tr>
<th>Cluster Sizes</th>
<th>TF-IDF</th>
<th>Word2vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
WSSE BASED EVALUATION

WSSE vs. Clusters (z700)
**RUNNING OVER DIFFERENT TWEET COLLECTIONS**

<table>
<thead>
<tr>
<th>Collection Clusters</th>
<th>541 #NAACPBombing</th>
<th>602 #Germanwings</th>
<th>668 #houstonflood</th>
<th>686 #Obamacare</th>
<th>694 #4thofjuly</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1374</td>
<td>31189</td>
<td>283</td>
<td>5333</td>
<td>6627</td>
</tr>
<tr>
<td>1</td>
<td>7504</td>
<td>26297</td>
<td>899</td>
<td>36070</td>
<td>5529</td>
</tr>
<tr>
<td>2</td>
<td>18375</td>
<td>5785</td>
<td>2605</td>
<td>16899</td>
<td>762</td>
</tr>
<tr>
<td>3</td>
<td>3166</td>
<td>11979</td>
<td>3672</td>
<td>69555</td>
<td>13663</td>
</tr>
<tr>
<td>4</td>
<td>1009</td>
<td>8328</td>
<td>1488</td>
<td>36383</td>
<td>1427</td>
</tr>
<tr>
<td>5</td>
<td>2437</td>
<td>6459</td>
<td>5884</td>
<td>19302</td>
<td>3961</td>
</tr>
</tbody>
</table>
RUNNING OVER DIFFERENT TWEET COLLECTIONS

- 541 #NAACPBombing
- 602 #Germanwings
- 668 #houstonflood
- 686 #Obamacare
- 694 #4thofjuly
## HBASE SCHEMA DESIGN

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Cluster no.</th>
<th>Cluster label</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>String (Tweet-id/URL)</td>
<td>Integer</td>
<td>String</td>
<td>Float</td>
</tr>
</tbody>
</table>
COLLECTIONS EVALUATED

- Tweets
  - 602, 668, 694, 541

- Webpages
  - 686 (TODO)
CLUSTERING EVALUATION USING TOPIC ANALYSIS DATA

1. Cluster labels
2. Hierarchy of mixed topics and clusters
3. Probability of a document belonging to the cluster

Outcome

- Combine document matrices
- Create topic frequency matrix per cluster
- Extract relevant topics per cluster using mean, frequency, deviation

Extract relevant topics per cluster using mean, frequency, deviation
EXPLANATION

- Combine document matrices
- Create topic frequency matrix per cluster
- Extract relevant topics per cluster using mean, frequency, deviation

How many times a topic occurs in a cluster?

<table>
<thead>
<tr>
<th>Cluster no</th>
<th>Topic frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T1=2, T5=2, T2=1, T3=1</td>
</tr>
<tr>
<td>2</td>
<td>T1=3, T3=3, T5=1, T2=5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster no</th>
<th>Topics per cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T3, T2</td>
</tr>
<tr>
<td>2</td>
<td>T1, T2</td>
</tr>
</tbody>
</table>
## Results

Collection 602 - tweets

<table>
<thead>
<tr>
<th>Cluster no (wordToVec)</th>
<th>Topics</th>
<th>Topic words</th>
<th>Cluster labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Topic 3, Topic 4</td>
<td>crash, germanwings, pilot, victims, plane, Lufthansa</td>
<td>Germanwings crash and pilot Lubitz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>germanwings, lubitz, copilot, flight, playing</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Topic 1, Topic 5</td>
<td>germanwings, riple, believe, click, bigarevel</td>
<td>Germanwings news</td>
</tr>
<tr>
<td></td>
<td></td>
<td>germanwings, world, charliehebdo, lives, garissa</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Topic 4, Topic 5</td>
<td>germanwings, lubitz, copilot, flight, playing</td>
<td>Germanwings crash news</td>
</tr>
<tr>
<td></td>
<td></td>
<td>germanwings, world, charliehebdo, lives, garissa</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Topic 1, Topic 4</td>
<td>germanwings, riple, believe, click, bigarevel</td>
<td>Pilot Lubitz of Germanwings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>germanwings, lubitz, copilot, flight, playing</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Topic 1, Topic 2</td>
<td>germanwings, riple, believe, click, bigarevel</td>
<td>Happenings and accidents</td>
</tr>
<tr>
<td></td>
<td></td>
<td>germanwings, copiloto, victias, accidente, vuelo</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Topic 1, Topic 4</td>
<td>germanwings, riple, believe, click, bigarevel</td>
<td>Pilot Lubitz of Germanwings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>germanwings, lubitz, copilot, flight, playing</td>
<td></td>
</tr>
</tbody>
</table>
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
1. Start from follower counts:
   \[ IF_{\text{account}} = \frac{\text{Number of followers}}{\text{Maximum follower count in SN}} \]

2. Tweet IF:
   \[ IF_{\text{tweet}} = \text{SUM (RT} \times IF_{\text{account}}\text{)} \text{ for each RT} \]

3. Account IF:
   \[ IF_{\text{account}} = \text{SUM (IF}_{\text{tweet}} \text{)} / \text{Tweet count} \]

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

Key Activities

Twitter Crawling: 1. User Information; 2. RT counts of each tweets.
Twitter connections summarization: find the RT and Mention (@) between accounts.

Milestones

Twitter Crawler: 1. User Information; 2. RT counts of each tweets.
Program the RT, Mention Network Builder.

Work

Data Collection → Social Network Building → Social Network Clustering → Important Factor Calculation

Nodes Grouping Using Clustering result.
Importance factor calculation based position in the social network (centroid or border of a cluster)

Calculation of the nodes-edge networks: nodes – accounts; edge – follow, RT, mention.
Social Network Drawing.

Clustering Based on Social Network Structure: Graph Clustering Given Nodes and Edges.

Social Network Visualization.
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