Research Pipeline

Interesting question

Collect data

Analyze data (fit statistical models)
Research Pipeline

Interesting question

Collect data

Analyze data (fit statistical models)
How do we make sense of this?
Some tools

- Topic Modeling

- Event Detection
Some tools - limitations

- Topic Modeling - very short documents

- Event Detection - detect only the onset of events
Our focus in this work...
Our focus

- Topic Modeling: very short documents
- Event Detection: detect only the onset of events
Model

\[ X \approx WH \]

\[
\begin{array}{ccc}
\text{doc-term} & \text{doc-top} & \text{top-term} \\
\downarrow & \downarrow & \rightarrow \\
\end{array}
\]
Model

\[ X \approx WH \]

Very sparse
Observation

The size of the vocabulary increases only marginally with increasing number of documents.

(Yan et. al. 2013)
Model

The term-by-term matrix $K$ is relatively denser.
Model

The term-by-term matrix $K$ is relatively denser.

Hence, decompose $K$

\[ K \approx Q^T Q \]

- term-term
- term-top
- top-term
Event Progression

Assume a set of documents arrive at every time step

\[ K^t \approx Q^t T Q^t \]
Event Progression

Assume a set of documents arrive at every time step

$$K^t \approx Q_t^T Q^t$$

Relate the current data to the past history

$$K^t \approx Q_t^{TT}$$
Event Progression

Assume a set of documents arrive at every time step

\[ K^t \approx Q^t T Q^t \]

Relate the current data to the past history

\[ K^t \approx Q^t T^t Q^{t-1} \]
Event Progression

\[ K^t \approx Q^T_t T^t Q^{-1}_t \]

top-top  top-term
Quick Aside:

\[ K^t \approx Q_t^{TT} T_t Q_{t-1} \]

The “tracking matrix” - helps connect the present to the past. This will help build the “timelines” for events.
\[ ||K^t - Q^{tT}Q^t||_F^2 + ||K^t - Q^{tT}T^tQ^{t-1}||_F^2 \]
Model

\[ \|K_t^2 - Q_t^T Q_t\|_F^2 + \|K_t - Q_t^T T_t Q_t^{-1}\|_F^2 \]

KNOWN
Optimization

- Collective Matrix Factorization (Singh and Gordon 2008)

- Details in the paper

- Implementation on github

https://github.com/kjanani/matrix_factorization/blob/master/matrix_factorization.py
Interesting Question

*Ebola Outbreak 2014. What really happened?*
Interesting Question: Goals

*Ebola Outbreak 2014. What really happened?*

- Want list of all events that occurred.

- How did they progress?
Tasks

- Topic Detection

- Event timelines
Topic Detection

- Estimated topics: Estimate $Q_t$ at every timestep. Each row is a distribution over words.

- Groundtruth topics: hashtags
Topic Detection Baselines

- Two baselines from the event detection literature

- A few classic topic modeling baselines (NMF, LDA e.t.c.)
### Topic Detection Results

<table>
<thead>
<tr>
<th>model type</th>
<th>model</th>
<th>NDCG</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Ours] MEP</td>
<td>0.2027</td>
<td>0.0953</td>
<td></td>
</tr>
<tr>
<td>event detection</td>
<td>trend-detect o-cluster</td>
<td>0.1823</td>
<td>0.0862</td>
</tr>
<tr>
<td>topic modeling</td>
<td>O-BTM</td>
<td>0.1745</td>
<td>0.091</td>
</tr>
<tr>
<td></td>
<td>nmf</td>
<td>0.1722</td>
<td>0.0864</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>0.1245</td>
<td>0.0589</td>
</tr>
</tbody>
</table>

**For k = 7**

<table>
<thead>
<tr>
<th>model type</th>
<th>model</th>
<th>NDCG</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Ours] MEP</td>
<td>0.1626</td>
<td>0.0706</td>
<td></td>
</tr>
<tr>
<td>event detection</td>
<td>trend-detect o-cluster</td>
<td>0.1502</td>
<td>0.0539</td>
</tr>
<tr>
<td>topic modeling</td>
<td>O-BTM</td>
<td>0.1459</td>
<td>0.0569</td>
</tr>
<tr>
<td></td>
<td>nmf</td>
<td>0.1306</td>
<td>0.0565</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>0.0837</td>
<td>0.0366</td>
</tr>
</tbody>
</table>

**For k = 10**

<table>
<thead>
<tr>
<th>model type</th>
<th>model</th>
<th>NDCG</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Ours] MEP</td>
<td>0.1430</td>
<td>0.0696</td>
<td></td>
</tr>
<tr>
<td>event detection</td>
<td>trend-detect o-cluster</td>
<td>0.1379</td>
<td>0.0667</td>
</tr>
<tr>
<td>topic modeling</td>
<td>O-BTM</td>
<td>0.1271</td>
<td>0.0412</td>
</tr>
<tr>
<td></td>
<td>nmf</td>
<td>0.1057</td>
<td>0.0463</td>
</tr>
<tr>
<td></td>
<td>lda</td>
<td>0.0660</td>
<td>0.0164</td>
</tr>
</tbody>
</table>
Event Timelines

How to come up with *timelines* of events?
Event Timelines

How to come up with *timelines* of events?

Look at the tracking matrix $T^t$
Tracking Matrix

\[ \left\| K^t - Q^{tT} Q^t \right\|_F^2 + \left\| K^t - Q^{tT} T^t Q^{t-1} \right\|_F^2 \]

- Tracking matrix is a square matrix of all positive entries
Tracking Matrix

\[ \| K^t - Q^{tT}Q^t \|_F^2 + \| K^t - Q^{tT}T^tQ^{t-1} \|_F^2 \]

- Tracking matrix is a square matrix of all positive entries
- It shows how the topics have changed from \( t - 1 \) to \( t \)
- Row-i tells how topic-i at time \( t \) is related to all the topics at time \( t - 1 \)
New events
Continuing events
Entropy $H(X)$

$$H(X) := -\sum_{i=1}^{n} P(x_i) \log(P(x_i))$$

- Quantifies the amount of “randomness”
- Range $[0, 2.32]$
Based on Entropy

- \(< 1 \rightarrow\) continuing events
- \(1 \leq H(X) \leq 2 \rightarrow\) evolving events
- \(> 2 \rightarrow\) new events
- Also ending events
Timeline Generation

- Look at heatmap
- Connect dots to the previous timesteps (if possible)
- Else, new events, or noise
Timeline

- long term events
- medium term events
- short term events

- Simpsons 1997
- UK exercise test
- Michael Milan
- Hospital worker Germany
- Smith Geno Swine
- Nurse Emory released
- Disney Ebola
- Cal giving speech
- Man airplane joke
- Obama travel ban
- Nigeria free
- Paul Allen
- Kid bad ass lie
- Dallas patient
- Health worker quarantine
Example continuing events - memes

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014/10/19</td>
<td>kim, kardashian, married, american, died</td>
</tr>
<tr>
<td>2014/10/20</td>
<td>kim, kardashian, married, american, died</td>
</tr>
<tr>
<td>2014/10/21</td>
<td>kim, kardashian, married, american, died</td>
</tr>
<tr>
<td>2014/10/22</td>
<td>kim, kardashian, married, american, died</td>
</tr>
<tr>
<td>2014/10/23</td>
<td>kim, kardashian, married, american, died</td>
</tr>
</tbody>
</table>
Example of evolving events

<table>
<thead>
<tr>
<th>Date</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014/10/07</td>
<td>kidney, dialysis</td>
</tr>
<tr>
<td>2014/10/08</td>
<td>thomas, eric, duncan, died, first, patient died, patient</td>
</tr>
<tr>
<td>2014/10/09</td>
<td>duncan, fever, nurse</td>
</tr>
<tr>
<td>2014/10/10</td>
<td>nurse, symptoms</td>
</tr>
<tr>
<td>2014/10/11</td>
<td>health, care, worker, positive</td>
</tr>
<tr>
<td>2014/10/12</td>
<td>health, care, worker, protocol</td>
</tr>
<tr>
<td>2014/10/13</td>
<td>nurse, dallas, nina, pham</td>
</tr>
<tr>
<td>2014/10/14</td>
<td>health, care, worker, 2nd, positive</td>
</tr>
<tr>
<td>2014/10/15</td>
<td>nurse, flight, ohio</td>
</tr>
<tr>
<td>2014/10/16</td>
<td>virus, flight, nina, pham</td>
</tr>
<tr>
<td>2014/10/17</td>
<td></td>
</tr>
</tbody>
</table>
Thank you!

Questions?
Social Media Era