Problem

- Large and expanding data source (streaming)
- Difficult to summarize, contextualize topical content

Existing Solutions:
- Google News, newmap.jp, Zite

Limitations:
- Isolated view of topic (snapshot view)
- No context around presented topics
Solution

- Introduce “Storytelling” graph
- Provide linkages between topics over time
- Emphasize time and connectedness
- Contextualize documents into stories
  - Importance
  - Lifetime
  - Momentum
  - Evolution
Project Goals

- Investigate potential of “Storytelling” graph
- Implement web application for data exploration
  - RSS Reader
- Implement backend processing system
  - System run for 60 days
  - Data sources: general news, business, sports
Requirements for graph creation:

- Extract topics from corpus of documents by day
- Topics are easily comparable from different days

- Many unsupervised options exist
- Not all are sufficiently powerful
- Not all fit streaming use case
Topic Modeling Alternatives

- **K-Means on tf-idf weighted document vectors**
  - Centroid like a topic
  - Pros: centroids easy to compare across days
  - Cons: one topic per document, polysemy

- **LSA on document-term matrix per day**
  - SVD factoring into latent vector representations
  - Pros: overcomes synonymy, polysemy
  - Cons: vectors unique to input, not comparable
PLSA

Pros:
- Outputs probability distributions, easy to compare
- Thematically richer mixture model

Cons:
- Topic distribution prone to over-fitting
- Predecessor to LDA, used in this project
Topic Modeling (LDA)

Generative technique to estimate “topic” probability distributions in documents

Thematically rich representation of documents

Input: N documents, K topics:

Output: 2 Distributions
  - Document – Topic $\theta_n$
  - Topic – Word $\Phi_k$

Using Mallet library
Graph Creation

\[ z^{t+1} \]
\[ z^{t+3} \]
\[ z^{t+2} \]
\[ z^{t+4} \]
\[ d^t \]
\[ d^{t+1} \]
\[ d^{t+2} \]
\[ d^{t+2} \]
\[ d^{t+3} \]
\[ d^{t+4} \]
Graph Creation

- **Standard form:** $G = V, E$
  - $V = \text{Documents} \cup \text{Topics}$
  - $E = \text{Topic-topic, Document-topic links}$

- **Structure inherited from conditional dependence**
  - Topic-topic grafts between days
  - Edges undirected, weighted
  - Nodes labeled by time
**Graph Creation**

**Edges weight rules:**
- \( e(d_i, d_j) = 0 \) (not allowed)
- \( e(d_i, z_k) = \theta_{i,k} \) (document-topic weight)
- \( e(z_i, z_j) = \frac{\Phi_i \cdot \Phi_j}{||\Phi_i|| \cdot ||\Phi_j||} \) (cosine similarity)

All edge weights between \([0, 1]\)

Topic-topic weights blatantly not probabilistic

Topics joined within 15 day window \( W \), could extend
**Query by Topic**

- **Input:** start topic $z_k^t$
- **Output:** sub-graph of topic neighbors $Z$

**Idea:**
- Breadth first search
- Topics in path with product of edge weights $> \text{thresh}_z$
- $O(|E|)$ where $|E| = TK^2W$
- $T = \text{max } \# \text{ of days}, \ K = \# \text{ of topics } / \text{ day}, W \text{ is window}$
- Typically much less edges in database
Define: Subgraph()
Input: Starting topic $z_k^t$,
      Topic threshold $thresh_z$
Output: Related topic set $Z$
Begin:
    Initialize $Z = \emptyset$,
    $V = \{z_k^t\}$,
    $F = \{< z_k^t, 1 >\}$
    While $F \neq \emptyset$:
        Remove $z$ from $F$
        For $z' \in \text{TopicNeighbors}(z)$:
            Let $w = e(z, z') \cdot \text{weight}(z)$
            If $z' \notin V$ and $w > thresh_z$:
                $Z = Z \cup < z', w >$
                $F = F \cup < z', w >$
                $V = V \cup z'$
        return $Z$
End
Algorithm topic sub-graph of topic $z_k^t$
Query by Document

- **Input:** start document $d_i$
- **Output:** neighboring documents $D$

**Idea:**
- Start from all prominent topics for $d_i$
- Query sub-graph of all neighboring topics
- Return all documents with topic edge $> \text{thresh}_d$
- $O(TK^2W|C|)$ where $|C|$ is size of corpus
- Again, typically much less edges in database
Define: Similar()
Input: Starting document $d^t_{i}$,
      Topic-topic threshold $thresh_z$,
      Document-topic threshold $thresh_d$
Output: Related document set $D$

Begin:
  Initialize $D = \{\}$,
  $F = \{\}$
  For $z \in \text{TopicNeighbors}(d^t_{i})$:
    If $e(d^t_{i}, z) > thresh_d$:
      $F = F \cup <z, 1>$
  For $z \in \text{Subgraph}(F, thresh_z)$:
    For $d \in \text{DocumentNeighbors}(z)$:
      Let $w = e(d, z)\times\text{weight}(z)$
      If $w > thresh_d$:
        $D = D \cup <d, z, w>$
  Return $D$
End

Algorithm for relevant documents $d^t_{i}$
Graph Metrics

Given topic sub-graph $S$, time frame $\Delta$

\[
\text{Lifetime}(S) = \max_{z \in S} \text{time}(z) - \min_{z \in S} \text{time}(z)
\]

\[
\text{Velocity}(S, \Delta) = \frac{|S^\Delta|}{|\Delta|}
\]

\[
\text{Mass}(S, \Delta) = |D_{S^\Delta}|
\]

\[
\text{Mass2}(S, \Delta) = \sum_{z \in S^\Delta} \alpha_z
\]

\[
\text{Momentum}(S, \Delta) = \text{Mass}(S, \Delta) \times \text{Velocity}(S, \Delta)
\]
Backend Process

Sync → Fetch → Topic → Edge → Import
Backend Process

- Nightly process scheduled by cron job

- Sync
  - Synchronize RSS feeds
  - Google Reader API

- Fetch
  - Download and normalize content
  - Apache Commons, boilerpipe

- Topic
  - Perform LDA topic modeling on day’s worth of content
  - Mallet
Backend Process

- Edge
  - Create edges between topics

- Import
  - Save content, topic distributions, and edge data
  - MySQL, Perl
Web Application

- Shows top 10 terms for 7 days
- Topics sorted by connectedness, prominence
- Tabular view with “drill-down” into story documents
- Graph view for fun to explore edges
- Stack: Bootstrap, Jquery, SVG, Memcached, Dancer
## Web Application

<table>
<thead>
<tr>
<th>Date</th>
<th>News Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012-08-15</td>
<td>Obama Romney campaign president republican Mitt presidential running america candidate</td>
</tr>
<tr>
<td>2012-08-16</td>
<td>Obama president Romney campaign republican Mitt republicans presidential voters political</td>
</tr>
<tr>
<td>2012-08-17</td>
<td>Ryan Obama Romney political president republican whine campaign paul biden</td>
</tr>
<tr>
<td>2012-08-18</td>
<td>Romney Ryan Obama president campaign paul week years Mitt running Obama tax president taxes Romney plan cuts budget congress bush</td>
</tr>
<tr>
<td>2012-08-19</td>
<td>Ryan Obama Medicare Romney president plan campaign Mitt program republican Gaza Benghazi Hamas missiles Obama big trust rohrabacher Dome dogs Gaza rohbaracher Dome dogs</td>
</tr>
<tr>
<td>2012-08-20</td>
<td>Romney Obama Romney campaign president republican Mitt gap presidential paul Gaza Benghazi Hamas missiles Obama big trust rohbaracher Dome dogs Gaza rohbaracher Dome dogs</td>
</tr>
<tr>
<td>2012-08-21</td>
<td>Romney Ryan Obama Romney campaign Mitt medicare republican candidate paul president Obama Romney tax campaign business American Bain lies taxes Gaza Benghazi Hamas missiles Obama big trust rohbaracher Dome dogs Gaza rohbaracher Dome dogs</td>
</tr>
</tbody>
</table>
## Web Application

### news / 1851

<table>
<thead>
<tr>
<th>2012-08-16</th>
<th>2012-08-17</th>
<th>2012-08-18</th>
<th>2012-08-19</th>
<th>2012-08-20</th>
<th>2012-08-21</th>
<th>2012-08-22</th>
</tr>
</thead>
<tbody>
<tr>
<td>obama president</td>
<td>ryan campaign</td>
<td>romney republican mitt</td>
<td>medicare campaign paul</td>
<td>romney obama political</td>
<td>ryan campaign paul</td>
<td>obama president mitt</td>
</tr>
<tr>
<td>president</td>
<td>president</td>
<td>republican white campaign</td>
<td>president plan</td>
<td>president</td>
<td>republican</td>
<td>president gop</td>
</tr>
<tr>
<td>biden</td>
<td>political</td>
<td>week years mitt running</td>
<td>program republican</td>
<td>Romney</td>
<td>Paul</td>
<td>convention</td>
</tr>
<tr>
<td>Md. Democrat to play Ryan in VP debate prep</td>
<td>Romney raises $10.2M with Ryan on team</td>
<td>Medicare debate takes a turn</td>
<td>Romney: Ryan is bad for comedy</td>
<td>Obama-Romney Medicare debate</td>
<td>Romney: Ryan targets Ryan on education</td>
<td>Romney: Ryan campaign takes strong lead</td>
</tr>
<tr>
<td>Biden handlers monitor VP after uproar at comments</td>
<td>Romney campaign.tailors message by swing states</td>
<td>Romney boasts of 'wiser' campaign spending</td>
<td>Obama now targets Ryan on education</td>
<td>Paul Ryan’s Debut: What Went Right, What Went Wrong?</td>
<td>Romney goes after Romney on education</td>
<td>Romney’s Pick Of Ryan Hasn’t Changed Race, Polls Signal</td>
</tr>
<tr>
<td>Opinion: Good for GOP, bad for comedy</td>
<td>Romney, Ryan Turn Medicare Attacks Back on Obama</td>
<td>Obama goes after Romney on education</td>
<td>Both Sides Can Claim Some Money Advantage In Presidential Race</td>
<td>AP poll: Obama-Romney race remains tight</td>
<td>Romney’s Pick Of Ryan Hasn’t Changed Race, Polls Signal</td>
<td>Romney’s Pick Of Ryan Hasn’t Changed Race, Polls Signal</td>
</tr>
</tbody>
</table>

- Obama stands behind Biden
- Gallup: Obama gets low marks on economy
- White House rejects McCain comments on Biden
- Ryan Rips Biden’s Latest Remark as ‘Desperate’
Evaluation
Question: Is LDA effective at summarizing a story?

Answer is qualitative
- Topics created from unsupervised process
- No ground truth, already optimized likelihood

Issues:
- Artifact topics related to data source(s)
- Ambiguous topics from bad content
- Identify bad topic manually, suppress similar
## Topic Evaluation

<table>
<thead>
<tr>
<th>Prominent “Artifact” Topics</th>
<th>Ambiguous Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. comments</td>
<td>1. video</td>
</tr>
<tr>
<td>2. usa</td>
<td>2. watch</td>
</tr>
<tr>
<td>3. today</td>
<td>3. 2012</td>
</tr>
<tr>
<td>4. news</td>
<td>4. 1</td>
</tr>
<tr>
<td>5. life</td>
<td>5. july</td>
</tr>
<tr>
<td>6. comment</td>
<td>6. full</td>
</tr>
<tr>
<td>7. blog</td>
<td>7. images</td>
</tr>
<tr>
<td>8. travel</td>
<td>8. purposes</td>
</tr>
<tr>
<td>9. sports</td>
<td>9. responsible</td>
</tr>
<tr>
<td>romney</td>
<td>10. provided</td>
</tr>
<tr>
<td>10. stories</td>
<td>10. content</td>
</tr>
<tr>
<td></td>
<td>11. external</td>
</tr>
<tr>
<td></td>
<td>12. internet</td>
</tr>
<tr>
<td></td>
<td>13. sites</td>
</tr>
<tr>
<td></td>
<td>14. 2012</td>
</tr>
<tr>
<td></td>
<td>15. network</td>
</tr>
<tr>
<td></td>
<td>16. provided</td>
</tr>
<tr>
<td></td>
<td>17. content</td>
</tr>
<tr>
<td></td>
<td>18. purposes</td>
</tr>
<tr>
<td></td>
<td>19. responsible</td>
</tr>
<tr>
<td></td>
<td>20. provided</td>
</tr>
<tr>
<td></td>
<td>21. content</td>
</tr>
</tbody>
</table>
Question: Is query by topic effective?

Requires threshold. How to set?

Answer:
- Again, qualitative, no true value
- Look at distribution of similarity values
- Most values near 0, examine > .3
- Spot check increasingly dissimilar distributions
- Sample storytelling graph
### Query By Topic

**Impact of Topic Similarity**
Top ten terms of similar distributions of decreasing similarity

<table>
<thead>
<tr>
<th>start</th>
<th>.9</th>
<th>.8</th>
<th>.6</th>
<th>.4</th>
<th>.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>isaac</td>
<td>isaac</td>
<td>isaac</td>
<td>storm</td>
<td>haiti</td>
<td>fire</td>
</tr>
<tr>
<td>louisiana</td>
<td>louisiana</td>
<td>lousiana</td>
<td>isaac</td>
<td>isaac</td>
<td>storm</td>
</tr>
<tr>
<td>hurricane</td>
<td>hurricane</td>
<td>hurricane</td>
<td>storm</td>
<td>isaac</td>
<td>crisis</td>
</tr>
<tr>
<td>orleans</td>
<td>orleans</td>
<td>orleans</td>
<td>national</td>
<td>haiti</td>
<td>fire</td>
</tr>
<tr>
<td>storm</td>
<td>storm</td>
<td>baithwaite</td>
<td>gulf</td>
<td>island</td>
<td>storm</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
<td>coast</td>
<td>portau prince</td>
<td>earthquake</td>
</tr>
<tr>
<td>braithwaite</td>
<td>residents</td>
<td>street</td>
<td>people</td>
<td>center</td>
<td>center</td>
</tr>
<tr>
<td>residents</td>
<td>parish</td>
<td>residents</td>
<td>topical</td>
<td>rain</td>
<td>rain</td>
</tr>
<tr>
<td>plaquemines</td>
<td>flooded</td>
<td>pontchartrain</td>
<td>rain</td>
<td>house</td>
<td>house</td>
</tr>
<tr>
<td>flooded</td>
<td>mississippi</td>
<td>floodwaters</td>
<td>winds</td>
<td>damaged</td>
<td>damaged</td>
</tr>
</tbody>
</table>
### Example of evolving story from topic perspective

<table>
<thead>
<tr>
<th>09-12</th>
<th>09-13</th>
<th>09-14</th>
<th>09-15</th>
<th>09-16</th>
</tr>
</thead>
<tbody>
<tr>
<td>teachers</td>
<td>school</td>
<td>teachers</td>
<td>teachers</td>
<td>teachers</td>
</tr>
<tr>
<td>chicago</td>
<td>teachers</td>
<td>school</td>
<td>union</td>
<td>chicago</td>
</tr>
<tr>
<td>strike</td>
<td>strike</td>
<td>chicago</td>
<td>union</td>
<td>school</td>
</tr>
<tr>
<td>unions</td>
<td>children</td>
<td>chicago</td>
<td>chicago</td>
<td>school</td>
</tr>
<tr>
<td>schools</td>
<td>children</td>
<td>children</td>
<td>school</td>
<td>strike</td>
</tr>
<tr>
<td>public</td>
<td>students</td>
<td>board</td>
<td>deal</td>
<td>schools</td>
</tr>
<tr>
<td>school</td>
<td>district</td>
<td>kids</td>
<td>students</td>
<td>contract</td>
</tr>
<tr>
<td>union</td>
<td>students</td>
<td>day</td>
<td>rally</td>
<td>emanuel</td>
</tr>
<tr>
<td>students</td>
<td>schools</td>
<td>students</td>
<td>teacher</td>
<td>delegates</td>
</tr>
<tr>
<td>teacher</td>
<td>kids</td>
<td>students</td>
<td>lewis</td>
<td></td>
</tr>
</tbody>
</table>
Query By Document

Question: Is query by document effective?

Sample datasets over 2 week period

Evaluate neighbors as cluster by WCSS

Using Storytelling graph, $0.6 = \text{thresh}_z, 0.15 = \text{thresh}_d$

Using cosine-similarity $> 0.45$ between documents

$$WCSS = \sum_{i \in D_S} \sum_{j} ||w_{i,j} - \mu_j||^2$$
Query By Document

Method of minimum WCSS neighbors by dataset

- **Dataset**
  - News
  - Sports
  - Business

- **Instances**
  - 0
  - 50
  - 100
  - 150
  - 200
  - 250
  - 300
  - 350

- **Legend**
  - Storytelling Graph
  - Document Similarity
Example querying the storytelling graph by document starting from:
“U.S. ambassador killed in Libya”

<table>
<thead>
<tr>
<th>Date</th>
<th>Summary</th>
</tr>
</thead>
</table>
| 09-12 | How the Benghazi attack unfolded  
Envoy to Libya dies in rocket blast  
U.S. vows to hunt down ambassador’s killers  
U.S. ambassador killed in Libya |
| 09-13 | Inside the attack in Benghazi  
How the deadly attack in Benghazi unfolded  
U.S. warships move toward Libya  
Benghazi casualties |
| 09-14 | Opinion: Some 'don’t get dissent’  
President Obama, Secretary Clinton Receive Bodies Of Americans Killed In Libya  
4 held in Libya attack |
Graph Metrics

- **Question:** Are graph metrics effective?
- **Sample date range and sort by momentum**
- **Plot momentum over time**
# Graph Metrics

<table>
<thead>
<tr>
<th>Mass</th>
<th>Mass$_\alpha$</th>
<th>Velocity</th>
<th>Momentum</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>244</td>
<td>0.634</td>
<td>9.0</td>
<td>2196.0</td>
<td>romney obama president republican mitt campaign</td>
</tr>
<tr>
<td>128</td>
<td>0.163</td>
<td>8.857</td>
<td>1133.714</td>
<td>gaza benghazi hamas dogs missiles obama dome walk</td>
</tr>
<tr>
<td>93</td>
<td>0.170</td>
<td>2.286</td>
<td>212.571</td>
<td>isaac louisiana hurricane storm orleans home residents</td>
</tr>
<tr>
<td>9</td>
<td>0.029</td>
<td>0.571</td>
<td>5.143</td>
<td>state penn football coach campus sandusky paterno</td>
</tr>
<tr>
<td>11</td>
<td>0.200</td>
<td>0.286</td>
<td>3.143</td>
<td>syrian government syria damascus rebels violence</td>
</tr>
<tr>
<td>6</td>
<td>0.015</td>
<td>0.429</td>
<td>2.571</td>
<td>police officers car henderson kidnapped arrested duran</td>
</tr>
<tr>
<td>8</td>
<td>0.020</td>
<td>0.286</td>
<td>2.286</td>
<td>bus party fernandez 16 head hospital teens overpass</td>
</tr>
</tbody>
</table>
Graph Metrics

Momentum of Obama, Gaza, Hurricane

- Obama
- Gaza
- Hurricane
Conclusion

Did the project work?
- Yes, surprisingly well.
- Sub-graphs intuitive, connectedness very useful

Project depended on good article extraction
- Boilerpipe library proved to be critical

Future work
- Incorporate NLP entity extraction
- Add search features
- More visualization work
Demo