Abstract

- Detecting topics in Twitter streams has been gaining an increasing amount of attention, as it can:
  - Be of great support for communities struck by natural disasters.
  - Assist companies and political parties understand users’ opinions and needs.
- Traditional approaches represent topics using terms that are not properly connected.
- In this work, we propose an Exemplar-based topic detection approach which:
  - Represents topics using a few selected tweets that allows for an easy interpretation of the topics.
- Experimental evaluation shows that the Exemplar-based approach achieves the best term precision, while maintaining good topic recall and running time compared to other approaches.

Exemplar-based Topic Detection

- Exemplar-based approach represents each topic by a tweet (exemplar).
- Other approaches such as LSA and K-means detected topics may contain:
  - Unstructured sentences.
  - Unrelated keywords.

Exemplar Selection Criterion

<table>
<thead>
<tr>
<th>Exemplar</th>
<th>Selection Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet 1</td>
<td>Its similarity distribution has low variance.</td>
</tr>
<tr>
<td>Tweet 2</td>
<td>Its similarity distribution has high variance.</td>
</tr>
<tr>
<td>Tweet 3</td>
<td>Good topic representative.</td>
</tr>
<tr>
<td>Tweet 4</td>
<td>Not similar to most of the others.</td>
</tr>
</tbody>
</table>

Exemplar Selection Algorithm

- Choose tweets that have the highest variances.
- These tweets are not guaranteed to discuss different topics.
- Choose the highest variances k tweets that are not similar to each other.
- Construct the set of exemplars E using the objective function:
  \[
  \text{var}(S_i) = \frac{1}{n} \sum_{j=1}^{i} (S_{ij} - \mu_i) \]
  where:
  - n is the total number of tweets.
  - \( S_{ij} \) is the similarity between \( t_i \) and \( t_j \).
  - \( \mu_i \) is the average similarity for \( t_i \).
- The proposed approach is compared against:
  - Latent Semantic Analysis (LSA).
  - Stochastic LSA (LSA Stoch.).
  - Non-negative matrix factorization: Rank-1 Downdate (R1D).
  - K-means.

Experimental Setup

- Two Twitter datasets are used for evaluation, which are collected by [1]:
  - US Elections contains 578,837 tweets.
  - Super Tuesday contains 353,650 tweets.
- The used measures are: Topic recall, term precision and term recall.

References