Approaches

- **SLSTC-J-R1**
  - Create word vectors with Word2vec (using Japanese Wikipedia and Nicopedia as corpora)
  - Weight each word vector using tf-idf
  - Represent each tweet as a sum of word vectors
  - Select top ten tweets from the repository that are similar to the generated reply.

- **SLSTC-J-R2**
  Based on word co-occurrence networks
  - **Network Generation**
    - For each word bigram in each tweet, an edge goes out from a word in the left word to the right word
    - For each post-reply pair, an edge goes out from a word in the post into a word in the reply
  - **Subnetwork Extraction**
    - Given a test post tweet, obtain a subnetwork from the above as follows:
      - $V': \subseteq V$: words in the test tweet, plus words connected by an edge $E'$ from a word in the new tweet
      - $E': \subseteq E$: edges involving words from $V'$
  - **PageRank Calculation**
    - Calculate the PageRank $PR(w)$ of each node $w \in V'$. $PR(w)$ is initially set to $1/d|V'|$ and updated through 100 iterations as follows:
      \[
      PR(w) = \frac{1 - d}{|V'|} + \sum_{w' \in V'(w)} \frac{d \times PR(w')}{|E'(w')|}
      \]
      $E'(w)$: set of edges from $w$
      $d$: parameter (set to 0.9 based on a preliminary experiment)

- **SLSTC-J-R3**
  Same as R2, except that we removed all continuous occurrences of "w" in the test post tweets
  (since "ww" "www" in Japanese tweets are similar in meaning to the English "lol" and can be considered as noise)

Results and failure analysis

Results

<table>
<thead>
<tr>
<th>Test post tweet</th>
<th>Nouns extracted</th>
<th>Nouns in subgraph with highest score</th>
<th>PageRank</th>
<th>IDF</th>
<th>PageRank*IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLSTC-J-R1</td>
<td>0.0381</td>
<td>0.0364</td>
<td>0.1644</td>
<td>0.1650</td>
<td></td>
</tr>
<tr>
<td>SLSTC-J-R2</td>
<td>0.0782</td>
<td>0.0332</td>
<td>0.3146</td>
<td>0.1795</td>
<td></td>
</tr>
<tr>
<td>SLSTC-J-R3</td>
<td>0.0054</td>
<td>0.0032</td>
<td>0.0391</td>
<td>0.0196</td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>0.4574</td>
<td>0.3583</td>
<td>0.7817</td>
<td>0.7050</td>
<td></td>
</tr>
<tr>
<td>MIN</td>
<td>0.0054</td>
<td>0.0032</td>
<td>0.0391</td>
<td>0.0196</td>
<td></td>
</tr>
</tbody>
</table>

Failure analysis

- Our systems returned the same nonrelevant tweets for many of test post tweets
- Each subnetwork includes too many words
- Morphological analysis often did not perform well due to spelling variations and symbols in the tweets

Future Work

- Improve the scoring scheme and setting a threshold for constructing subnetworks
- Improve the morphological analysis dictionary
- Remove symbols (e.g. "", "")
- Normalize expressions that are characteristic in microblogs