Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings

Luchen Tan, Haotian Zhang, Charles L.A. Clarke and Mark D. Smucker
Information Retrieval Group, University of Waterloo

Overview

Methods
Mikolov et al. (2013a) finds that the relative positions of semantically related words are preserved across languages. This paper adapts this result to explore differences between corpora written in a single language.

Wikipedia word representation vector matrix: 
\[ A = [a_1^T, a_2^T, ..., a_n^T]^T \]

Twitter word representation vector matrix: 
\[ B = [b_1^T, b_2^T, ..., b_n^T]^T \]

Linear transformation matrix \( M \) that maps vector space \( A \) to \( B \) can be learned.

\[ AM = B \]

Transformation objective:
\[ \min_M \sum_{i=1}^{n} \| a_i M - b_i \| ^2 \]

Sort transformed vectors according to cosine distance:
\[ \text{Sim}(a_i^t, b_i), n \leq i \leq c. \]

Data

- We choose a word vector dimensionality of 300, which is typical for training large datasets with Word2Vec.
- We choose 10 words for the context window size.

Examples - Words with the greatest adjusted distance

<table>
<thead>
<tr>
<th>Word</th>
<th>Twitter Most Similar</th>
<th>Wikipedia Most Similar</th>
</tr>
</thead>
<tbody>
<tr>
<td>bc</td>
<td>because bcus bcuz cuz cos</td>
<td>bce macedon hellenistic euthydemus ptolemaic</td>
</tr>
<tr>
<td>ill</td>
<td>ill imma ima will you</td>
<td>unwell sick frail fated bedridden</td>
</tr>
<tr>
<td>cameron</td>
<td>cam nash followmecam camerons callmecam</td>
<td>gillies duncan mckay mitchell bryce</td>
</tr>
<tr>
<td>mentions</td>
<td>unfollow reply respond strangerswelcomended offened</td>
<td>mentions mentioned mentioning reference attested</td>
</tr>
<tr>
<td>miss</td>
<td>miss love missss missssss ismiss</td>
<td>pageant pageants titleholder titleholders pageantopolis</td>
</tr>
<tr>
<td>yup</td>
<td>yep yupp yeah yea yep</td>
<td>chevak yupik gwaii tinglit nunivak</td>
</tr>
<tr>
<td>taurus</td>
<td>capricorn sagittarius pisces gemini scorpio</td>
<td>poniatovii scorpio subcompact sagittarius chevette</td>
</tr>
</tbody>
</table>

Validation

The vectors of some common English words are visualized by applying principal component analysis (PCA).

Adjusted Distance

Words with higher frequencies have higher average cosine similarity than those words with lower frequencies. Adjusted distance corrects for this finding.

\[ D_{\text{adjusted}}(t) = \text{Sim}(\epsilon_{\text{medium}}) - \text{Sim}(t) \]

Result

Rank biased overlap measures similarity between two similar word ranked lists for a specific word (Webber et al., 2010).

\[ RBO = (1 - \varphi) \sum_{k=1}^{\infty} \frac{1}{k} |A_{1,k} \cap B_{1,k}| \]

Corpora | #Documents | #Vocabulary |
---------|------------|-------------|
Wikipedia | 3,776,418  | 7,267,802   |
Twitter   | 263,572,856| 13,622,411  |