Modern Methods for Exploring Text Data

Peter Baumgartner
Data Scientist @ RTI International
GASP 2019
How is it tokenizing?
What tokens is it excluding?
Can I differentiate nouns, verbs, and adjectives?
Can I combine tokens with the same root word in a meaningful way?
Which tokens are unique for this corpus?
Can I see a token used in context?
How will this scale to larger amounts of text?
Can I cluster words by their use?
SAMPLE DATASET

Post titles and text from the r/statistics and r/askstatistics communities on reddit from December 2015 – March 2019.

30,693 total posts.

61% from r/statistics
39% from r/askstatistics
"Can anyone tell me what a p-value is?"
**Token Attributes**

**Lemma**: be

**POS**: VERB

**Prob**: 0.0088
## Data Processing & Lemma Statistics

<table>
<thead>
<tr>
<th>Lemma</th>
<th>POS</th>
<th>Stopword</th>
<th>Corpus Count</th>
<th>Corpus Prob</th>
<th>English Prob</th>
<th>Corpus/English Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>manual</td>
<td>NOUN</td>
<td>FALSE</td>
<td>29</td>
<td>0.000009</td>
<td>0.000023</td>
<td>0.388</td>
</tr>
<tr>
<td>injure</td>
<td>VERB</td>
<td>FALSE</td>
<td>13</td>
<td>0.000004</td>
<td>0.000001</td>
<td>7.033</td>
</tr>
<tr>
<td>methods</td>
<td>NOUN</td>
<td>FALSE</td>
<td>58</td>
<td>0.000018</td>
<td>0.000277</td>
<td>0.063</td>
</tr>
<tr>
<td>irregular</td>
<td>ADJ</td>
<td>FALSE</td>
<td>12</td>
<td>0.000004</td>
<td>0.000015</td>
<td>0.243</td>
</tr>
<tr>
<td>forests</td>
<td>NOUN</td>
<td>FALSE</td>
<td>10</td>
<td>0.000003</td>
<td>0.000016</td>
<td>0.193</td>
</tr>
</tbody>
</table>

\[ n\text{ lemmas} = 74,358 \]
Interactive Visualization 1: Exploring lemma counts and uniqueness by parts of speech

**WORD EMBEDDINGS**

For each word, calculate a *location* in vector space such that words that appear in similar contexts are *located* near each other.

more on word2vec: [https://jalammar.github.io/illustrated-word2vec/](https://jalammar.github.io/illustrated-word2vec/)
WORD EMBEDDINGS & DIMENSION REDUCTION

Word Embeddings

able
advance
advice
...
...
year
wrong
write

2D Projection

able
advance
advice
...
...
year
wrong
write

UMAP

Project an $n$-dimensional space down to 2 dimensions, such that both local and global structure is retained.

projection dimensions
# MERGE PROJECTION BACK TO DATASET

```
sample
```

<table>
<thead>
<tr>
<th>Lemma</th>
<th>POS</th>
<th>...</th>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>manual</td>
<td>NOUN</td>
<td>...</td>
<td>0.934</td>
<td>0.734</td>
</tr>
<tr>
<td>injure</td>
<td>VERB</td>
<td>...</td>
<td>0.723</td>
<td>0.222</td>
</tr>
<tr>
<td>methods</td>
<td>NOUN</td>
<td>...</td>
<td>0.147</td>
<td>0.063</td>
</tr>
<tr>
<td>irregular</td>
<td>ADJ</td>
<td>...</td>
<td>0.237</td>
<td>0.243</td>
</tr>
<tr>
<td>forests</td>
<td>NOUN</td>
<td>...</td>
<td>0.717</td>
<td>0.182</td>
</tr>
</tbody>
</table>
Interactive Visualization 2: Exploring projections of word embeddings from word2vec

https://modern-text-exploration.netlify.com/w2v-umap.html