Applied Natural Language Processing

Info 256
Lecture 3: Finding Distinctive Terms (Sept. 2, 2021)

David Bamman, UC Berkeley
Panel B: Phrases Used More Often by Republicans

**Two-Word Phrases**
- stem cell
- natural gas
- death tax
- illegal aliens
- class action
- war on terror
- embryonic stem
- tax relief
- illegal immigration
- date the time
- personal accounts
- Saddam Hussein
- pass the bill
- private property
- border security
- President announces
- human life
- Chief Justice
- human embryos
- increase taxes
- retirement accounts
- government spending
- national forest
- minority leader
- urge support
- cell lines
- cord blood
- action lawsuits
- economic growth
- food program

**Three-Word Phrases**
- embryonic stem cell
- hate crimes legislation
- adult stem cells
- oil for food program
- personal retirement accounts
- energy and natural resources
- global war on terror
- hate crimes law
- change hearts and minds
- global war on terrorism
- Circuit Court of Appeals
- death tax repeal
- housing and urban affairs
- million jobs created
- national flood insurance
- oil for food scandal
- private property rights
- temporary worker program
- class action reform
- Chief Justice Rehnquist
- Tongass national forest
- pluripotent stem cells
- Supreme Court of Texas
- Justice Priscilla Owen
- Justice Janice Rogers
- American Bar Association
- growth and job creation
- natural gas natural
- Grand Ole Opry
- reform social security

Schwartz et al. (2013), "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach"
Which are the words most likely to be from Android and most likely from iPhone?

http://varianceexplained.org/r/trump-tweets/
Distinctive terms

- Finding distinctive terms is useful:
  - As a data exploration exercise to understand larger trends in individual word differences.
  - As a pre-processing step of feature selection.
  - When the two datasets are A and ¬A, these terms also provide insight into what A is about.
  - Many methods for finding these terms! (Developed in NLP, corpus linguistics, political science, etc.)
Difference in proportions

For word $w$ written by author with label $k$ (e.g., \{democrat, republican\}), define the frequency to be the normalized count of that word

$$f_{w,k} = \frac{C(w, k)}{\sum_{w'} C(w', k)}$$

where $C(w, k)$ is the count of word $w$ in group $k$, and $\sum_{w'} C(w', k)$ is the count of all words in group $k$.

$$f_{w,k} = \text{dem} - f_{w,k} = \text{repub}$$
Monroe et al. (2009), "Fightin’ Words"
Difference in proportions

• The difference in proportions is a conceptually simple measure and easily interpretable.

• Drawback: tends to emphasize words with high frequency (where even comparatively small differences in word usage between groups is amplified).

• Also, no measure whether a difference is statistically meaningful. We have uncertainty about the what the true proportion is for any group.
• $\chi^2$ (chi-square) is a statistical test of dependence—here, dependence between the two variables of word identity and corpus identity.

• For assessing the difference in two datasets, this test assumes a 2x2 contingency table:

<table>
<thead>
<tr>
<th></th>
<th>word</th>
<th>~word</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus 1</td>
<td>7</td>
<td>104023</td>
</tr>
<tr>
<td>corpus 2</td>
<td>104</td>
<td>251093</td>
</tr>
</tbody>
</table>
Does the word *robot* occur significantly more frequently in science fiction?

\[
\chi^2
\]

<table>
<thead>
<tr>
<th></th>
<th>robot</th>
<th>¬robot</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sci-fi</td>
<td>104</td>
<td>1004</td>
<td>10.3%</td>
</tr>
<tr>
<td>¬sci-fi</td>
<td>2</td>
<td>13402</td>
<td>0.015%</td>
</tr>
</tbody>
</table>
For each cell in contingency table, sum the squared difference between observed value in cell and the expected value assuming independence.

\[
\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]
<table>
<thead>
<tr>
<th></th>
<th>robot</th>
<th>¬robot</th>
<th>sum</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>sci-fi</td>
<td>104</td>
<td>1004</td>
<td>1108</td>
<td>0.076</td>
</tr>
<tr>
<td>¬sci-fi</td>
<td>2</td>
<td>13402</td>
<td>13404</td>
<td>0.924</td>
</tr>
<tr>
<td>sum</td>
<td>106</td>
<td>14406</td>
<td></td>
<td></td>
</tr>
<tr>
<td>frequency</td>
<td>0.007</td>
<td>0.993</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Among 14512 words, we would expect to see 7.69 occurrences of \textit{robot} in sci-fi texts.

Assuming independence:

\[
P(\text{robot, scifi}) = P(\text{robot}) \times P(\text{scifi})
\]
\[
= 0.007 \times 0.076 = 0.00053
\]
What $\chi^2$ is asking is: how different are the observed counts from the counts we would expect given complete independence?

<table>
<thead>
<tr>
<th></th>
<th>robot</th>
<th>¬robot</th>
</tr>
</thead>
<tbody>
<tr>
<td>sci-fi</td>
<td>104</td>
<td>1004</td>
</tr>
<tr>
<td>¬sci-fi</td>
<td>2</td>
<td>13402</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>robot</th>
<th>¬robot</th>
</tr>
</thead>
<tbody>
<tr>
<td>sci-fi</td>
<td>7.69</td>
<td>1095.2</td>
</tr>
<tr>
<td>¬sci-fi</td>
<td>93.9</td>
<td>13315.2</td>
</tr>
</tbody>
</table>
\[ \chi^2 \]

- With algebraic manipulation, simpler form for 2x2 table \( O \) (cf. Manning and Schütze 1999)

\[ \chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \]
The $\chi^2$ value is a statistic of dependence with a probability governed by a $\chi^2$ distribution; if this value has low enough probability in that measure, we can reject the null hypothesis of the independence between the two variables.
5% probability mass from 3.84 forward; if $\chi^2$ is in this region, then we reject independence as being too unlikely (at $\alpha = 0.05$)
Chi-square is ubiquitous in corpus linguistics (and in NLP as a measure of collocations).

A few caveats for its use:

- Each cell should have an *expected* count of at least 5
- Each observation is independent
A drawback, however, is due to the burstiness of language: the tendency for the same words to clump together in texts.

Chi-square is testing for independence of two variables (word identity and corpus identity), but it assumes each mention of the word is independent from the others.
• Is Dracula really a word that distinguishes these two corpora?

• It distinguishes one text, but otherwise doesn’t appear in the corpus at all.
Mann-Whitney rank sums test

- Mann-Whitney is a test of the difference in some quantity of interest in two datasets. Null hypothesis: if you select a random sample from group A and another from group B, just as likely that A will be greater than B as less than B.
Mann-Whitney

<table>
<thead>
<tr>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ranks

<table>
<thead>
<tr>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 11 | 12 | 13 | 14 |   |
Mann-Whitney

\[ R_1 = 7 + 9 + 10 + 11 + 12 + 13 + 14 = 76 \]
Mann-Whitney

\[ R_1 = 7 + 9 + 10 + 11 + 12 + 13 + 14 = 76 \]

\[ U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} \]

- Once we have this U value, we can ask whether it’s significantly different from the average value we would expect if there’s no difference between the two groups at all. (We can do so by converting U to a z-score using a Normal approximation and checking significance).
In corpus linguistics, each measurement is the count of a word in a fixed-sized chunk of text (e.g., 500 words).

This lets us accommodate a more realistic assumption about the burstiness of language.
417 mentions of Dracula in one book

Not a significant difference in ranks
Log-odds ratios with priors

- The odds of a word is another informative measure:

\[
\frac{\text{probability of event occurring in corpus } X}{\text{probability of event not occurring in corpus } X} = \frac{\frac{\text{count of word } w \text{ in corpus } X}{\text{count of all words } \neg w \text{ in corpus } X}}{\frac{\# \text{ words in } X}{\# \text{ words in } X}}
\]

\[
\frac{\# \text{ mentions of } "\text{recall}"}{\# \text{ mentions of } \neg "\text{recall}"} = \frac{14}{1,000,000} = 0.000014
\]
Log-odds ratios with priors

- Two get a measure of difference, we can compare two odds in different corpora

\[
\frac{\text{# mentions of "recall"}}{\text{# mentions of } \neg \text{"recall"}} = \frac{14}{1,000,000} = 0.000014 \quad \frac{42}{2,000,000} = 0.000021
\]

- The odds ratio gives us one way of combining these into a single score

\[
\frac{14}{2,000,000} = 0.667
\]
Log-odds ratios with priors

• But this is bounded by $(0, \infty)$ and not easy to interpret with respect to the boundary (1) separating a word being likelier in corpus than another.

\[
\frac{14}{1,000,000} \div \frac{42}{2,000,000} = 0.667
\]

• We can work with the log instead, which transforms this into the space $(-\infty, \infty)$, with 0 as a boundary

\[
\log\left(\frac{14}{1,000,000} \div \frac{42}{2,000,000}\right) = -0.4054
\]
Log-odds ratios with priors

• What if we have 0 counts?

• We can add pseudocounts! e.g., assume vocabulary size of 10,000 words, 100 here = 10,000 * 0.01 to account for total pseudocount mass added, and we remove 0.01 from the denominators since the denominator is the count of ¬word.

\[
\log \left( \frac{14}{1,000,000} \right) = \text{\textcolor{red}{7.94}}
\]

\[
\log \left( \frac{14+0.01}{1,000,000+100-0.01} \right) = \text{\textcolor{red}{7.94}}
\]
Log-odds ratios with priors

\[
\log \left( \frac{14 + 0.01}{1,000,000 + 100 - 0.01} \frac{42 + 0.01}{2,000,000 + 100 - 0.01} \right) = -0.4050
\]

\[
= \log \left( \frac{14 + 0.01}{1,000,000 + 100 - 0.01} \right) - \log \left( \frac{42 + 0.01}{2,000,000 + 100 - 0.01} \right)
\]

How confident are we about these estimates?
Log-odds ratios with priors

- Transform them into z-scores by dividing them by the standard deviation.

\[
\approx \log \left( \frac{14 + 0.01}{1,000,000 + 100 - 0.01} \right) - \log \left( \frac{42 + 0.01}{2,000,000 + 100 - 0.01} \right)
\]

\[
\sqrt{\frac{1}{14 + 0.01} + \frac{1}{42 + 0.01}}
\]

The larger the term counts (e.g., 14, 42), the more confident we can be that the difference is meaningful.
Other methods

- There are many other methods for learning distinguishing words between two corpus; major classes:
  - Model-based methods that assume parametric forms + Bayesian priors (for smoothing) [Monroe et al. 2009]
  - Methods using classification to learn informative features that separate classes.
Activity

• Hypothesize terms that will be different between 2020 Democrat and Republican platforms.

• Execute chi-square to find terms that are different

• Compare to Mann-Whitney for this data.