Comparing Text Corpora

Instructor: Steve Wilson

11-Nov-2020

Initial Text Analysis

• Scenario: you are given access to a new dataset
  • 2 corpora, each contains thousands of plain text files
  • You want to understand and quantify:
    • What is the content of these documents? What are they about?
    • How does the content of these corpora differ?

• What are some things you might try first?
Lecture Objectives

• Analyze text corpora
  • Content analysis background
  • Word-level differences
  • Dictionaries and Lexicons
  • Topic modeling
  • Annotation + classification

Content Analysis

• Goal: given some documents determine
  • What are the types of content present? (themes/topics)
  • Which documents contain which topics?
• Traditionally a manual process
  1. Read a subset of documents, define themes/topics
  2. Determine consistent coding methodology
  3. Read all documents and label them according to codes
  4. Check agreement between human coders
  5. Settle disagreements via a third-party
  6. Analyze resulting annotations
Content Analysis

• Can this process be automated?
  • Yes, to an extent
• **Should** this process be automated?
  • Humans are better than machines at this task (for now?)
  • Computers are *much, much* faster
    • Avg. human reading speed: 250 wpm
    • Assume 1K words/document, 50K documents…
      • Average person needs > 4 months to read
      • This is a *relatively small* corpus for modern NLP
  • Modern computers can process millions of words/second

Automated Content Analysis

• Single corpus/class
  • Word frequency analysis
  • Dictionaries & Lexicons
  • Topic modelling
• Multiple corpora/classes
  • Word-level differences
  • Dominance Scores
  • Topic-level differences
Word Level Analysis

Word frequency analysis

• Very simple starting point
  1. Preprocess as usual (lowercasing? stemming?...)
  2. Count words
  3. Normalize by document length
  4. Average across all documents
**Word-level Differences**

- Which words best characterize a corpus?
  - Need a reference corpus
- Some methods to do this:
  - Mutual information
  - Chi squared
- Can also be used for *feature selection*

---

**Mutual Information**

- $I(X;Y)$
  - How much can I learn about $X$ by observing $Y$?
  - Is the same as *information gain*
  - Is *not* the same as *pointwise mutual information*
- We want to learn about important words in our corpus
- What should $X$ and $Y$ be?
  - $X = U =$ document contains term $t$ (Boolean)
  - $Y = C =$ class is the target class (Boolean)

$$I(U;C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$
Mutual Information

\[ I(U;C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)} \]

- Given count data for 2 classes, can be computed as:

\[ I(U;C) = \frac{N_{11}}{N} \log_2 \frac{N_{1,1}}{N_{1,N}} + \frac{N_{01}}{N} \log_2 \frac{N_{N_{0,1}}}{N_{N_{0,N}}} \]
\[ + \frac{N_{10}}{N} \log_2 \frac{N_{N_{1,0}}}{N_{N_{1,N}}} + \frac{N_{00}}{N} \log_2 \frac{N_{N_{0,0}}}{N_{N_{0,N}}} \]

Example:

- What is \( I(U;C) \) given these values?

<table>
<thead>
<tr>
<th>( e_t = e_{\text{export}} = 1 )</th>
<th>( e_t = e_{\text{poultry}} = 1 )</th>
<th>( e_t = e_{\text{poultry}} = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_{11} = 49 )</td>
<td>( N_{10} = 27,652 )</td>
<td>( N_{01} = 141 )</td>
</tr>
<tr>
<td>( N_{00} = 774,106 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Mutual Information for News Data**

<table>
<thead>
<tr>
<th>UK</th>
<th>China</th>
<th>poultry</th>
</tr>
</thead>
<tbody>
<tr>
<td>london</td>
<td>0.1925</td>
<td></td>
</tr>
<tr>
<td>uk</td>
<td>0.0755</td>
<td></td>
</tr>
<tr>
<td>british</td>
<td>0.0596</td>
<td></td>
</tr>
<tr>
<td>stg</td>
<td>0.0555</td>
<td></td>
</tr>
<tr>
<td>britain</td>
<td>0.0469</td>
<td></td>
</tr>
<tr>
<td>plc</td>
<td>0.0357</td>
<td></td>
</tr>
<tr>
<td>england</td>
<td>0.0238</td>
<td></td>
</tr>
<tr>
<td>pence</td>
<td>0.0212</td>
<td></td>
</tr>
<tr>
<td>pounds</td>
<td>0.0149</td>
<td></td>
</tr>
<tr>
<td>english</td>
<td>0.0126</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>coffee</th>
<th>elections</th>
<th>sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>coffee</td>
<td>0.0111</td>
<td></td>
</tr>
<tr>
<td>bags</td>
<td>0.0042</td>
<td></td>
</tr>
<tr>
<td>growers</td>
<td>0.0025</td>
<td></td>
</tr>
<tr>
<td>kg</td>
<td>0.0019</td>
<td></td>
</tr>
<tr>
<td>colombia</td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>brazil</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>export</td>
<td>0.0014</td>
<td></td>
</tr>
<tr>
<td>exporters</td>
<td>0.0013</td>
<td></td>
</tr>
<tr>
<td>exports</td>
<td>0.0013</td>
<td></td>
</tr>
<tr>
<td>crop</td>
<td>0.0012</td>
<td></td>
</tr>
</tbody>
</table>

Example: Manning, Raghavan, and Schütze, 2008

**Chi-squared**

- Hypothesis testing approach
- $H_0$: Term appearance is independent from a document’s class
  - i.e., $P(U=1,C=1) = P(U=1)P(C=1)$
- Compute:
  $$X^2(D, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$
- Or to directly plug in values like before:
  $$X^2(D, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})}$$
**Chi-squared**

\[ X^2(D,t,c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})} \]

- **Example**
  - What is the value of \( X^2 \) given the example data?

<table>
<thead>
<tr>
<th>( e_c = e_{poultry} )</th>
<th>( e_c = e_{poultry} = 1 )</th>
<th>( e_c = e_{poultry} = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e_t = e_{export} = 1 )</td>
<td>( N_{11} = 49 )</td>
<td>( N_{10} = 27,652 )</td>
</tr>
<tr>
<td>( e_t = e_{export} = 0 )</td>
<td>( N_{01} = 141 )</td>
<td>( N_{00} = 774,106 )</td>
</tr>
</tbody>
</table>
Dictionaries and Lexicons

• What if we know what we are looking for?
• Dictionaries (lexicons) are prebuilt mappings
  • Category -> word list
  • E.g., a tiny sentiment lexicon:
    • Positive: good, great, happy, amazing, wonderful, best, incredible
    • Negative: terrible, horrible, bad, awful, nasty, gross, worst, poor

• Domain can be important
  • “unpredictable movie plot” ✓
  • “unpredictable coffee pot” ❌

• How to get a score per category?

\[
\frac{\text{num\_dictionary\_words\_in\_document}}{\text{num\_total\_words\_in\_document}}
\]

• That’s it!
• Can also be used as machine learning features

• A more advanced approaches to quantifying categories
  (optional reading)
  • https://www.ncbi.nlm.nih.gov/pubmed/28364281
Some Dictionaries

- LIWC (Pennebaker et al. 2015)
- General Inquirer (Stone 1997)
- Roget’s Thesaurus Categories
- VADER (Hutto and Gilbert, 2014)
- Sentiwordnet (Esuli and Sebastiani 2006)
- Wordnet Domains (Magnini and Cavaglia, 2000)
- EmoLex (Mohammad and Turney, 2010)
- Empath (Fast et al., 2016)
- Personal Values Lexicon (Wilson et al., 2018)
- ...

Reactions to Rumor Tweets with EmoLex

Vosoughi, Roy, and Aral, 2018
**Dominance Scores**

- The dominance score for a category w.r.t. a corpus:

\[
\frac{\text{category\_score\_in\_target\_corpus}}{\text{category\_score\_in\_background\_corpus}}
\]

- From Mihalcea and Pulman, 2009

---

**LIWC category dominance scores**

<table>
<thead>
<tr>
<th>Truthful</th>
<th>Deceptive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interviews</strong></td>
<td><strong>Trials</strong></td>
</tr>
<tr>
<td>Class</td>
<td>Score</td>
</tr>
<tr>
<td>Metaphor</td>
<td>2.98</td>
</tr>
<tr>
<td>Money</td>
<td>2.74</td>
</tr>
<tr>
<td>Inhibition</td>
<td>2.74</td>
</tr>
<tr>
<td>Home</td>
<td>2.13</td>
</tr>
<tr>
<td>Humans</td>
<td>2.02</td>
</tr>
<tr>
<td>Family</td>
<td>1.96</td>
</tr>
</tbody>
</table>

Pérez-Rosas et al, 2015
Topic Level Analysis

Intro to Topic Modelling

• Goals are similar to traditional content analysis:
  • What are the main themes/topics in this corpus?
  • Which documents contain which topics?
Topic Models

The New York Times

Expected Soon: First-Ever Photo of a Black Hole
Have astronomers finally recorded an image of a black hole? The world will know on Wednesday.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Percent of Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astrophysics</td>
<td>40</td>
</tr>
<tr>
<td>Photography</td>
<td>30</td>
</tr>
<tr>
<td>Optimism</td>
<td>20</td>
</tr>
</tbody>
</table>

Steve Wilson, TTDS 2020/2021
Example from David Blei
**Dimensionality Reduction**

- **p** (number of words)
- **k** (number of topics)

**Data**

- **n**

**Data with Topic Model**

- **n**

**Topic Modeling**

- **Corpus**
- **Document-Topic Matrix**
  - **k**
  - **d**
- **Topic-Word Matrix**
  - **v**
  - **k**
Topic Models

- Most often used for text data, but can also be applied in other settings:
  - Bioinformatics (Liu et al. 2016)
  - Computer code (McBurney et al. 2014)
  - Music (Hu and Saul 2009)
  - Network data (Cha and Cho 2014)
Topic Modeling Methods

• Most popular: Latent Dirichlet Allocation (LDA)
  • Introduced by David Blei, Andrew Ng, and Michael Jordan (2003)

• Other methods include
  • pLSI
  • PCA-based methods
  • Non-negative matrix factorization
  • Deep learning based topic modeling
  • ...

Steve Wilson, TTDS 2020/2021
Latent Dirichlet Allocation (LDA)

- More details coming up in next lecture…