Text Technologies for Data Science
INFR11145

Comparing Text Corpora I

Instructor: Steve Wilson

13-Nov-2019

Pre-Lecture

• Today
  • Lecture: Comparing Text Corpora
**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, can be computed as:

$$I(U;C) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1}.N_{1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_{0}.N_{1}} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1}.N_{0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0}.N_{0}}$$

Source: Manning, Raghavan, and Schütze, 2008
Mutual Information

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

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

<table>
<thead>
<tr>
<th>(e_t = e_{export} = 1)</th>
<th>(e_t = e_{export} = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(e_c = e_{poultry} = 1)</td>
<td>(N_{11} = 49)</td>
</tr>
<tr>
<td>(e_c = e_{poultry} = 0)</td>
<td>(N_{01} = 141)</td>
</tr>
</tbody>
</table>

Example: Manning, Raghavan, and Schütze, 2008

Mutual Information for News Data

<table>
<thead>
<tr>
<th>UK</th>
<th>China</th>
<th>poultry</th>
</tr>
</thead>
<tbody>
<tr>
<td>coffee</td>
<td>bags</td>
<td>growers</td>
</tr>
<tr>
<td>0.0111</td>
<td>0.0042</td>
<td>0.0025</td>
</tr>
</tbody>
</table>

Example: Manning, Raghavan, and Schütze, 2008

Steve Wilson, TTDS 2019/2020
**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:
  \[
  X^2(D, t, c) = \sum_{\omega \in \{0,1\}} \sum_{\tau \in \{0,1\}} \frac{(N_{\omega\tau} - E_{\omega\tau})^2}{E_{\omega\tau}}
  \]
- Or to directly plug in values like before:
  \[
  X^2(D, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})}
  \]

**Example**
- What is the value of $X^2$ given the example data?

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

• How to get a score per category?

\[
\frac{\text{num}_{\text{dictionary\_words\_in\_document}}}{\text{num}_{\text{total\_words\_in\_document}}}
\]

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

• A more advanced approaches to quantifying categories (optional reading)

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></th>
<th>Truthful</th>
<th>Trials</th>
<th>Deceptive</th>
<th>Interviews</th>
<th>Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class</td>
<td>Score</td>
<td>Class</td>
<td>Score</td>
<td>Class</td>
</tr>
<tr>
<td>Metaphor</td>
<td>2.98</td>
<td>You</td>
<td>3.99</td>
<td>Assent</td>
<td>4.81</td>
</tr>
<tr>
<td>Money</td>
<td>2.74</td>
<td>Family</td>
<td>3.07</td>
<td>Past</td>
<td>2.59</td>
</tr>
<tr>
<td>Inhibition</td>
<td>2.74</td>
<td>Home</td>
<td>2.45</td>
<td>Sexual</td>
<td>2.00</td>
</tr>
<tr>
<td>Home</td>
<td>2.13</td>
<td>Humans</td>
<td>1.87</td>
<td>Other</td>
<td>1.87</td>
</tr>
<tr>
<td>Humans</td>
<td>2.02</td>
<td>Posemo</td>
<td>1.81</td>
<td>Motion</td>
<td>1.68</td>
</tr>
<tr>
<td>Family</td>
<td>1.96</td>
<td>Insight</td>
<td>1.64</td>
<td>Negemo</td>
<td>1.44</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

*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.
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)
**Dimensionality Reduction**

- **n** (number of documents) \(\rightarrow\) **p** (number of words) \(\rightarrow\) Data
- **k** (number of topics) \(\rightarrow\) Data with Topic Model

**Topic Modeling**

- **Corpus** \(\rightarrow\) **Topic Modeling Method**
- **d** \(\rightarrow\) **k** (Document-Topic Matrix)
- **v** \(\rightarrow\) **k** (Topic-Word Matrix)
**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
  - ...
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
  • ...

Reading

• Manning: IR book, section 13.5