Text Mining

Exercise 7
Text Preprocessing

• Tokenisation
  • Break text into single words or n-grams
  • “example text”
    • (“example”, “text”)
    • (“exam”, “xamp”, “ampl”, “mple”, “ple”, “le t”, “e te”, “ tex”, “text”)

• Stopword Removal
  • Remove frequent words that may confuse your algorithm
  • “this is an example” -> “example”

• Stemming
  • Finding the root/stem of a word helps matching similar words
  • “user”, “users”, “used”, “using” -> “use”
Operators: Tokenize

• Input Port
  • Document

• Output Port
  • Tokenised Document

• Parameters
  • Mode (how to create tokens)

• Splits a document into tokens
Operators: Filter Stopwords (English)

- Input Port
  - Tokenised Document

- Output Port
  - Tokenised Document

- Parameters
  - None

- Removes stopwords

- Different operators for different languages
Operators: Stem (Porter)

- Input Port
  - Tokenised Document

- Output Port
  - Tokenised Document

- Parameters
  - None

- Replaces tokens with their stems
- Different operators for different stemming methods
Feature Generation from Text

• Documents are treated as bags of words (tokens)
  • Each token becomes a feature
  • The order of tokens is ignored

• Different techniques to determine feature values (feature vector creation)
  • **Binary Term Occurrence**: 1 if the token is present, 0 otherwise
  • **Term Occurrence**: Absolute frequency of the token, i.e., 5
  • **Term Frequency**: Relative frequency of the token, i.e., 5%
  • **Term Frequency – Inverse Document Frequency**:
    • More weight if the token is rare
    • Less weight if the token is frequent
Feature Generation Examples – Binary Term Occurrences

• Sample document set:
  • d1 = “Saturn is the gas planet with rings.”
  • d2 = “Jupiter is the largest gas planet.”
  • d3 = “Saturn is the Roman god of sowing.”

• Documents as vectors:

<table>
<thead>
<tr>
<th></th>
<th>saturn</th>
<th>is</th>
<th>the</th>
<th>gas</th>
<th>planet</th>
<th>with</th>
<th>rings</th>
<th>jupiter</th>
<th>largest</th>
<th>roman</th>
<th>god</th>
<th>of</th>
<th>sowing</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Feature Generation Examples – Term Frequency

• Sample document set:
  • d1 = “Saturn is the gas planet with rings.”
  • d2 = “Jupiter is the largest gas planet.”
  • d3 = “Saturn is the Roman god of sowing.”

• Documents as vectors:

<table>
<thead>
<tr>
<th></th>
<th>saturn</th>
<th>is</th>
<th>the</th>
<th>gas</th>
<th>planet</th>
<th>with</th>
<th>rings</th>
<th>jupiter</th>
<th>largest</th>
<th>roman</th>
<th>god</th>
<th>of</th>
<th>sowing</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>1/7</td>
<td>1/7</td>
<td>1/7</td>
<td>1/7</td>
<td>1/7</td>
<td>1/7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>0</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>0</td>
<td>0</td>
<td>1/6</td>
<td>1/6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>1/7</td>
<td>1/7</td>
<td>1/7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/7</td>
<td>1/7</td>
<td>/71</td>
<td>1/7</td>
<td>1/7</td>
</tr>
</tbody>
</table>
Feature Generation Examples – TF-IDF

• Sample document set:
  • d1 = “Saturn is the gas planet with rings.”
  • d2 = “Jupiter is the largest gas planet.”
  • d3 = “Saturn is the Roman god of sowing.”

• Documents as vectors:

<table>
<thead>
<tr>
<th></th>
<th>saturn</th>
<th>is</th>
<th>the</th>
<th>gas</th>
<th>planet</th>
<th>with</th>
<th>rings</th>
<th>jupiter</th>
<th>largest</th>
<th>roman</th>
<th>god</th>
<th>of</th>
<th>sowing</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.07</td>
<td>0.07</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0</td>
<td>0.08</td>
<td>0.08</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Operators: Process Documents from Files

- **Input Port**
  - Word Vector (optional)

- **Output Ports**
  - Example Set (Vectorised Documents)
  - Word Vector

- **Parameters**
  - Directories: Which files to load & which label to assign
  - Vector creation method
  - Pruning (next slide)
Feature Selection

• High dimensional data!
• Not all features help!

• Pruning: Remove too frequent or too infrequent tokens
  • Percentual: ignore words that appear in less / more than a given percentage of all documents
  • Absolute: ignore words that appear in less / more than a given number of documents
  • By Rank: ignore a given percentage of the most frequent / infrequent words
Similarity Measures for Documents: Jaccard Coefficient

• Jaccard Coefficient:
  • For asymmetric binary attributes: the 1 state is more important than the 0 state
    \[ Jaccard(x_i, x_j) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \]

<table>
<thead>
<tr>
<th></th>
<th>saturn</th>
<th>is</th>
<th>the</th>
<th>gas</th>
<th>planet</th>
<th>with</th>
<th>rings</th>
<th>jupiter</th>
<th>largest</th>
<th>roman</th>
<th>god</th>
<th>of</th>
<th>sowing</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With stopwords</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaccard(D1,D2) = \frac{4}{2 + 3 + 4} = 0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaccard(D1,D3) = \frac{3}{4 + 4 + 3} = 0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaccard(D2,D3) = \frac{2}{5 + 4 + 2} = 0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without stopwords</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaccard(D1,D2) = \frac{2}{2 + 2 + 2} = 0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaccard(D1,D3) = \frac{1}{3 + 3 + 1} = 0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaccard(D2,D3) = \frac{0}{4 + 4 + 0} = 0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Similarity Measures for Documents: Cosine Similarity

- Cosine Similarity
  - Dot product only considers combinations that are both non-zero
  - Normalised by length of both vectors

\[
\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| \cdot |d_2|}
\]

<table>
<thead>
<tr>
<th></th>
<th>saturn</th>
<th>is</th>
<th>the</th>
<th>gas</th>
<th>planet</th>
<th>with</th>
<th>rings</th>
<th>jupiter</th>
<th>largest</th>
<th>roman</th>
<th>god</th>
<th>of</th>
<th>sowing</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.07</td>
<td>0.07</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0.08</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
</tr>
</tbody>
</table>

- **With stopwords**
  - \(\text{Cosine}(D1, D2) = 0.13\)
  - \(\text{Cosine}(D1, D3) = 0.05\)
  - \(\text{Cosine}(D2, D3) = 0.00\)

- **Without stopwords**
  - \(\text{Cosine}(D1, D2) = 0.17\)
  - \(\text{Cosine}(D1, D3) = 0.08\)
  - \(\text{Cosine}(D2, D3) = 0.00\)
Today’s Datasets

• Corpus 4-docs:
  • Doc1: “David Cameron Joins Talks On Euro Crisis”
  • Doc2: “Real Madrid Slips Into First With a Hat Trick by Ronaldo”
  • Doc3: “An Occupation for the 99 Per Cent” (Occupy Wall Street)
  • Doc4: “Málaga vs. Real Madrid Barcelona vs. Sevilla”

• Corpus 30-docs (newsgroups¹):
  • sci.space
  • soc.religion.christian
  • talk.politics.guns

• Corpus 300-docus:
  • misc.forsale
  • rec.sport.baseball
  • rec.sport.hockey

• Job Postings:
  • Category + Posting Text

¹ https://en.wikipedia.org/wiki/Usenet_newsgroup