Vector Space
Scoring
Introduction to Information Retrieval
CS 221
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org
Vector Space Model

- Define: Vector Space Model
  - Representing a set of documents as vectors in a common vector space.
  - It is fundamental to many operations
    - (query, document) pair scoring
    - document classification
    - document clustering
  - Queries are represented as a document
  - A short one, but mathematically equivalent
Vector Space Scoring

Vector Space Model

- Define: **Vector Space Model**
- A document, $d$, is defined as a vector: $\vec{V}(d)$
  - One component for each term in the dictionary
  - Assume the term is the tf-idf score

$$\vec{V}(d)_t = (1 + \log(tf_{t,d})) * \log\left( \frac{|corpus|}{df_{t,d}} \right)$$

- A corpus is many vectors together.
- A document can be thought of as a point in a multi-dimensional space, with axes related to terms.
## Vector Space Scoring

### Vector Space Model

- Recall our Shakespeare Example:

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>worser</td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Vector Space Scoring

Vector Space Model

- Recall our Shakespeare Example:

\[ \vec{V}(d_1) \]

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>worser</td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>
### Vector Space Model

- Recall our Shakespeare Example:

<table>
<thead>
<tr>
<th></th>
<th>( \vec{V}(d_1) )</th>
<th>( \vec{V}(d_2) )</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony and Cleopatra</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>worser</td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>
**Vector Space Model**

- Recall our Shakespeare Example:

<table>
<thead>
<tr>
<th></th>
<th>$\vec{V}(d_1)$</th>
<th>$\vec{V}(d_2)$</th>
<th>$\vec{V}(d_3)$</th>
<th>$\vec{V}(d_6)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antony</strong></td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Brutus</strong></td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>Caesar</strong></td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>Calpurnia</strong></td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Cleopatra</strong></td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>mercy</strong></td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>worse</strong></td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
</tr>
</tbody>
</table>

**Friday, February 5, 2010**
**Vector Space Scoring**

**Vector Space Model**

- Recall our Shakespeare Example:

<table>
<thead>
<tr>
<th></th>
<th>( \vec{V}(d_1) )</th>
<th>( \vec{V}(d_2) )</th>
<th>( \vec{V}(d_6) )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antony</strong></td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Brutus</strong></td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Caesar</strong></td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Calpurnia</strong></td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Cleopatra</strong></td>
<td>17.7</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td><strong>mercy</strong></td>
<td>0.5</td>
<td>0.0</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>worser</strong></td>
<td>1.2</td>
<td>0.0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**Antony and Cleopatra**

**Julius Caesar**

**The Tempest**

**Hamlet**

**Othello**

**Macbeth**

Friday, February 5, 2010
Vector Space Scoring

Vector Space Model

- Recall our Shakespeare Example:
Recall our Shakespeare Example:

\[
\begin{array}{cccccc}
\vec{V}(d_1) & \vec{V}(d_2) & \text{The Tempest} & \text{Hamlet} & \text{Othello} & \vec{V}(d_6) \\
\hline
\text{Antony} & 13.1 & 11.4 & 0.0 & 0.0 & 0.0 \\
\text{Brutus} & 3.0 & 8.3 & 0.0 & 1.0 & 0.0 \\
\text{Caesar} & 2.3 & 2.3 & 0.0 & 0.5 & 0.3 \\
\text{Calpurnia} & 0.0 & 11.2 & 0.0 & 0.0 & 0.0 \\
\text{Cleopatra} & 17.7 & 0.0 & 0.0 & 0.0 & 0.0 \\
\text{mercy} & 0.5 & 0.0 & 0.7 & 0.9 & 0.9 \\
\text{worse} & 1.2 & 0.0 & 0.6 & 0.6 & 0.6 \\
\end{array}
\]
Vector Space Scoring

Vector Space Model

- Recall our Shakespeare Example:

\[ \text{worser} \rightarrow \text{mercy} \]

- Antony and Cleopatra
- Tempest
- Othello
- Macbeth
- Julius Caesar
Vector Space Scoring

Query as a vector

- So a query can also be plotted in the same space
- "worser mercy"
- To score, we ask:
  - How similar are two points?
  - How to answer?
Vector Space Scoring

Score by magnitude

- How to answer?
- Similarity of magnitude?
- But, two documents, similar in content, different in length can have large differences in magnitude.
Vector Space Scoring

Score by angle

• How to answer?
  • Similarity of relative positions, or
  • difference in angle
    • Two documents are similar if the angle between them is 0.
    • As long as the ratios of the axes are the same, the documents will be scored as equal.
  • This is measured by the dot product

\[
\vec{V}(d_1) \cdot \vec{V}(d_2) \cdot \vec{V}(d_3) \cdot \vec{V}(d_4) \cdot \vec{V}(d_5)
\]
Vector Space Scoring

Score by angle

• Rather than use angle
• use cosine of angle
• When sorting cosine and angle are equivalent
• Cosine is monotonically decreasing as a function of angle over (0 ... 180)
Big picture

• Why are we turning documents and queries into vectors
  • Getting away from Boolean retrieval
  • Developing ranked retrieval methods
  • Developing scores for ranked retrieval
  • Term weighting allows us to compute scores for document similarity
• Vector space model is a clean mathematical model to work with
Big picture

- Cosine similarity measure
  - Gives us a symmetric score
    - if $d_1$ is close to $d_2$, $d_2$ is close to $d_1$
  - Gives us transitivity
    - if $d_1$ is close to $d_2$, and $d_2$ close to $d_3$, then
      - $d_1$ is also close to $d_3$
  - No document is closer to $d_1$ than itself
  - If vectors are normalized (length = 1) then
    - The similarity score is just the dot product (fast)
Queries in the vector space model

- Central idea: the query is a vector
- We regard the query as a short document
- We return the documents ranked by the closeness of their vectors to the query (also a vector)

\[ \text{sim}(q, d_i) = \frac{\vec{V}(q) \cdot \vec{V}(d_i)}{|\vec{V}(q)| |\vec{V}(d_i)|} \]

- Note that q is very sparse!
Cosine Similarity Score

- Also called *cosine similarity*

\[
\vec{V}(d_1) \cdot \vec{V}(d_2) = \cos(\theta) \frac{\|\vec{V}(d_1)\| \|\vec{V}(d_2)\|}{\|\vec{V}(d_1)\| \|\vec{V}(d_2)\|}
\]

\[
\cos(\theta) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{\|\vec{V}(d_1)\| \|\vec{V}(d_2)\|}
\]

\[
sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{\|\vec{V}(d_1)\| \|\vec{V}(d_2)\|}
\]
Cosine Similarity Score

- Define: *dot product*

\[ \vec{V}(d_1) \cdot \vec{V}(d_2) = \sum_{i=t_1}^{t_n} (\vec{V}(d_1)_i \vec{V}(d_2)_i) \]

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>worser</td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>

\[
\vec{V}(d_1) \cdot \vec{V}(d_2) = (13.1 \times 11.4) + (3.0 \times 8.3) + (2.3 \times 2.3) + (0 \times 11.2) + (17.7 \times 0) + (0.5 \times 0) + (1.2 \times 0) = 179.53
\]
Cosine Similarity Score

- Define: Euclidean Length

\[ |\vec{V}(d_1)| = \sqrt{\sum_{i=t_1}^{t_n} (\vec{V}(d_1)_i \vec{V}(d_1)_i)} \]

Vector Space Scoring

<table>
<thead>
<tr>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>8.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>2.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>worser</td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
</tr>
</tbody>
</table>

\[ |\vec{V}(d_1)| = \sqrt{(13.1 \times 13.1) + (3.0 \times 3.0) + (2.3 \times 2.3) + (17.7 \times 17.7) + (0.5 \times 0.5) + (1.2 \times 1.2)} \]
\[ = 22.38 \]
## Vector Space Scoring

### Cosine Similarity Score

- **Define:** Euclidean Length

\[
|\vec{V}(d_1)| = \sqrt{\sum_{i=t_1}^{t_n} (\vec{V}(d_1)_i \vec{V}(d_1)_i)}
\]

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>2.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>8.3</td>
<td>0.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>2.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>Worse</td>
<td>1.2</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>

\[
|\vec{V}(d_1)| = \sqrt{(11.4 \times 11.4) + (8.3 \times 8.3) + (2.3 \times 2.3) + (11.2 \times 11.2)}
\]

\[
= 18.15
\]
Cosine Similarity Score

- Example

\[
sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}
\]

\[
= \frac{179.53}{22.38 \times 18.15}
\]

\[
= 0.442
\]
Vector Space Scoring

Exercise

• Rank the following by decreasing cosine similarity.
  
  • Assume tf-idf weighting:
    
    • Two docs that have only frequent words in common
      
      • (the, a, an, of)
    
    • Two docs that have no words in common

    • Two docs that have many rare words in common
      
      • (mocha, volatile, organic, shade-grown)