Lesson 4: Descriptive Modelling of Similarity of Text
Unit 3: Vector space models for similarity

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties
Completing this unit you should …

- Be familiar with the vector space model for text documents
- Be aware of term frequency and (inverse) document frequency
- Have reviewed the definitions of base and dimension
- Realize that the angle between two vectors can be seen as a similarity measure
A model based on vector spaces

1) Model documents as vectors of words
A model based on vector spaces

1) Model documents as vectors of words

2) Calculating distance between vectors

Vector space Model → Distance between vectors
A model based on vector spaces

1) Model documents as vectors of words

2) Calculating distance between vectors

3) Interpreting

Vector space Model

Distance between vectors
A model based on vector spaces

1) Model documents as vectors of words

2) Calculating

3) Interpreting

Be able to make statements about similarity of documents

Vector space Model

Distance between vectors
Pay close attention to notation

• Words \( W = \{w_1, w_2, \ldots, w_n\} \)

• Word Vectors \( V = \langle \vec{w}_1, \vec{w}_2, \ldots, \vec{w}_n \rangle \)

• Document \( D_j \in W^* \) are a sequence of words
  – So \( D_j = w_{i_1} w_{i_2} \ldots w_{i_m} \) has a length of \( m \)

• Document vector \( \vec{d}_j = \sum_{i=1}^{n} tf(w_i, D_j) \vec{w}_i \)
Usually tf-idf is considered instead of tf!

- The document frequency is defined as

\[ df(w_i) = |\{D_j | w_i \text{ in } D_j\}| \]

- Inverse document frequency is defined as

\[ idf(w_i) = \log \frac{|D|}{df(w_i)} \]

resulting in

\[ tfidf(w_i, D_j) = tf(w_i, D_j) \times \log \frac{|D|}{df(w_i)} \]

- In the videos and slides for simplicity of numbers we will only use the term frequency
Example (generic language)

• Let us assume 3 documents
  – \( D_1 = a\ a\ a\ b\ b\ a\ a\ b\ a\ b \)
  – \( D_2 = a\ b\ b \)
  – \( D_3 = a\ a\ b \)

• In our artificial language we have just two words “a” and “b”

• Which documents are similar to each other?
Choose a vector space and base

• Let \( V = \langle \vec{a}, \vec{b} \rangle \) be the vector space spanned by the words “a” and “b”

• \( \vec{a} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \) is a base vector for word “a”

• Similarly for word “b” we have \( \vec{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \)

• Chosing the base was a modelling choice!
Calculate the modelled document vectors $\mathbf{d}_1$

- $D_1 = \text{a a a b b a a b a b}$
  - $tf(a, D_1) = 6$
  - $tf(b, D_1) = 4$

- Let us now create the document vectors

\[
\mathbf{d}_1 = \sum_{i=1}^{2} tf(w_i, D_1)\mathbf{w}_i = tf(a, D_1)\mathbf{a} + tf(b, D_1)\mathbf{b}
\]

\[
= 6 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 4 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}
\]
Calculate the modelled document vectors \( \mathbf{d}_2 \)

- \( \mathbf{D}_2 = \mathbf{a} \mathbf{b} \mathbf{b} \)
  - \( t\mathbf{f}(a, \mathbf{D}_2) = 1 \)
  - \( t\mathbf{f}(b, \mathbf{D}_2) = 2 \)
- Let us now create the document vectors

\[
\mathbf{d}_2 = \sum_{i=1}^{2} t\mathbf{f}(w_i, \mathbf{D}_2)\mathbf{w}_i = t\mathbf{f}(a, \mathbf{D}_2)\mathbf{a} + t\mathbf{f}(b, \mathbf{D}_2)\mathbf{b}
\]

\[
= 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}
\]
Calculate the modelled document vectors $d_3$

- $D_3 = \text{a a b}$
  - $tf(a, D_3) = 2$
  - $tf(b, D_3) = 1$

Let us now create the document vectors

$$d_3 = \sum_{i=1}^{2} tf(w_i, D_3) \vec{w}_i = tf(a, D_3)\vec{a} + tf(b, D_3)\vec{b}$$

$$= 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$
Digression: Drawing the document vectors

- $D_1 = \text{a a a b b a a b a b}$, $D_2 = \text{a b b}$, $D_3 = \text{a a b}$

\[
\vec{d}_1 = \begin{pmatrix} 6 \\ 4 \end{pmatrix}, \quad \vec{d}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \vec{d}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}
\]
Which vectors are closest too each other?

- \( D_1 = a \ a \ a \ b \ b \ a \ a \ b \ \ a \ b \ \ a \ b \), \( D_2 = a \ b \ b \), \( D_3 = a \ a \ b \)

\[
\vec{d}_1 = \begin{pmatrix} 6 \\ 4 \end{pmatrix}, \quad \vec{d}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \vec{d}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}
\]
Two ways of calculating the distance between two vectors \( \mathbf{d}_i \) and \( \mathbf{d}_j \)?

- **Euclidean distance**
  - “Take a rule and measure”
  - Take difference: \( \mathbf{d} = \mathbf{d}_i - \mathbf{d}_j \)
  - Calculate length of difference \( ||\mathbf{d}|| \)

- **Cosine distance**
  - “Take the angle between vectors”
Calculate Euclidean distance between two document vectors \( d_i \) and \( d_j \)

- Take the difference \( \vec{d} = d_i - d_j \)
- Calculate the length of the difference

\[
||\vec{d}||^2 = \sum_{k=1}^{n} (d_k)^2 = \sum_{k=1}^{n} ((d_i)_k - (d_j)_k)^2
\]
Calculate Euclidean distance between two document vectors $d_i$ and $d_j$

- Take the difference $d' = d_i - d_j$
- Calculate the length of the difference

$$\|d'\|^2 = \sum_{k=1}^{n} (d_k)^2 = \sum_{k=1}^{n} ((d_i)_k - (d_j)_k)^2$$

The k-th component of vector $d_i$

$$(d_i)_k = tf(w_k, D_i)$$ by definition
Calculate Euclidean distance between two document vectors $d_i$ and $d_j$

- Take the difference $\vec{d} = \vec{d}_i - \vec{d}_j$
- Calculate the length of the difference

$$||\vec{d}||^2 = \sum_{k=1}^{n} (d_k)^2 = \sum_{k=1}^{n} ((\vec{d}_i)_k - (\vec{d}_j)_k)^2$$

$$= \sum_{k=1}^{n} (tf(w_k, D_i) - tf(w_k, D_j))^2$$

- For every word $w_k$ we compare how often it appears in document $D_i$ and document $D_j$
Calculate Euclidean distance between two document vectors $d_i$ and $d_j$

- Take the difference $\tilde{d} = d_i - d_j$
- Calculate the length of the difference

$$||\tilde{d}||^2 = \sum_{k=1}^{n} (d_k)^2 = \sum_{k=1}^{n} ((d_i)_k - (d_j)_k)^2$$

$$= \sum_{k=1}^{n} (tf(w_k, D_i) - tf(w_k, D_j))^2$$

- For every word $w_k$ we compare how often it appears in document $D_i$ and document $D_j$
Euclidean distances for our example

- $D_1 = a \ a \ a \ b \ b \ a \ a \ b \ a \ b$,  
- $D_2 = a \ b \ b$,  
- $D_3 = a \ a \ b$

\[
\| \vec{d}_1 - \vec{d}_2 \|_e = 5.4 \quad \| \vec{d}_1 - \vec{d}_3 \|_e = 5 \quad \| \vec{d}_2 - \vec{d}_3 \|_e = 1.4
\]
Calculate Cosine distance between two document vectors \( \mathbf{d}_i \) and \( \mathbf{d}_j \)

- Calculate the scalar product \( s = \langle \mathbf{d}_i, \mathbf{d}_j \rangle \)

\[
\langle \mathbf{d}_i, \mathbf{d}_j \rangle = \sum_{k=1}^{n} (\mathbf{d}_i)_k (\mathbf{d}_j)_k \\
= \sum_{k=1}^{n} tf(w_k, D_i) tf(w_k, D_j)
\]

Remember: \((\mathbf{d}_i)_k = tf(w_k, D_i)\) is zero most of the time.
Calculate Cosine distance between two document vectors $d_i$ and $d_j$

- Calculate the scalar product $s = \langle \vec{d}_i, \vec{d}_j \rangle$
- Divide it by the product of lengths of both vectors (length as in Euclidean distance)

$$\cos(\theta) = \frac{\langle \vec{d}_i, \vec{d}_j \rangle}{||\vec{d}_i|| \ast ||\vec{d}_j||}$$

$$\Rightarrow \theta = \cos^{-1} \left( \frac{\langle \vec{d}_i, \vec{d}_j \rangle}{||\vec{d}_i|| \ast ||\vec{d}_j||} \right)$$
Cosine distances for our example

- $D_1 = a \ a \ a \ b \ b \ a \ a \ b \ a \ b$
- $D_2 = a \ b \ b$
- $D_3 = a \ a \ b$

\[
\| \vec{d}_1 - \vec{d}_2 \|_c = 0.52 \\
\| \vec{d}_1 - \vec{d}_3 \|_c = 0.12 \\
\| \vec{d}_2 - \vec{d}_3 \|_c = 0.64
\]

$\vec{d}_1 = (6, 4)$
$\vec{d}_2 = (1, 2)$
$\vec{d}_3 = (2, 1)$
Comparing cosine and Euclidean distance

- $D_1 = a\ a\ a\ b\ b\ a\ a\ b\ a\ b$, $D_2 = a\ b\ b$, $D_3 = a\ a\ b$

<table>
<thead>
<tr>
<th></th>
<th>Euklid</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td></td>
<td>\vec{d}_1 - \vec{d}_2</td>
</tr>
<tr>
<td>$</td>
<td></td>
<td>\vec{d}_2 - \vec{d}_3</td>
</tr>
<tr>
<td>$</td>
<td></td>
<td>\vec{d}_1 - \vec{d}_3</td>
</tr>
</tbody>
</table>

- Different choices of metric can yield very different results
- Choice of metric is part of the model!
- Usually cosine distance is considered
Thank you for your attention!

Contact:
Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de
Copyright:

- This Slide deck is licensed under creative commons 3.0. share alike attribution license. It was created by Rene Pickhardt. You can use share and modify this slide deck as long as you attribute the author and keep the same license.
- https://commons.wikimedia.org/wiki/File:Synoptic_word-for-word.png By Alecmonroy (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY 3.0 (http://creativecommons.org/licenses/by/3.0)], via Wikimedia Commons