Data Mining and Machine Learning

Vector Representation of Documents

Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Objectives

- To explain vector representation of documents

- To understand cosine distance between vector representations of documents

Vector Notation for Documents

- Suppose that we have a set of documents

  $D = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \dots, \boldsymbol{d_N}\}$

  think of this as the corpus for IR

- Suppose that the number of different words in the whole corpus is V (vocabulary size)

- Now suppose a document $\boldsymbol{d}$ in D contains $\boldsymbol{M}$ different terms: $\{\boldsymbol{t_{i(1)}}, \boldsymbol{t_{i(2)}}, \dots, \boldsymbol{t_{i(M)}}\boldsymbol{)}$
- Finally, suppose term $\boldsymbol{t_{i(m)}}$ occurs $\boldsymbol{f_{i(m)}}$ times

Vector Notation

- The vector representation ***vec*(*d*)** of ***d*** is the ***V*** dimensional vector:

$$\left(0,\ldots,0,w_{i(1),d},0,\ldots,0,w_{i(2),d},0,\ldots,0,w_{i(M),d},0,\ldots,0\right)$$

$i(1)^{th}$ place

$i(2)^{th}$ place

$i(M)^{th}$ place

Notice that this is the weighting – i.e. the term frequency times the inverse document frequency $w_{i(1),d} = f_{i(1),d} \times IDF(i(1))$ from text IR

Uniqueness

- Is the mapping between documents and vectors one-to-one?
- In other words:
  - if d_1 , d_2 are documents, is it true that $vec(d_1) = vec(d_2)$ if and only if $d_1 = d_2$?
- If λ is a scalar and $vec(d_1) = \lambda vec(d_2)$ what does this tell you about d_1 and d_2?

Example

- $\boldsymbol{d_1}$ = the cat sat on the cat's mat → cat sat cat mat
- $\boldsymbol{d_2}$ = the dog chased the cat → dog chase cat
- $\boldsymbol{d_3}$ = the mouse stayed at home → mouse stay home
- Vocabulary:
  - cat, chase, dog, home, mat, mouse, sat, stay
- To calculate the vector representations of these documents first calculate the TF-IDF weights

Example (continued)

	d1	d2	d3	Nd	IDF	w(t,d1)	w(t,d2)	w(t,d3)
cat	2	1		2	0.41	0.81	0.41	
chase		1		1	1.1		1.1	
dog		1		1	1.1		1.1	
home			1	1	1.1			1.1
mat	1			1	1.1	1.1		
mouse			1	1	1.1			1.1
sat	1			1	1.1	1.1		
stay			1	1	1.1			1.1

Example (continued)

$$vec(d_1) = \begin{bmatrix} 0.81 \\ 0 \\ 0 \\ 0 \\ 1.1 \\ 0 \\ 1.1 \\ 0 \end{bmatrix} \quad vec(d_2) = \begin{bmatrix} 0.41 \\ 1.1 \\ 1.1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad vec(d_3) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1.1 \\ 0 \\ 1.1 \\ 0 \\ 1.1 \end{bmatrix}$$

Document length revisited

- Recall that the length of a vector

$$x = (x_1, \ldots, x_N)$$

is given by:

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_N^2}$$

Document length

- In the case of a ‘document vector’

$$vec(d) = \left(0,\ldots,0, w_{i(1)d}, 0,\ldots,0, w_{i(2)d}, 0,\ldots\ldots, w_{i(M)d}, 0\ldots,0\right)$$

$$\left\| vec(d) \right\| = \sqrt{w_{i(1)d}^2 + w_{i(2)d}^2 + \ldots + w_{i(M)d}^2} = \left\| d \right\|$$

Document Similarity

- Suppose ***d*** is a document and ***q*** is a query
  - If ***d*** and ***q*** contain the same words in the same proportions, then ***vec***(***d***) and ***vec***(***q***) will point in the same direction
  - If ***d*** and ***q*** contain different words, then ***vec***(***d***) and ***vec***(***q***) will point in different directions
  - Intuitively, the greater the angle between ***vec***(***d***) and ***vec***(***q***) the less similar the document ***d*** is with the query ***q***

Cosine similarity

- Define the **Cosine Similarity** between document $\boldsymbol{d}$ and query $\boldsymbol{q}$ by:

$$CSim(\boldsymbol{q},\boldsymbol{d}) = \cos\theta$$

where θ is the <u>angle</u> between $\boldsymbol{vec}(\boldsymbol{q})$ and $\boldsymbol{vec}(\boldsymbol{d})$

- Similarly, define the **Cosine Similarity** between documents $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ by:

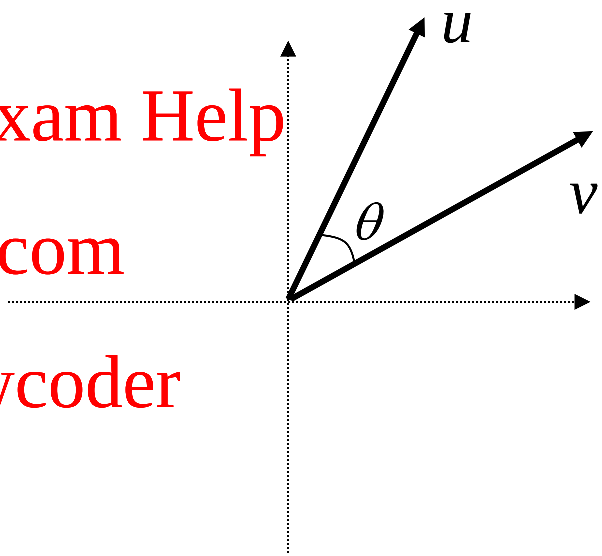
$$CSim(\boldsymbol{d_1},\boldsymbol{d_2}) = \cos\theta$$

where θ is the <u>angle</u> between $\boldsymbol{vec}(\boldsymbol{d_1})$ and $\boldsymbol{vec}(\boldsymbol{d_2})$

Cosine Similarity & Similarity

- Let $u=(x_1,y_1)$ and $v=(x_2,y_2)$ be vectors in 2 dimensions, then

$$\cos(\theta) = \frac{x_1 x_2 + y_1 y_2}{\|u\|\|v\|} = \frac{u \cdot v}{\|u\|\|v\|}$$



- In fact, this result holds for vectors in any N dimensional space

Cosine Similarity & Similarity

- Hence, if ***q*** is a query, ***d*** is a document, and θ is the angle between ***vec***(***q***) and ***vec***(***d***), then:

Cosine similarity

$$CSim(q,d) = \cos(\theta) = \frac{vec(q) \cdot vec(d)}{\|q\|\|d\|} = \frac{\sum_{t \in q \cap d} w_{tq} \cdot w_{td}}{\|q\|\|d\|}$$

$$= Sim(q,d)$$

Similarity

Summary

- Vector space representation of documents

- Cosine distance between vector representations of documents