Intelligent Data Analysis 2020

Lecture 4

Vector Representation of Documents

Martin Russell

UNIVERSITY OF BIRMINGHAM

Objectives

- To explain **vector representation** of documents
- To understand **cosine distance** between vector representations of documents
- To understand, intuitively, how **Latent Semantic Analysis (LSA)** can
  – Discover latent **topics** in a corpus and represent them in terms of words
  – Achieve dimension reduction for document vectors
  – Represent words in terms of topics

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

Vector Notation

- The **vector representation** $\boldsymbol{vec(d)}$ of $\boldsymbol{d}$ is the $\boldsymbol{V}$ dimensional vector:

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

$i(1)^{\text{th}}$ place

$i(2)^{\text{th}}$ place

$i(M)^{\text{th}}$ place

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

$\text{freq}(\cdot)\cdot[\ln(ND) - \ln(ND_t)]$.

- $vec(d)$ is the **document vector** for d

UNIVERSITY OF BIRMINGHAM

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 $d_1 = d_2$? False
- If λ is a scalar and $vec(d_1) = \lambda vec(d_2)$ what does this tell you about d_1 and d_2?

$vec(d_1) = 2\,vec(d_2)$

twice

Example

- d_1 = the cat sat on the cat's mat → cat sat mat cat
- d_2 = the dog chased the cat → dog chase cat
- d_3 = the mouse stayed at home → mouse stay home
- Vocabulary: Corpus $M = 8$.
  - cat, chase, dog, home, mat, mouse, sat, stay
- To calculate the vector representations of these documents first calculate the TF-IDF weights

Example (continued)

				# doc		TF - IDF		
	d1	d2	d3	Nd	IDF	w(t,d1)	w(t,d2)	w(t,d3)
cat	2	1		~~3~~ 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)

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

Document length revisited

- Recall that the **length (norm)** of a vector

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

is given by:

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

UNIVERSITY OF BIRMINGHAM

Document length

- In the case of a **document vector**

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

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

- Where $\|d\|$ is the length of the document d from last week's lecture

Document Similarity

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

$\angle \uparrow$, sim( ) $\downarrow$

UNIVERSITY OF
BIRMINGHAM

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 **angle** 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:

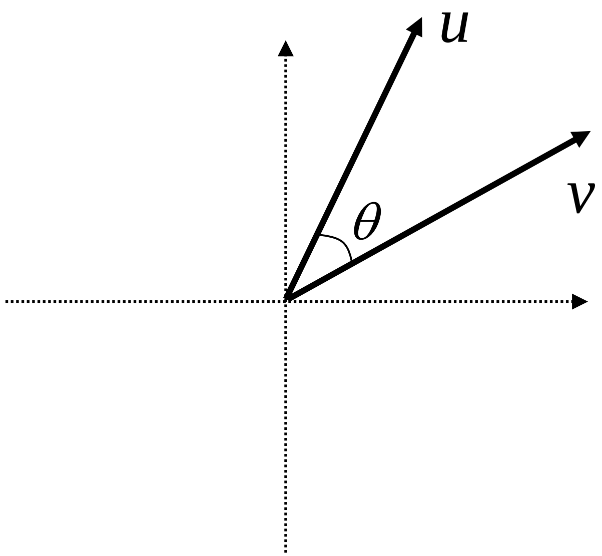
  $$CSim(\boldsymbol{d_1},\boldsymbol{d_2}) = \cos\theta$$

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

Cosine Similarity & Similarity

- Recall that if $u=(x_1,y_1)$ and $v=(x_2,y_2)$ are vectors in 2 dimensions, then

$$\cos(\theta) = \frac{x_1x_2 + y_1y_2}{\|u\|\|v\|} = \frac{u \cdot v}{\|u\|\|v\|}$$



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

Cosine Similarity & Similarity

- Hence, if $\boldsymbol{q}$ is a query, $\boldsymbol{d}$ is a document, and θ is the angle between $\boldsymbol{vec}(\boldsymbol{q})$ and $\boldsymbol{vec}(\boldsymbol{d})$, 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

- Vectorisation of documents
- Cosine similarity is equivalent to TF-IDF similarity
- Document length revisited