https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018
Indian Institute of Information Technology, Sri City

Life is like Google. You just need to know what you are looking for.
– Source Unknown.

Boolean Retrieval

- A model of IR in which
  - Query is a Boolean expression of terms (for example, t1 AND t2 OR t3 AND NOT t4)
- Each document is a set of words

What is a Set?

- {1, 2, 3}
- {A, B, C}
- {apple, banana, orange}
- {}

Which of the Following are Sets?

- {1, 2, 3, 4, 5, 6, 5, 7, 8, 9, 10, 11, 12, 13}
- {A, B, C, D, E, F, G, H, I, I, J, K, L, M, N, O}
- {apple, banana, orange, apple, banana, orange}

Bag

- {1, 2, 3, 4, 5, 6, 5, 7, 8, 9, 10, 11, 12, 13}
- {A, B, C, D, E, F, G, H, I, I, J, K, L, M, N, O}
- {apple, banana, orange, apple, banana, orange}

Quiz

- In our simple **Boolean Retrieval** model, we modeled documents as a set of words.
  - True
  - False

Quiz

- In our simple **Boolean Retrieval** model, we modeled documents as a set of words.
  - **True**
  - False

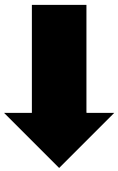
Boolean Retrieval

- In the **Boolean Retrieval** model, we modeled documents as a **set** of words.
  - True
  - False
- Let query q = "IIIT Sri City". Let document, d1 = "IIIT Sri City" and d_2 = "IIIT Delhi". Our term-document matrix looked as follows:

	IIIT	Sri	City	Delhi
q	1	1	1	0
d_1	1	1	1	0
d_2	1	0	0	1

Set of Words Representation

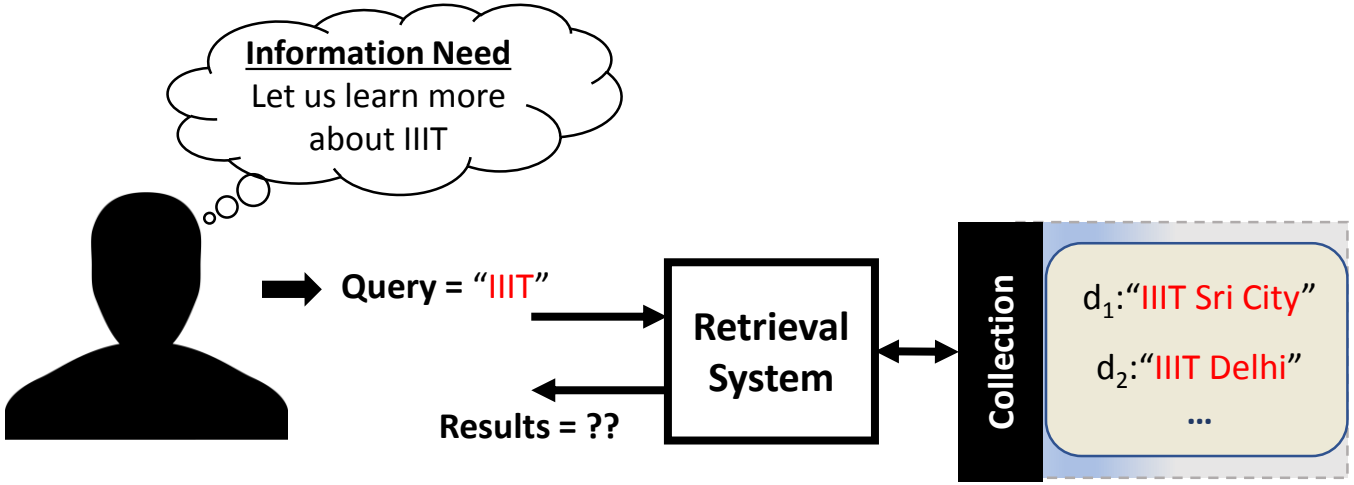
- "IIIT Sri City" → {IIIT, Sri, City}
- "IIIT Sri City, Sri City" → {IIIT, Sri, City}



	IIIT	Sri	City
q	1	1	1

Leads to same term-document matrix

Ranked Retrieval



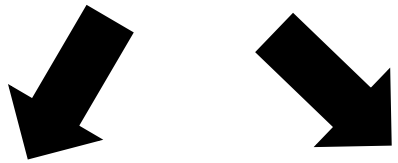
Results are ranked. But, which IIIT should come first?

Boolean Retrieval

In the **Ranked Retrieval** model, we may want to model documents as **bag** of words.

Bag of Words Representation

- "IIIT Sri City" → {IIIT, Sri, City}
- "IIIT Sri City, Sri City" → [IIIT, Sri, Sri, City, City]



IIIT Sri City

	IIIT	Sri	City
q	1	1	1

IIIT Sri City, Sri City

	IIIT	Sri	City
q	1	2	2

Leads to different term-document matrix

Set Similarity

- Similarity between two sets is easy

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard Similarity

Quiz

- What is the Jaccard similarity between
  - {1,2,3} and {4,5,6}?
  - {1,2,3} and {1,2,4}?
  - {1,2,3} and {1,2,3}?

Quiz

- What is the Jaccard similarity between
  - {1,2,3} and {4,5,6} = 0
  - {1,2,3} and {1,2,4} = $\frac{|\{1,2\}|}{|\{1,2,3,4\}|}$ = 0.5
  - {1,2,3} and {1,2,3} = 1

Quiz

- What is the Jaccard similarity between
  - “IIIT is Great” and “IIITD is Great”?

Quiz

- What is the Jaccard similarity between
  - “IIIT is Great” and “IIITD is Great”?
- Same as Jaccard Similarity between {IIIT, is, Great} and {IIITD, is, Great}. Equals 0.5.

Quiz

- If we represent sentences as vectors, what is the Jaccard similarity between
  - (1,0,1,1) and (0,1,1,1) ?

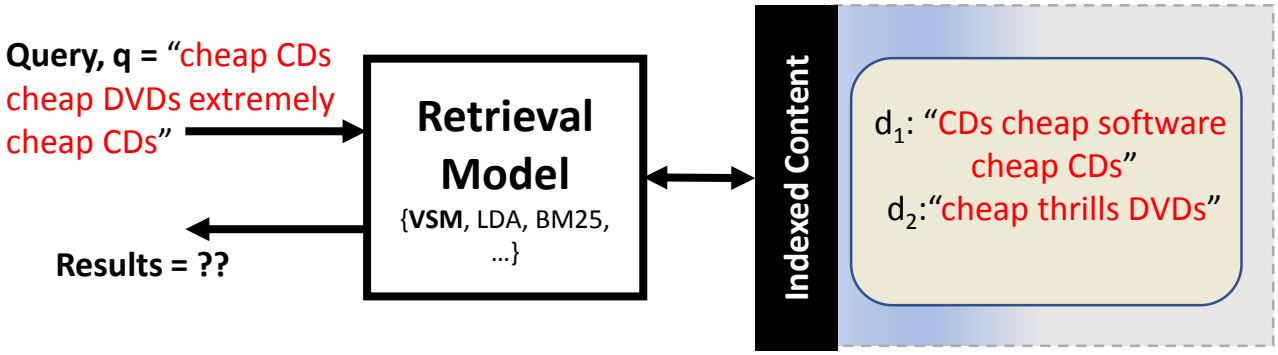
Quiz

- If we represent sentences as vectors, what is the Jaccard similarity between
  - (1,0,1,1) and (0,1,1,1) ?

- Same as before (Answer = 0.5):

	IIIT	IIITD	is	Great
d_1	1	0	1	1
d_2	0	1	1	1

Which Document to Retrieve?

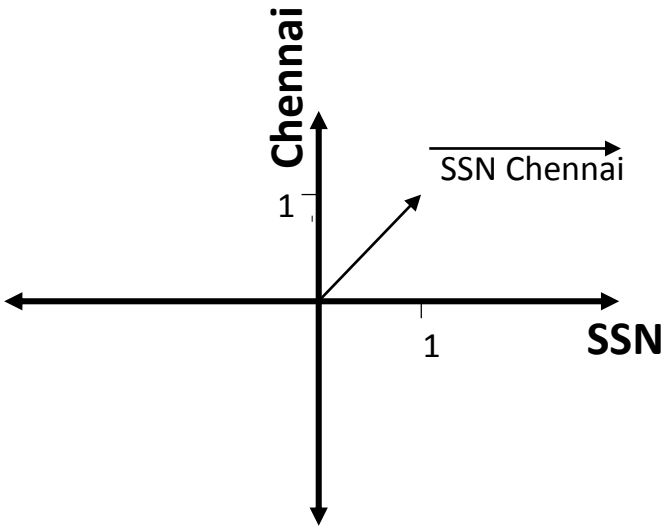


	cheap	CDs	DVDs	extremely	software	thrills
q	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1

How to Score Documents for a Query?

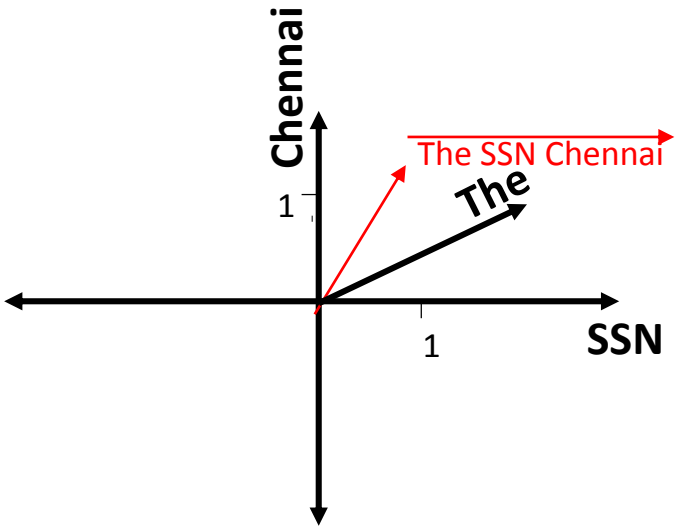
Sentences are vectors

- “SSN Chennai” as a vector



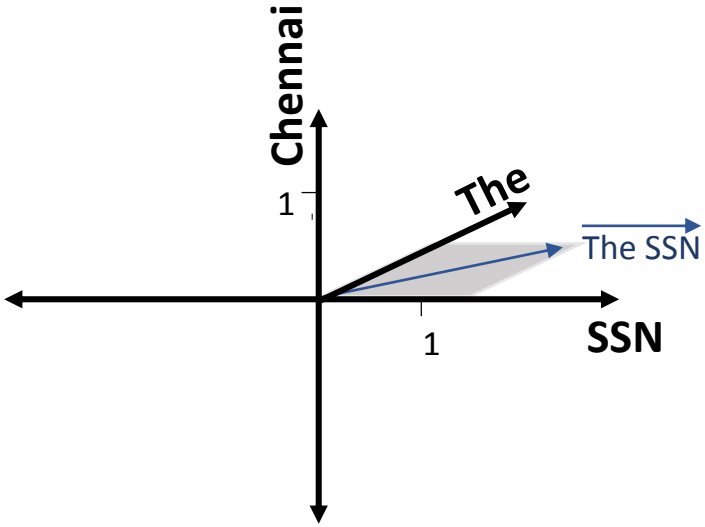
Sentences are vectors

- "The SSN Chennai" is a 3-dimensional vector



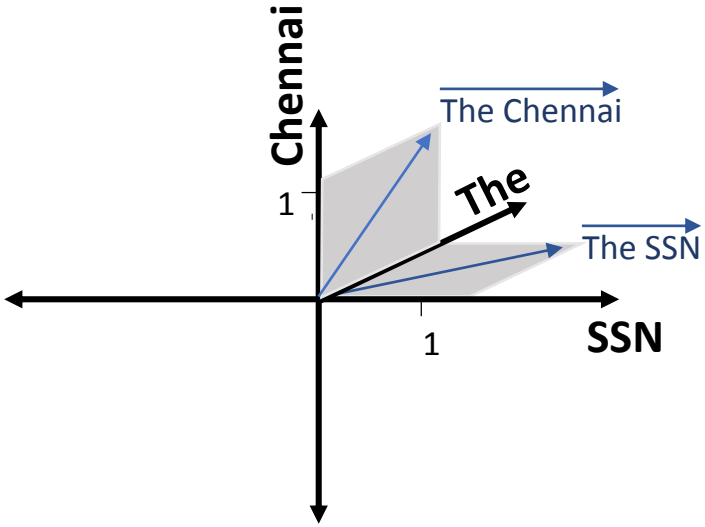
Sentences are vectors

- On this 3D space, “The SSN” vector will lie on the x (The) and z (SSN) plane.



Comparing Sentences

- We can compare sentences using the angle between vectors

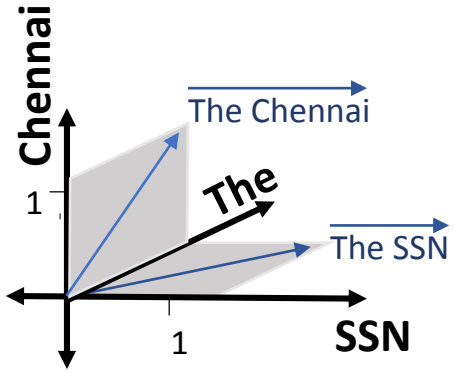


Angle between two vectors

- What is the angle between $\overrightarrow{\text{The}}$ and $\overrightarrow{\text{SSN}}$ vectors?
- What is the angle between $\overrightarrow{\text{SSN}}$ and $\overrightarrow{\text{Chennai}}$ vectors?
- What is the angle between $\overrightarrow{\text{The SSN}}$ and $\overrightarrow{\text{The SSN}}$ vectors?

Mathematical Notation

- We represent vectors as follows:
  - Vector = (dimension1, dimension2, dimension3, …)
    - First, define the dimensions
    - Next, put “1” if the word is present in the sentence, else “0”
- Example



Vector = (dimension1, d2, d3, …)

In our case,
vector = (The, SSN, Chennai)

So,

$\overrightarrow{\text{The Chennai}} = (1,0,1)$

$\overrightarrow{\text{The SSN}} = (1,1,0)$

Similarity Score

- D1 = “Chennai”
- D2 = “Chennai”

- Quiz
  - On a scale of 0 – 1, how similar are D1 and D2?
    - 0 ➔ Dissimilar
    - 1 ➔ Identical

Similarity Score

- D1 = “Chennai”
- D2 = “IIIT”

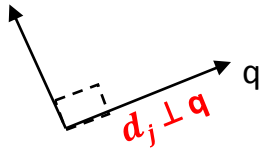
- Quiz
  - What is the angle between D1 and D2 vectors?
  - On a scale of 0 – 1, how similar are D1 and D2?

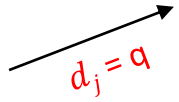
How to Convert 0 to 90 → 1 to 0

Revisiting Trigonometry

Converting from "0 – 90" to "1 – 0"

- For convenience, We convert the angles 0 – 90 to values 1 - 0
  - When vectors are same, we want to output 1.
  - When vectors are perpendicular we want to output 0.





0 – 90 to 1 – 0: How?

	0°	30°	45°	60°	90°
sin θ	0	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{3}}{2}$	1
cos θ	1	$\frac{\sqrt{3}}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	0
tan θ	0	$\frac{1}{\sqrt{3}}$	1	$\sqrt{3}$	Not defined

Back to Trigonometry

- If x and y are non-unit vectors, what is the cosine of angle between them (cos Θ)?

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos(\theta)$$

Matching Documents to Queries

- Document as a vector of term-weights

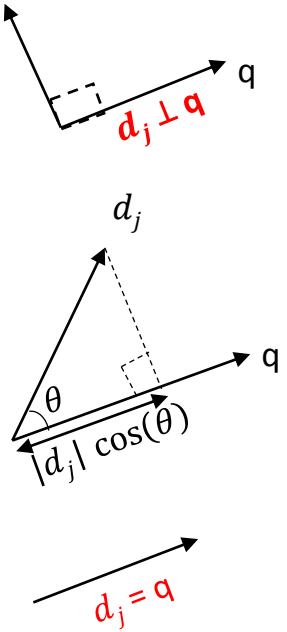
$$d_j = (w_{1j}, w_{2j}, \ldots w_{nj})$$

- Query as a vector of term-weights

$$q = (w_1, w_2, \ldots wm)$$

- Similarity between these vectors can be represented as

$$Cosine\ Similarity = \cos(\theta) = \frac{d_j . q}{||d_j||\ ||q||}$$



Example

Let query q = "BITS Pilani".

Let document, d_1 = "BITS Pilani Goa Campus" and d_2 = "IIIT Delhi".

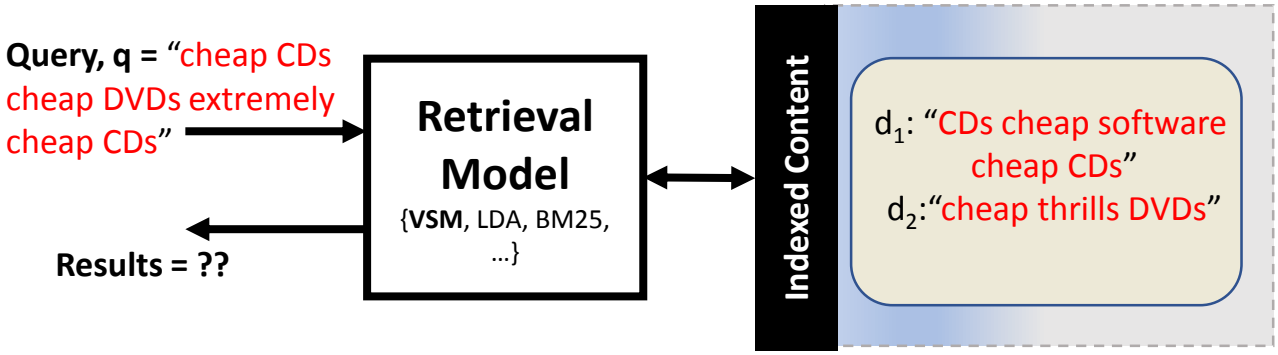
	BITS	Pilani	Goa	Campus	IIIT	Delhi
q	1	1	0	0	0	0
d_1	1	1	1	1	0	0
d_2	0	0	0	0	1	1

In our VSM, q = (1,1,0,0,0,0), d_1= (1,1,1,1,0,0) and d_2 = (0,0,0,0,1,1)

similarity(d_1, q) = $$\frac{d_1.q}{||d_1||\,||q||} = \frac{1.1 + 1.1}{\sqrt{1^2+1^2+1^2+1^2}\,\sqrt{1^2+1^2}} = 0.71.$$

similarity(d_2, q) = $$\frac{d_2.q}{||d_2||\,||q||} = 0.$$

Which Document to Retrieve?



	cheap	CDs	DVDs	extremely	software	thrills
q	3	2	1	1	0	0
d_1	2	2	0	0	1	0
d_2	1	0	1	0	0	1

$sim(q,d_1) = 0.86$

$sim(q,d_2) = 0.59$

Quiz

- What is the dot product of two vectors (1,0,1,1,1) and (1,1,1,0,0) ?

(1 * 1) + (0 * 1) + (1 * 1) + (1 * 0)+ (1 * 0) = 2

Quiz

- What is the dot product of two vectors (1,1,1,1,1) and (0,0,0,0,0) ?

1.0 + 1.0 + 1.0 + 1.0 + 1.0 = 0

Using . to represent multiplication.

Quiz

- What is the dot product of two vectors (1,0,1,0,1) and (0,1,0,1,0) ?

1.0 + 0.1 + 1.0 + 0.1 + 1.0 = 0

Quiz

- What is the dot product of two vectors (2,2,2,2,2) and (1,1,1,1,1) ?

2.1 + 2.1 + 2.1 + 2.1 + 2.1 = 5(2.1) = 10

Back to Trigonometry

- If x and y are non-unit vectors, what is the cosine of angle between them (cos Θ)?

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos(\theta)$$

Cosine Similarity

Let query q = "BITS Pilani".

Let document, d_1 = "BITS Pilani Goa Campus" and d_2 = "IIIT Delhi".

	BITS	Pilani	Goa	Campus	IIIT	Delhi
q	1	1	0	0	0	0
d_1	1	1	1	1	0	0
d_2	0	0	0	0	1	1

In our VSM, q = (1,1,0,0,0,0), d_1= (1,1,1,1,0,0) and d_2 = (0,0,0,0,1,1)

similarity(d_1, q) = $\frac{d_1.q}{||d_1||\,||q||} = \frac{1.1 + 1.1}{\sqrt{1^2+1^2+1^2+1^2}\,\sqrt{1^2+1^2}} = 0.71.$

similarity(d_2, q) = $\frac{d_2.q}{||d_2||\,||q||} = 0.$

Cosine Similarity

What is the similarity between the following sentences?

“**Julie loves me more than Linda loves me**”

“**Jane likes me more than Julie loves me**”

Step 1 – Term Frequency Computation

- Distinct terms from both sentences:
  - me Julie loves Linda than more likes Jane

Term	Frequency in d1	Frequency in d2
me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1

Step 2 – Compute Cosine Similarity

$$\textbf{similarity}(\mathbf{d_1}, \mathbf{d_2}) = \frac{d_1 . d_2}{||d_1||\,||d_2||} = \frac{4+0+1+0+0+2+1+1}{||d_1||\,||d_2||}$$

L2 norm of d1 = $||d_1|| = \sqrt{2^2+0^2+1^2+1^2+0^2+2^2+1^2+1^2} = \sqrt{12}$

L2 norm of d2 = $||d_2|| = \sqrt{2^2+1^2+1^2+0^2+1^2+1^2+1^2+1^2} = \sqrt{10}$

Putting it all together, $\textbf{similarity}(\mathbf{d_1}, \mathbf{d_2}) = \frac{9}{\sqrt{12}\sqrt{10}} = 0.82$

*Read about Least Absolute Deviations and Least Squares