Big Data – Text Processing Principles and Practices

Venkatesh Vinayakarao

http://vvtesh.co.in

Big Data is Ubiquitous

Big Data Landscape

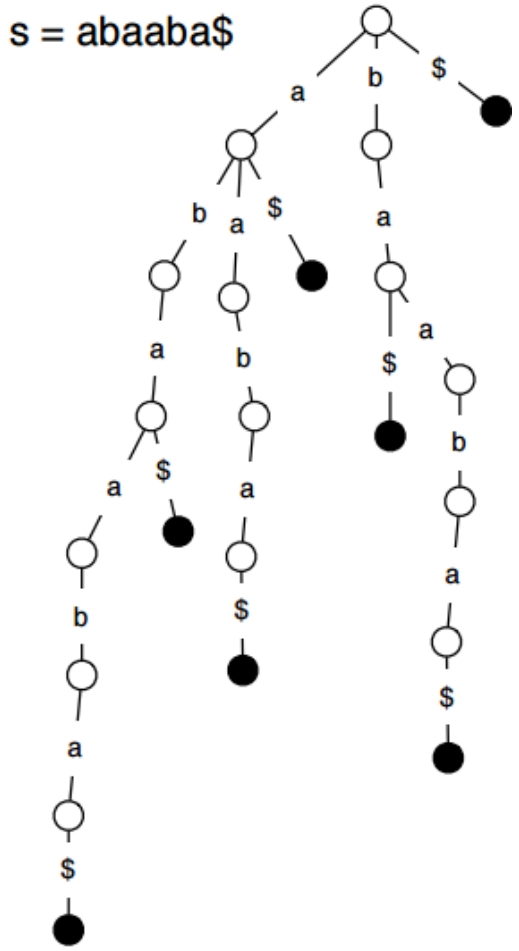
Source: © Matt Turck (@mattturck) and ShivonZilis (@shivonz)

Traditional Text Processing

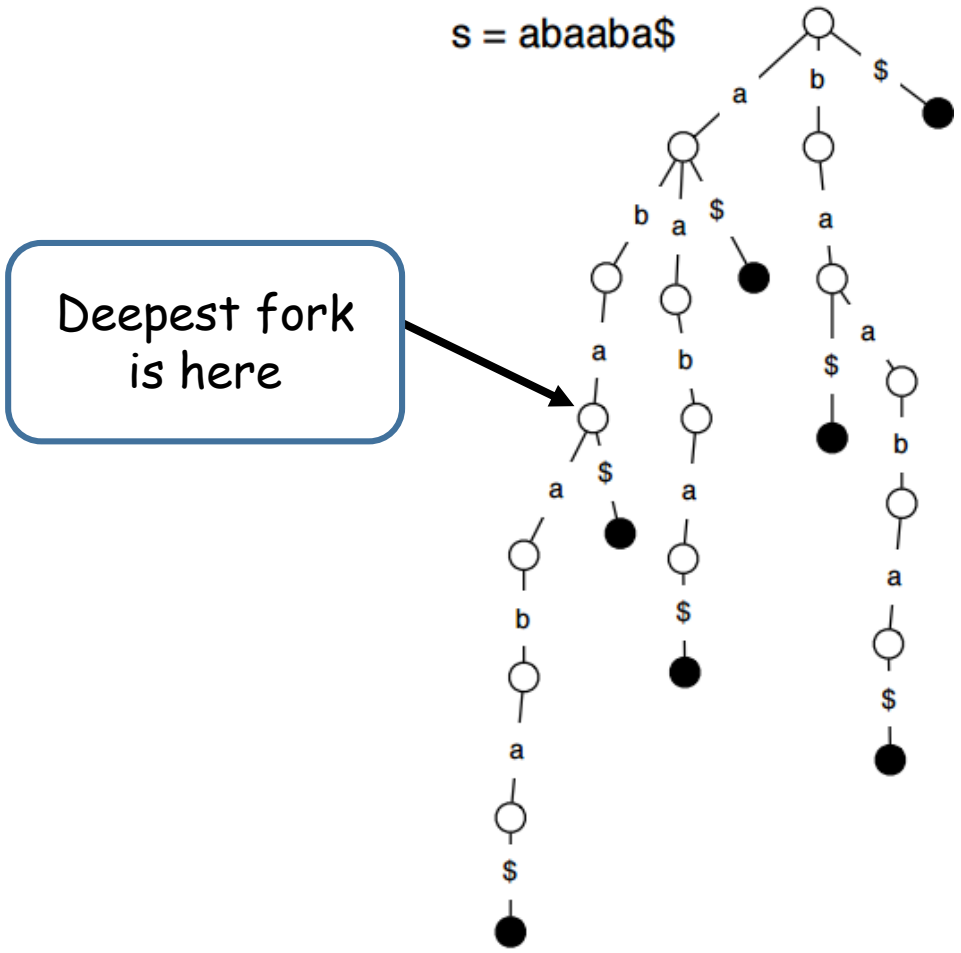
Flex your brain!

- How can you find if a given string S is a substring of another string T?
- How can you find the number of times S occurs in T?
- Is S a suffix of T?
- Find the longest repeating substring of T.
- Given two strings X and Y, find the longest common substring of X and Y.

Suffix Tree



Longest Repeating Pattern



Flex your brain!

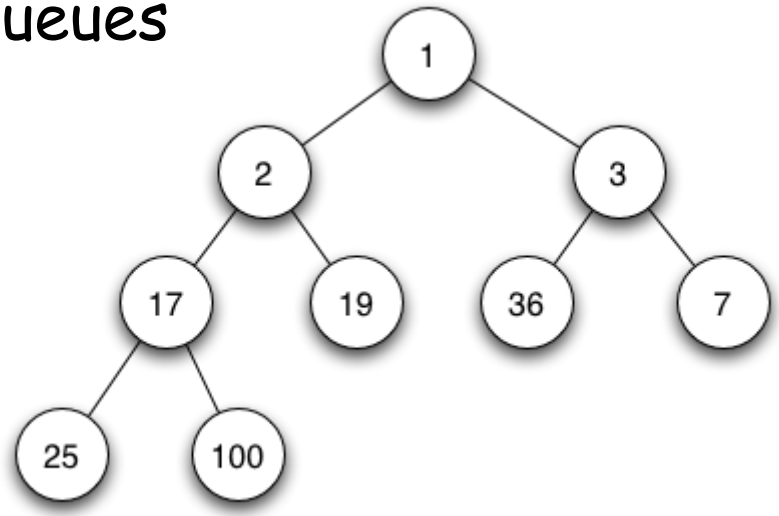
- Draw suffix trees for
  - banana
  - ssnsace
  - elephant

Flex your brain!

- How can you find if a given string S is a substring of another string T?
- How can you find the number of times S occurs in T?
- Is S a suffix of T?
- Find the longest repeating substring of T.
- Given two strings X and Y, find the longest common substring of X and Y.

Heaps

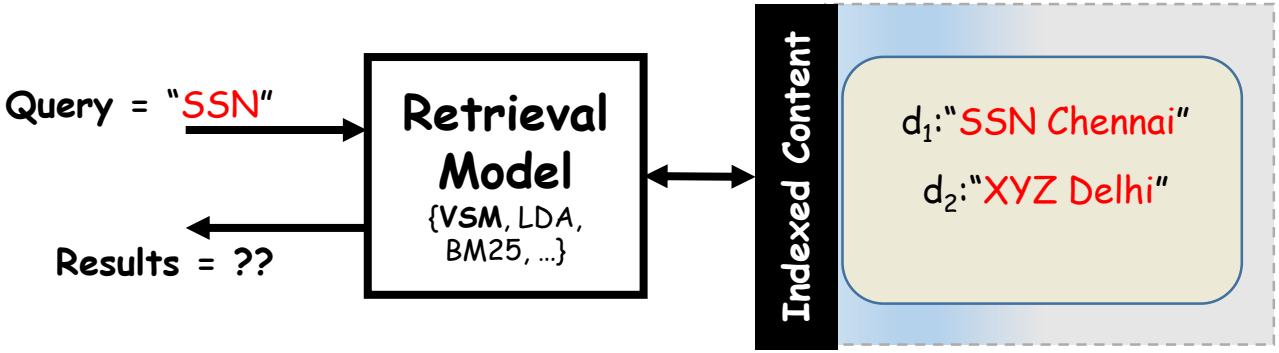
- Applications
  - kth smallest (or largest) element in an array
  - Sort an array
  - Construct priority queues
  - Find median



Modern Text Processing

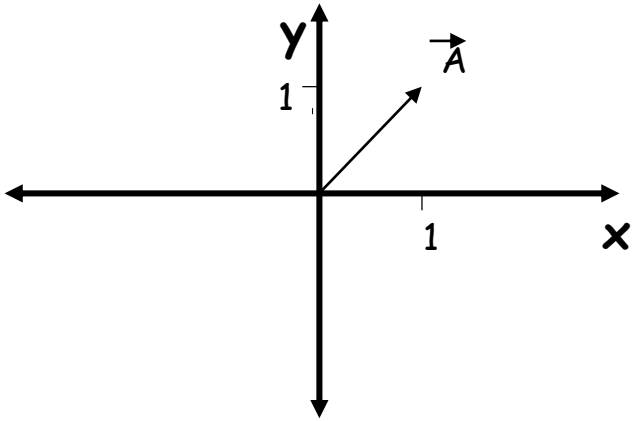
Vector Space Model

Which Document to Retrieve?



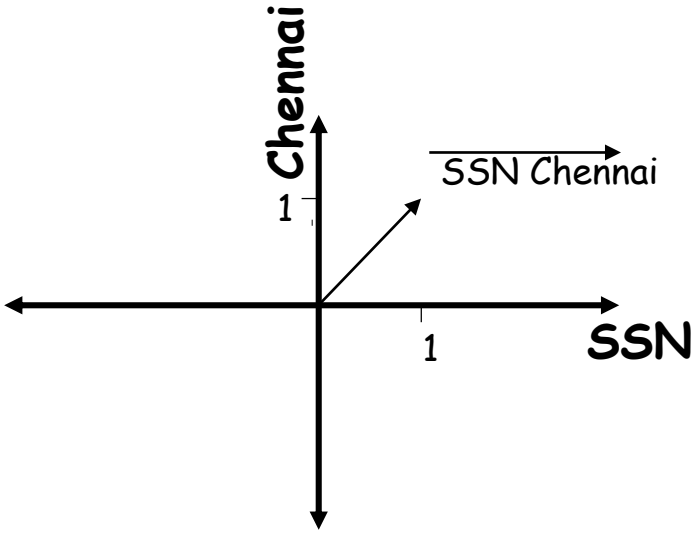
Vectors

- Geometric entity which has magnitude and direction



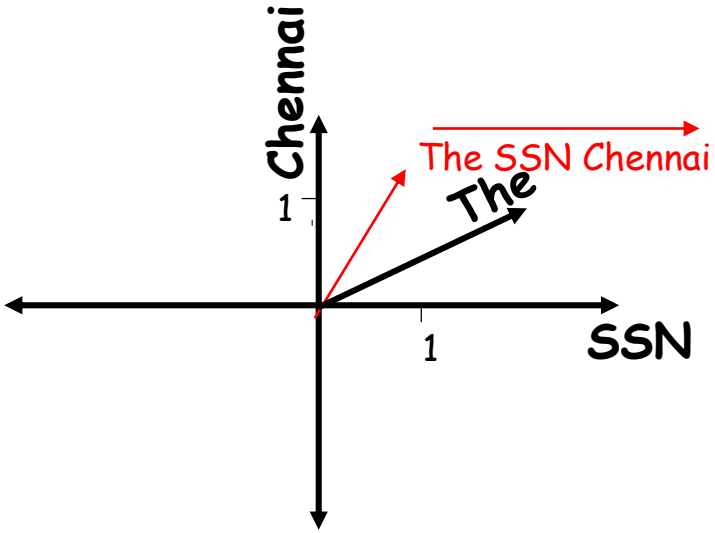
Sentences are vectors

- "SSN Chennai" as a vector



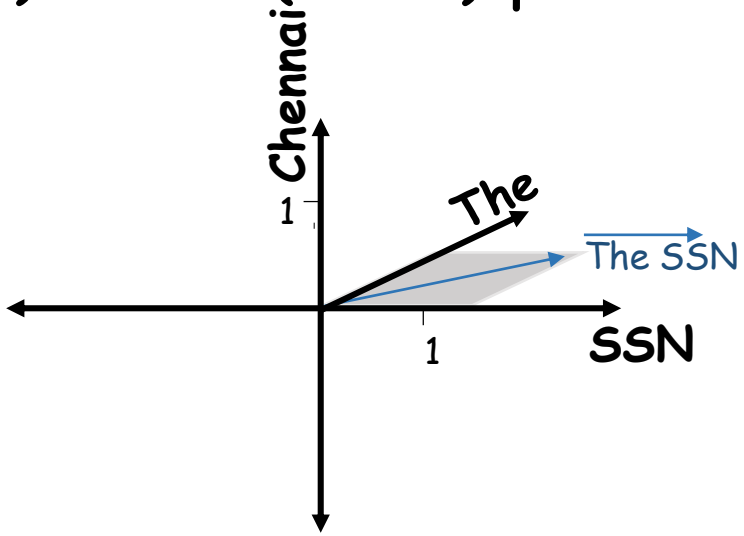
Sentences are vectors

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



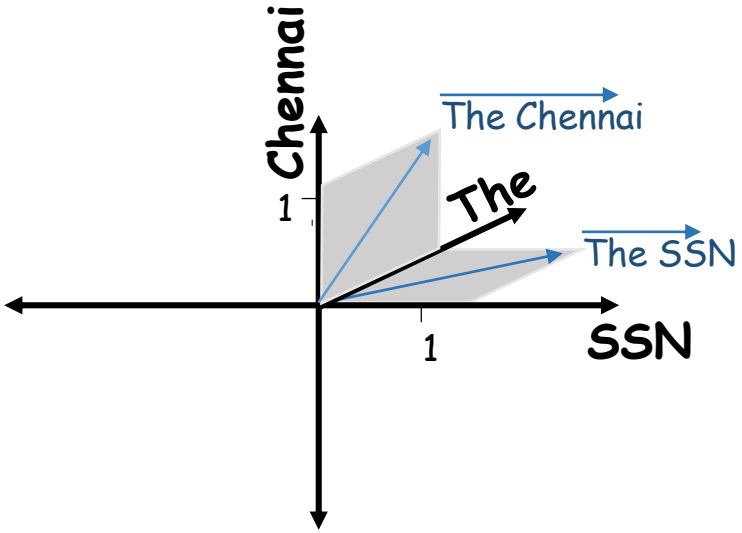
Sentences are vectors

- 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

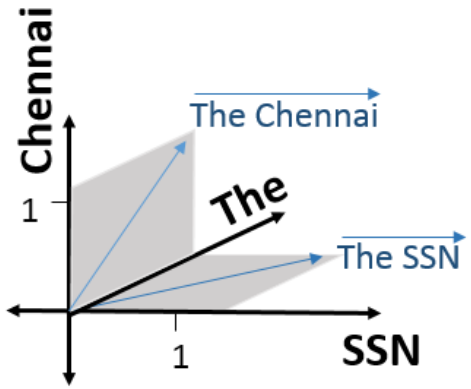


Angle between two vectors

- 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,

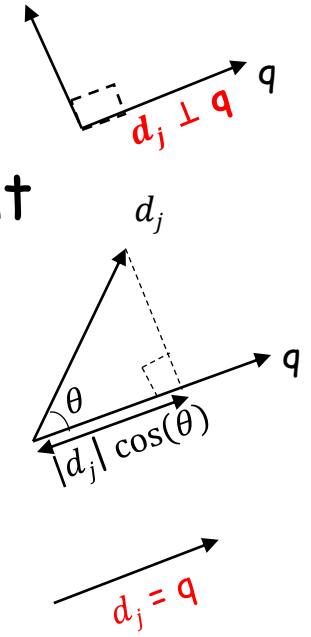
$\overrightarrow{\text{The Chennai}}$ = (1,0,1)

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

Converting from "0 - 90" to "0 - 1"

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

	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

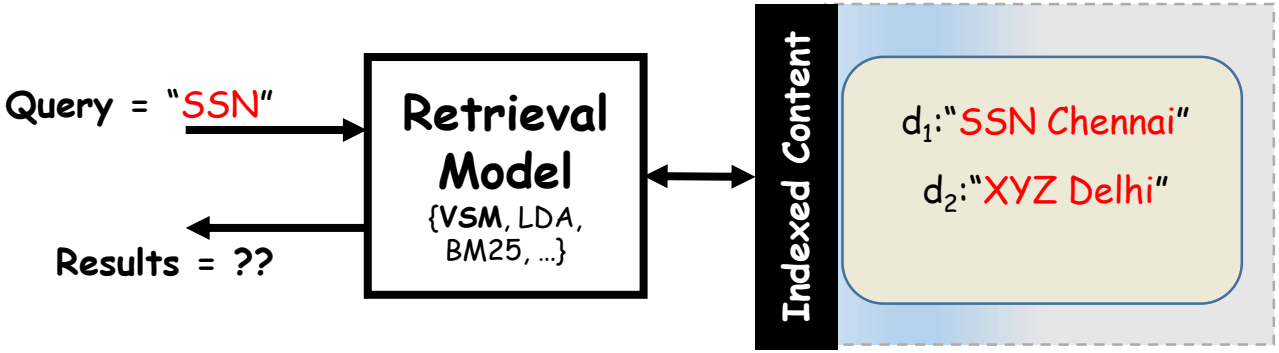


A Way to Calculate cosθ

- $\cos(\theta) = \frac{x.y}{||x||\,||y||}$
- Here,
  - x.y is the "dot product" of x and y vectors.
- So, similarity between "The SSN" and "SSN Chennai"

  = $\frac{1.0 + 1.1 + 0.1}{\sqrt{1^2+1^2+0^2}\sqrt{0^2+1^2+1^2}} = \frac{1}{\sqrt{2}\sqrt{2}}$ = 0.5

Which Document to Retrieve?



Example

Let query q = "SSN".

Let document, d_1 = "SSN Chennai" and d_2 = "XYZ Delhi".

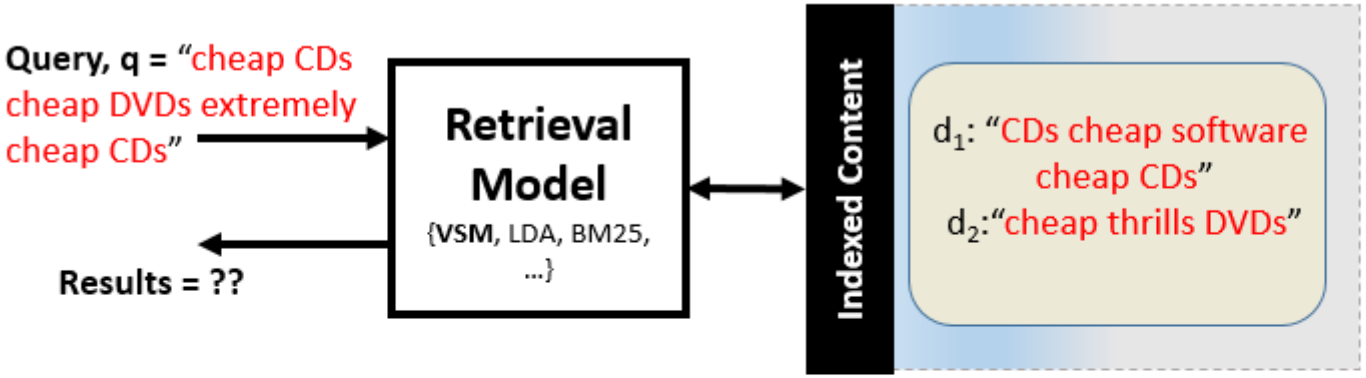
	SSN	Chennai	XYZ	Delhi
q	1	0	0	0
d_1	1	1	0	0
d_2	0	0	1	1

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

similarity(d_1, q) = $\frac{d_1.q}{||d_1||\,||q||} = \frac{1.1+1.0+0.0+0.0}{\sqrt{1^2+1^2}\sqrt{1^2}} = \frac{1}{\sqrt{2}} = 0.71$

similarity(d_2, q) = $\frac{d_2.q}{||d_2||\,||q||} = \frac{1.0+0.0+0.1+0.1}{\sqrt{1^2+1^2}\sqrt{1^2}}$ = 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

$sim(q,d_1) = 0.86$

$sim(q,d_2) = 0.59$

Summary

- Traditional Text Processing
  - Data Structures – Suffix Trees, Heaps, …
- Modern Text Processing
  - Vector Space Model
- Remember
  - Data processing goes beyond common sense… we need techniques and tools.
  - Products are good to learn. But, principles are even more important. Don't ignore them.

Thank You