Overview of Text Mining

Credit: some slides are adopted from Prof. Ray Wang's course from UNC Chapel Hill

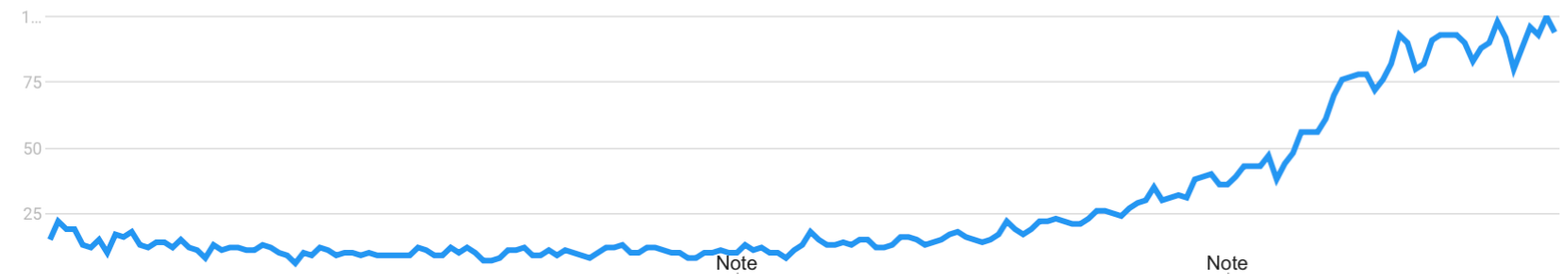
Google Trends | Explore

machine learning
Search term

+ Compare

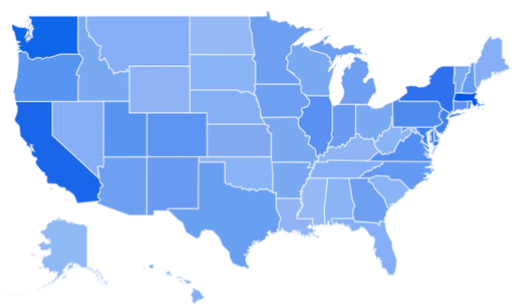
United States | 2004 - present | All categories | Web Search

Interest over time



Interest by subregion

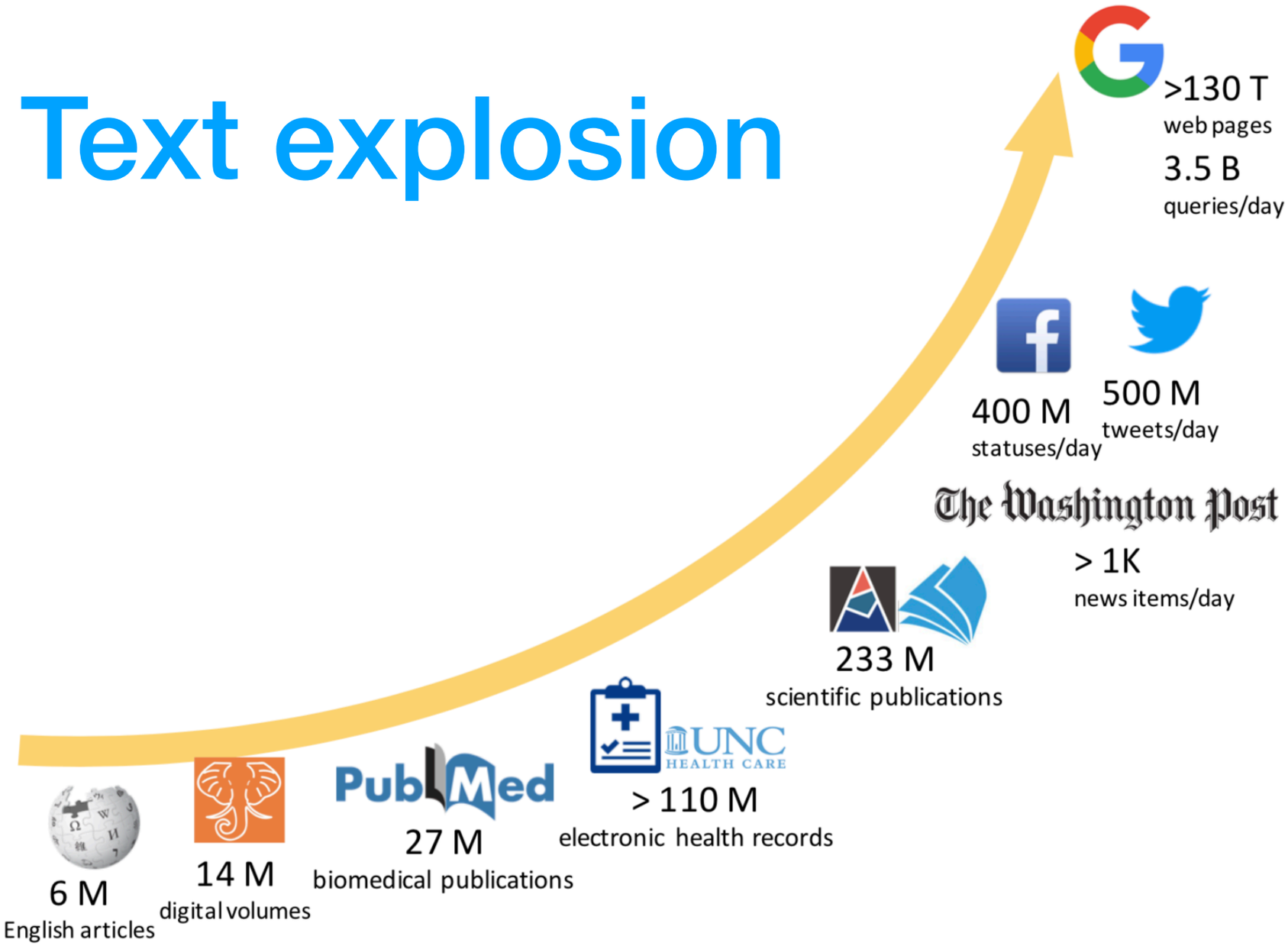
Subregion



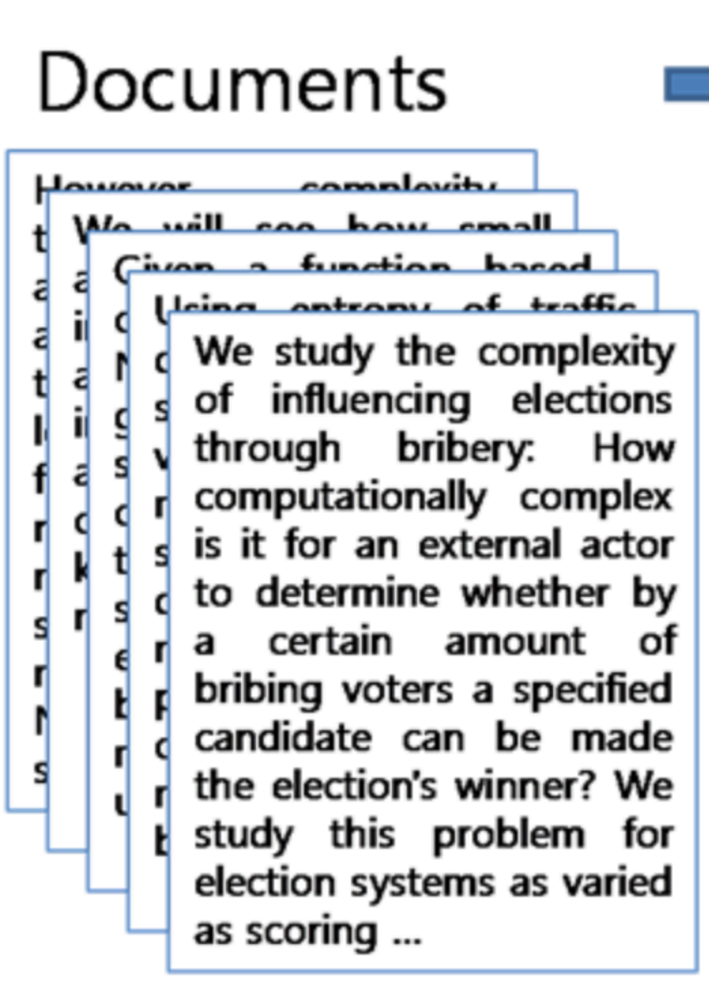
1	Massachusetts	100
2	Washington	88
3	California	86
4	New York	75
5	Maryland	66

Showing 1-5 of 51 subregions

Text explosion



The anatomy of a matrix for text mining.



Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

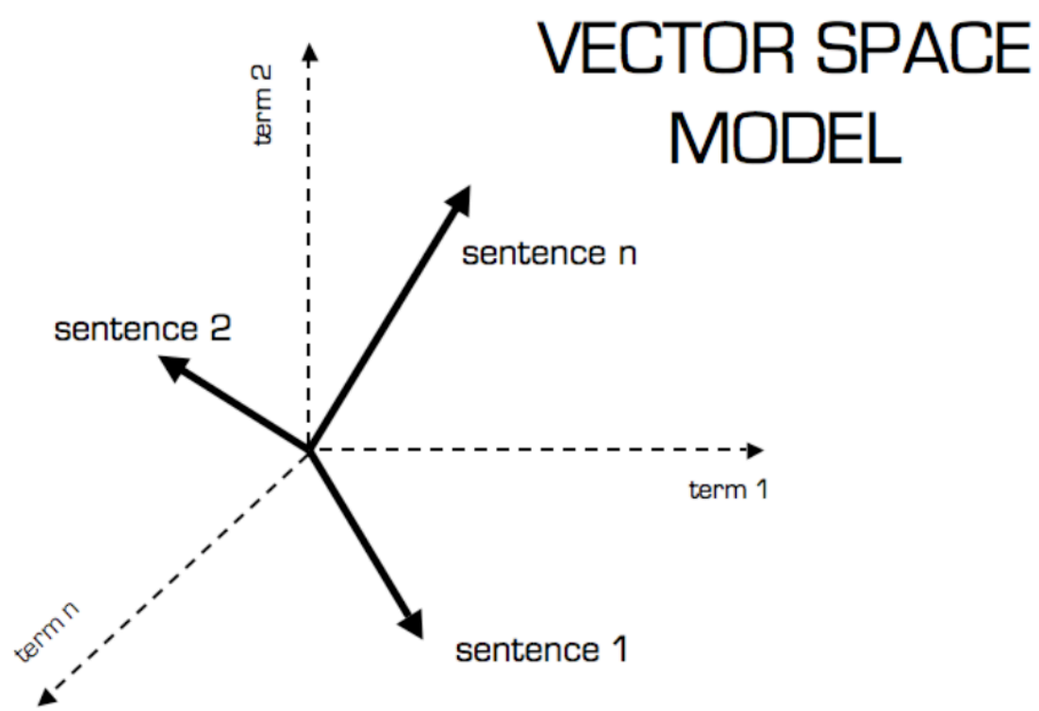
Term-document matrix

The Vector Space View

- Data object: a user's ratings of movies, a piece of text, a patient's test results
- Attribute: a numerical property of the object
- A data object is represented as n-dimensional vector of attributes
  - Each attribute corresponds to one dimension of vector space
  - Numerical value on each dimension

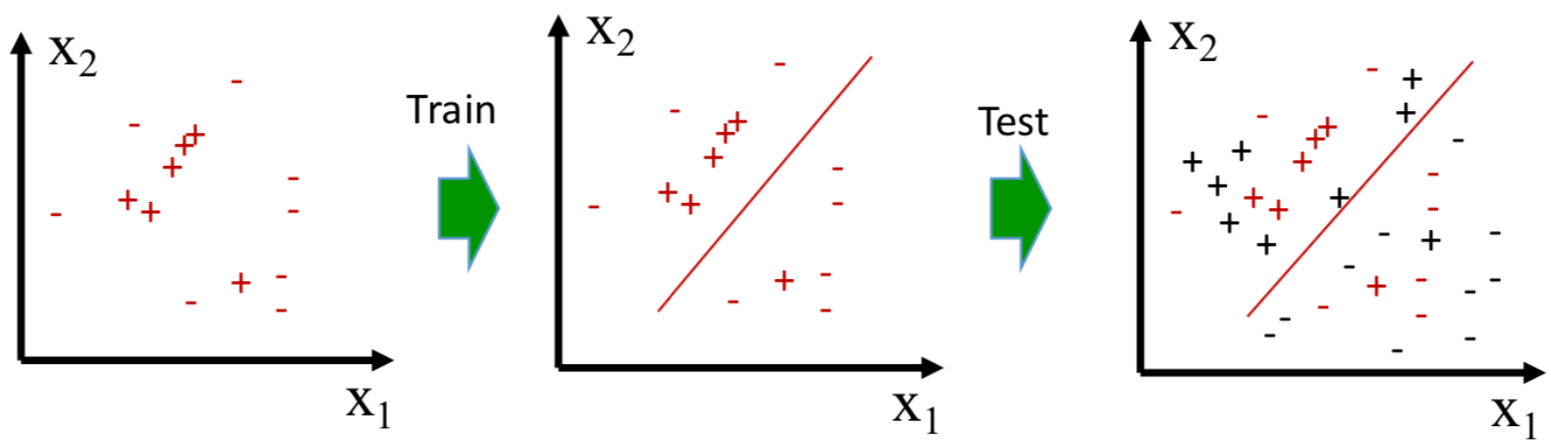
$$\vec{X} = [x_1, x_2, \cdots, x_n]$$

- x_i is the numerical value of X at the i^{th} dimension (attribute)
- Therefore, the entire dataset can be represented as a matrix (a collection of vectors)



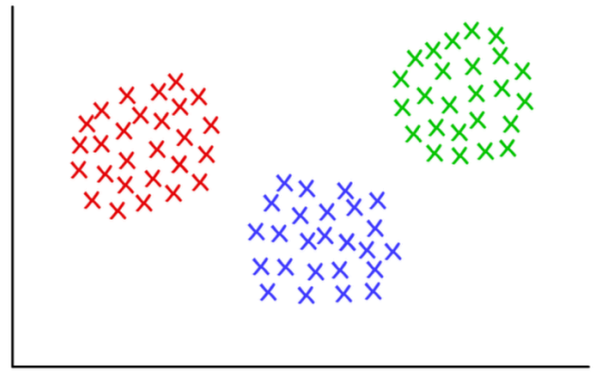
Classification

- Data: a set/vector of attributes (features)
- Classifier: attributes (X) → categorical class label (Y)
- Key technique: supervised learning
  - availability of a training set of examples with class labels
  - Training: learn classifier from training data
  - Test: predict the class label of unseen data



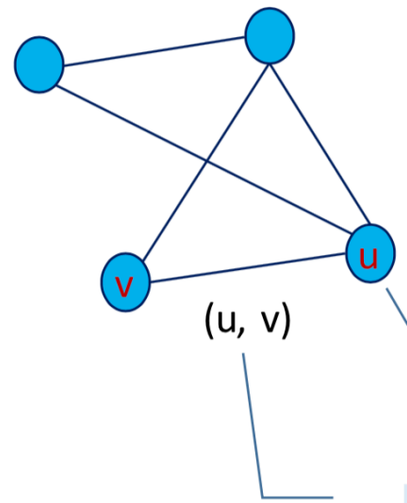
Clustering

- Similar to classification, but no predefined classes (or training data)
- Attributes (X) → assignment to one of multiple clusters/groups
- Key technique: unsupervised learning
- Usually requires a metric of similarity/distance between data points
- Intuition:
  - Similarity within cluster
  - Dissimilarity across clusters



What is a Network?

- A network is a collection of points joined by lines



Mathematically (Topologically):
"Network" ≡ "Graph"

G = (V, E)
V is a set of vertices, u, v in V are vertices
E is a set of edges, (u, v) in E is an edge

Vertex, node, site, actor, …

Edge, arc, link, bond, tie, relation …

One thing these last two methods have in common…

Distance/Similarity Calculation

- The relevance of two vectors can be calculated based on distance/similarity measures
- s: $\mathbf{x}, \mathbf{y} \rightarrow [0, 1]$

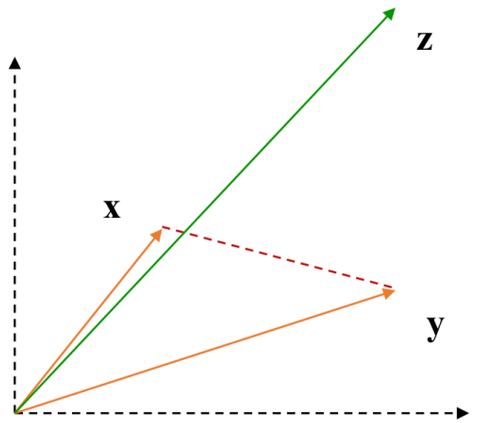
$$\mathbf{x} = \langle x_1, x_2, \cdots, x_n \rangle$$
$$\mathbf{y} = \langle y_1, y_2, \cdots, y_n \rangle$$

- s($\mathbf{x}$, $\mathbf{y}$) = ?
  - The more dimensions in common, the larger the similarity
  - The closer values in each dimension, the larger the similarity

Similarity Measures

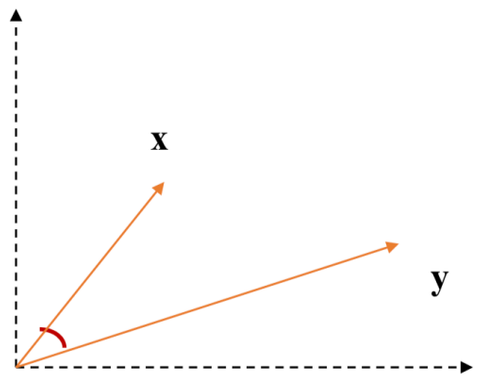
- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



- Cosine similarity

$$cos(\mathbf{x}, \mathbf{y}) = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{\sqrt{x_1^2 + \cdots + x_n^2} \cdot \sqrt{y_1^2 + \cdots + y_n^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$



We will get into the details of all these methods over the next few weeks. Work hard, and have fun!