WEEK 3
VECTOR SPACE MODELS.

- Why learn vector space models? → to identify similarity:

Where are you heading?
Where are you from?
different meaning

What is your age?
How old are you?
same meaning

Applications:

> capture the dependencies between words
> information extraction to answer the questions
> machine translation
> chatbots

→ Represent words and documents as vectors
→ Representation that captures relative meaning

• Word by word & word by doc.

Vector representation designs (co-occurrence):

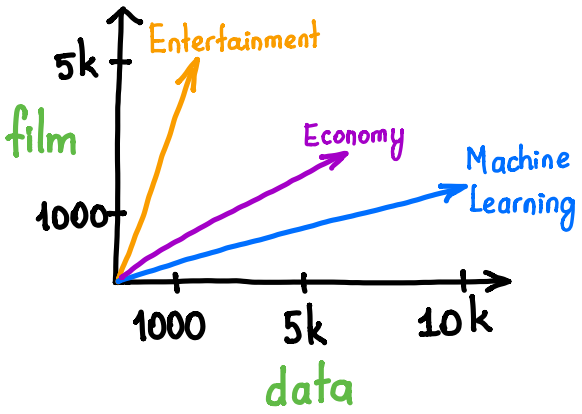
1 → word by word design: # of times they occur together within a distance k

"I like simple data"
"I prefer simple raw data"

	simple	raw	like	I
data	2	1	1	0

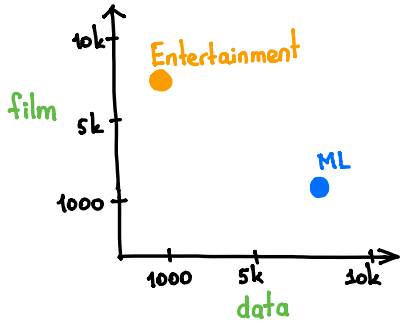
2 → word by document design: # of times a word occurs within a certain category.

	Entertainment	Economy	Machine Learning.
data	500	6620	9320
film	7000	4000	1000



Measures of "similarity":
angle distance

• Euclidean distance



A : [500, 7000]

B: [9320, 1000]

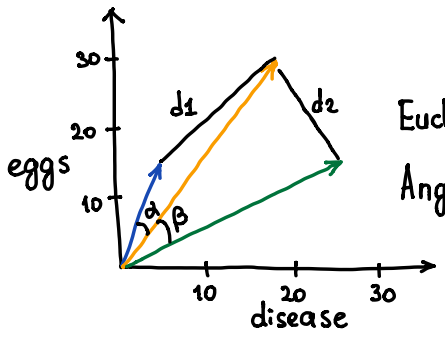
$d(B,A) = \sqrt{8820^2 + (-6000)^2} = 10667$

$d(\vec{v}, \vec{w}) = \sqrt{\sum_{i=1}^{n} (v_i - w_i)^2}$ → norm of $(\vec{v} - \vec{w})$

• Cosine Similarity

Euclidean distance and cosine similarity,

→ the cosine of angle.



Euclidean distance: $d_2 < d_1$

Angles comparison: $\beta > \alpha$

Use cosine similarity (over euclidean distance) when corpora are in different sizes.

→ Vector norm $\|\vec{v}\| = \sqrt{\sum_{i=1}^{n} v_i^2}$

→ Dot product $\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} v_i \cdot w_i$

$\vec{v} \cdot \vec{w} = \|\vec{v}\| \, \|\vec{w}\| \cos(\beta)$

$\cos(\beta) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\| \, \|\vec{w}\|}$

- When $\beta = 90°$ → $\cos(\beta) = 0$ → dissimilar / orthogonal
- When $\beta = 0°$ → $\cos(\beta) = 1$ → similar / identical

• Manipulating words in vector spaces.

USA → Washington DC

Russia → ? (Russia + (Washington DC − USA))

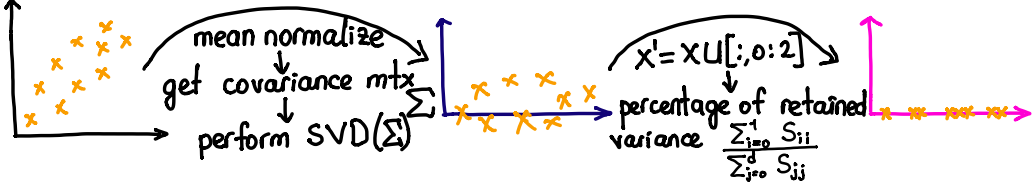
→ use known relationships to make predictions.

• PCA: Principal Component Analysis

- PCA is an algorithm used for dimensionality reduction
- Original space → Uncorrelated features → Dimensionality reduction
- Visualization to see words relationships in the vector space

• Eigenvectors & eigenvalues

- Eigenvector: uncorrelated features for the data (U)
- Eigenvalue: the amount of information retained by each feature (S)



- Eigenvectors give the direction of uncorrelated features.
- Eigenvalues are the variance of the new features.
- Dot product gives the projection on uncorrelated features.