Vector Space Models and Vector Similarity

Some slides and screenshots taken from https://www.coursera.org/learn/classification-vector-spaces-in-nlp

Why use vector space models

- How old are you ?
- What is your age ?
- => different words, same meaning…

- We want to try and capture the meaning of sentences/words, while not being too sensitive to the forms of the words used, but to their meaning !
- For QA, information extraction…

Why use vector space models

- Capture depencies between words :

- I like to eat apples.
- I like to eat pears.
- Using the context of apples and pears, we can deduce that these are both food ! We see they are « surrounded » by the same words and occur in similar positions.

- Going too fast is dangerous, but going slow is not dangerous…
- Given the context, we can deduce fast and slow are antonyms !

Vectors

- Vectors are used as a way to represent the information found in a word or a sentence (/document).

- They are an effective way of transforming words and their relative meaning into mathematical objects to feed to an algoithm.

Fundamental Concept

« You shall know a word by the company it keeps » (Firth, 1957)

Indeed, the vector is built by observing the context around the word and this captures the word’s relative meaning !

How do we construct these vectors ?

- Using a coocurence matrix
- To extract vector representations of
  - A word
  - A Document
  - Depending on the application

- These are called *designs*

Word by Word design

- The co-occurrence of 2 different words is defined by the *# of times they occur together within a certain distance/window k*

I like simple data

I prefer simple raw data

k=2

	simple	raw	like	I
data	2	1	1	0

Practice

- « In general, I love music. But I love pop music more than any other musical genre. To me, music is my greatest love. »

- What is the value for the co-occurrence of “love” and “ music”, if k=2?

Word by Document Design

- Number of times a word *occurs within a specific category*
- Imagine our corpus is divided into three topics :

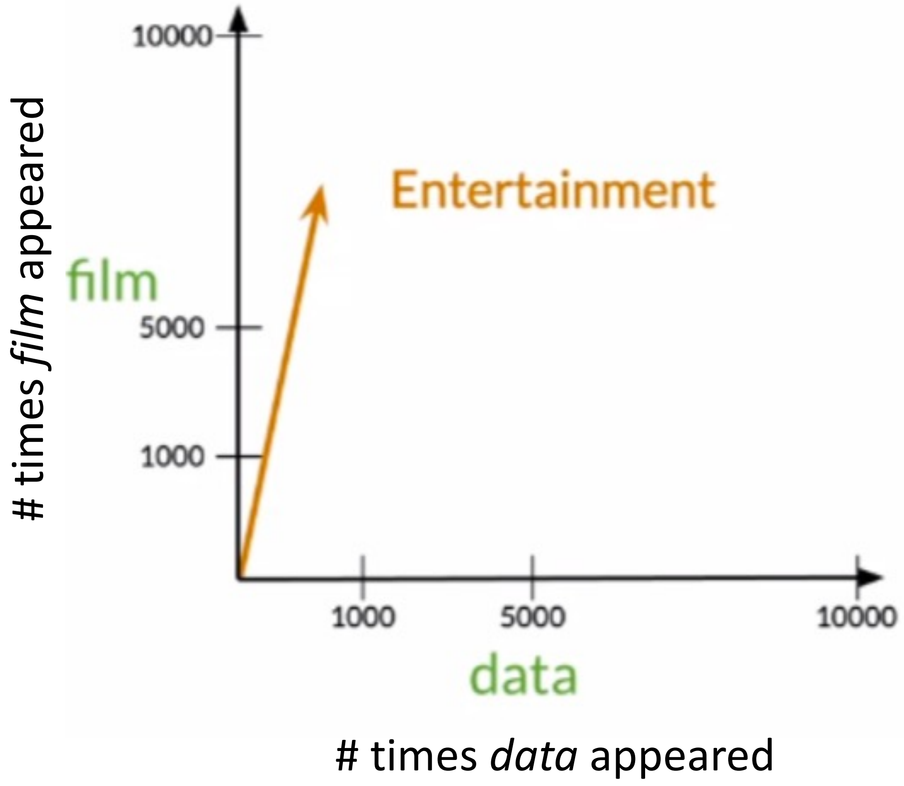
	Entertainment	Economy	Machine Learning
data	500	6620	9320
film	7000	4000	1000

Vector Space

- Given our matrix in the previous slide, we could represent the words *data* and *film* using the rows of our matrix, which would give us two 3-D vectors.

- To make things more visual let’s take the vectors for the **topics** (2-D vectors), using the columns :

Vector space

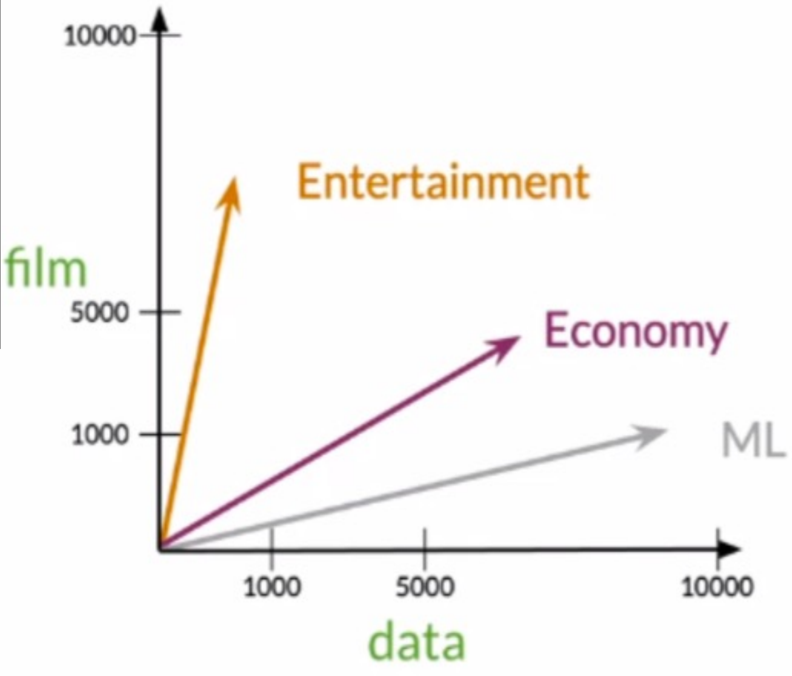


	Entertainment	Economy	ML
data	500	6620	9320
film	7000	4000	1000

Vector Spaces

- We can determine relationships between types of documents
- We can see that the documents about economy and ML are more similar than those about entertainement…

Vector Space



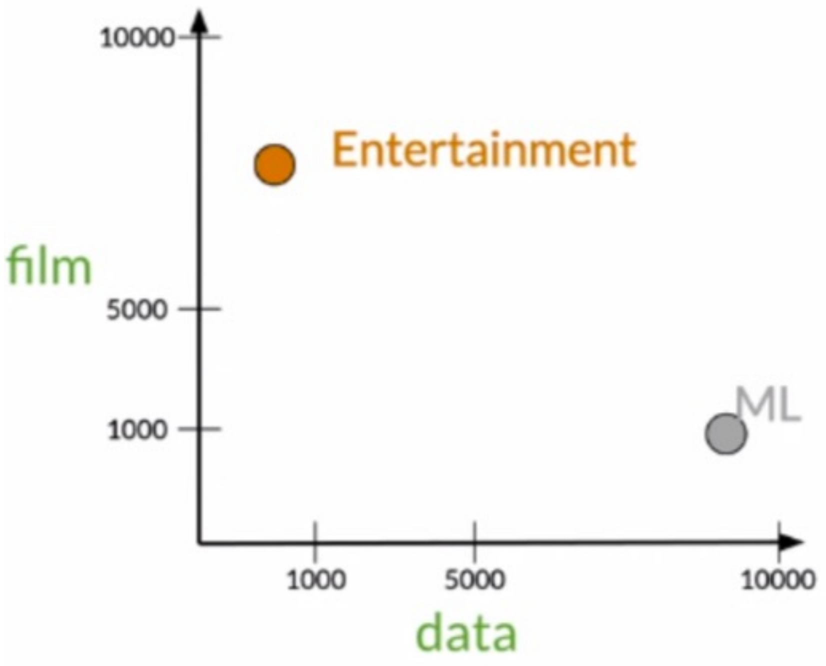
	Entertainment	Economy	ML
data	500	6620	9320
film	7000	4000	1000

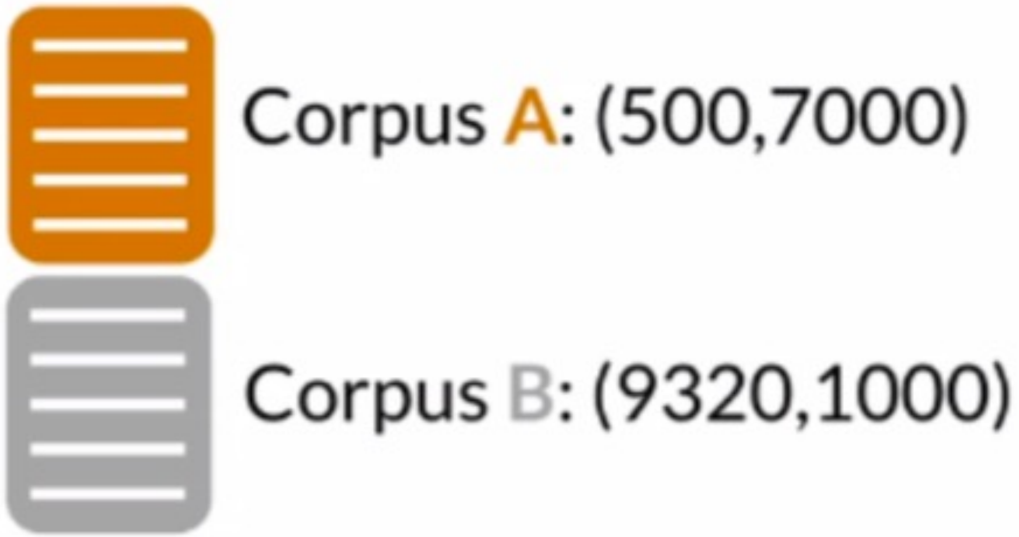
How do we measure the degree of similarity ?

- Distance between vectors

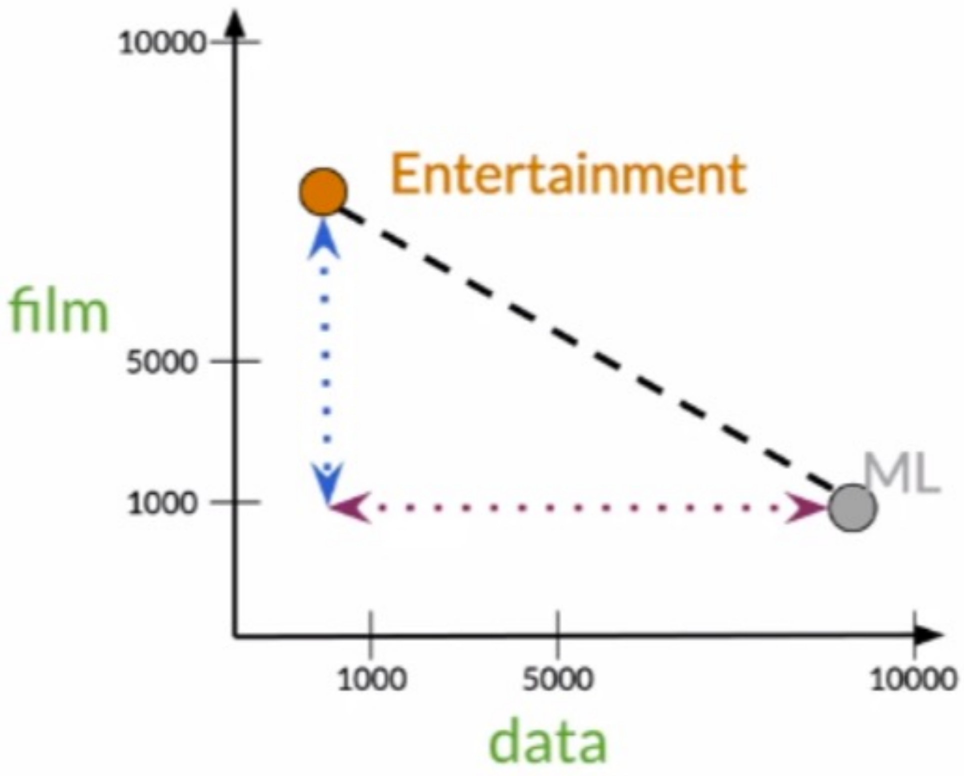
- Angle between vectors

Euclidian distance between 2 points/vectors





Euclidian distance



$$d(B,A) = \sqrt{\underline{(B_1 - A_1)}^2 + \underline{(B_2 - A_2)}^2}$$

- 1st term : distance between their x coordinates
- 2nd term : distance between the y coordinates

$$c^2 = a^2 + b^2$$

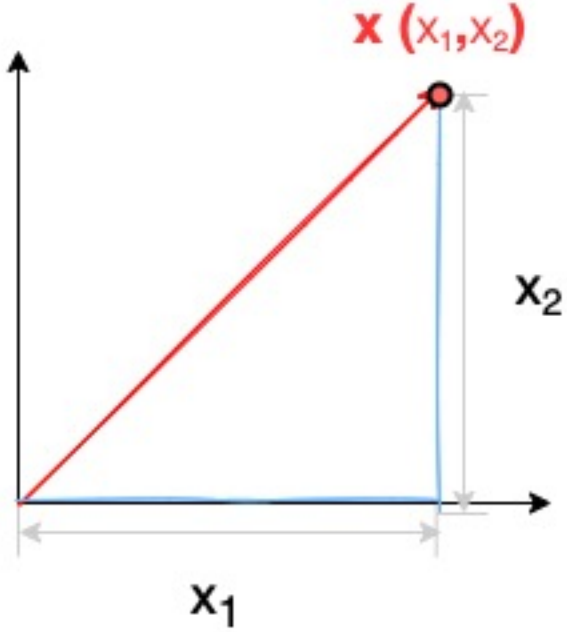
Vector Norm – Euclidian Norm

- How can we calculate the length of a vector ?

- $\| \boldsymbol{a} \| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} = \sqrt{\sum_{i=1}^{n} a_i^2}$

- Euclidian norm or L2 norm

- Measures the shortest distance from the origin



Distance / Vector norm

- Finding the distance between 2 vectors comes down to

- *calculating the length of the vector that allows you to reach the 2nd vector from the tip of the 1st*

- This is the vector $\mathbf{c} = \boldsymbol{b} - \boldsymbol{a}$

$$\| \boldsymbol{c} \| = \sqrt{c_1^2 + c_2^2 + \cdots + c_n^2}$$

$$= \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + \cdots + (b_n - a_n)^2}$$

Euclidian distance for an n-dimensional matrix

- We can now generalize to any number of dimensions

	data	boba ($\vec{w}$)	ice-cream ($\vec{v}$)
AI	6	0	1
drinks	0	4	6
food	0	6	8

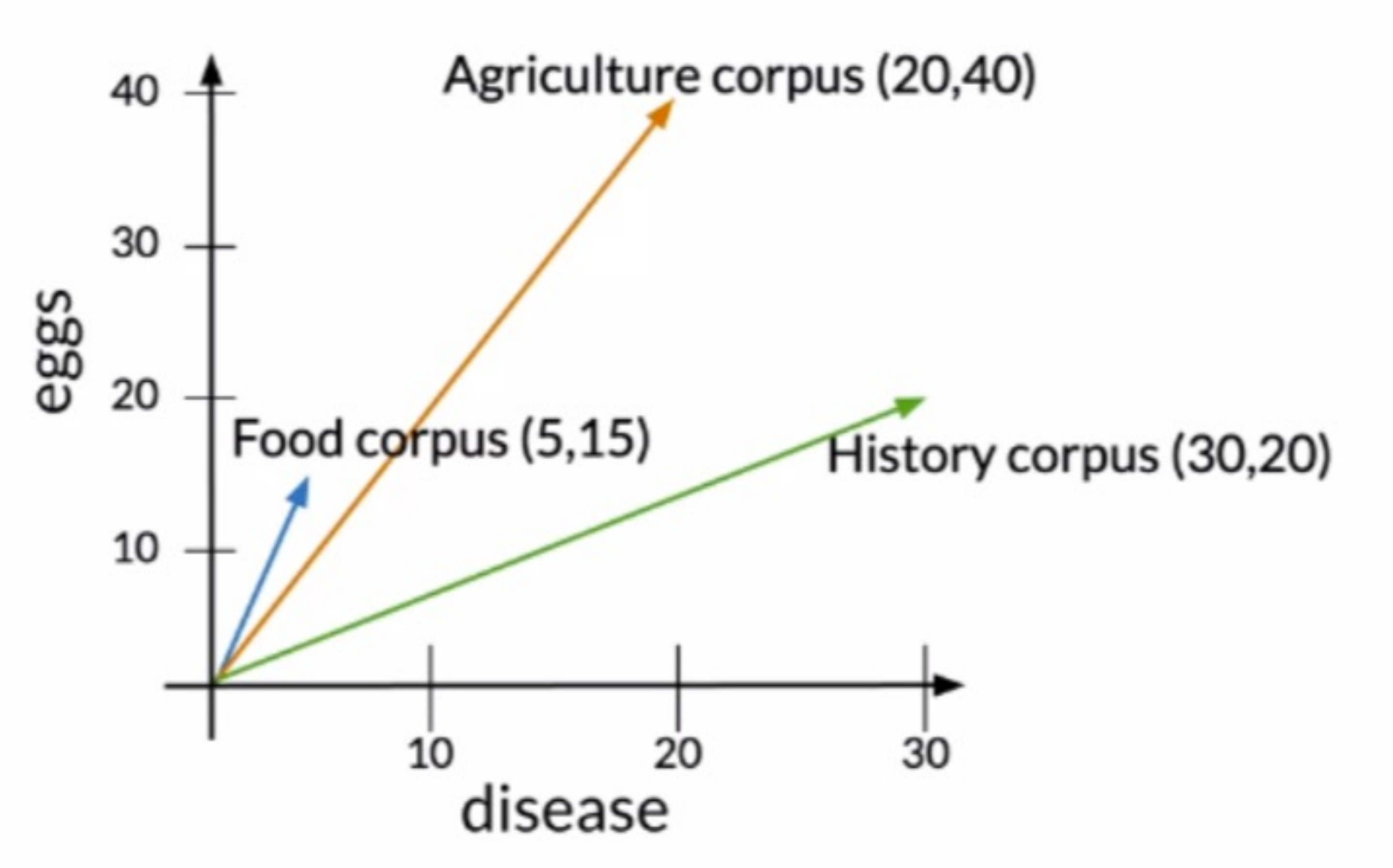
$$= \sqrt{(1-0)^2 + (6-4)^2 + (8-6)^2}$$

$$= \sqrt{1+4+4} = \sqrt{9} = 3$$

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

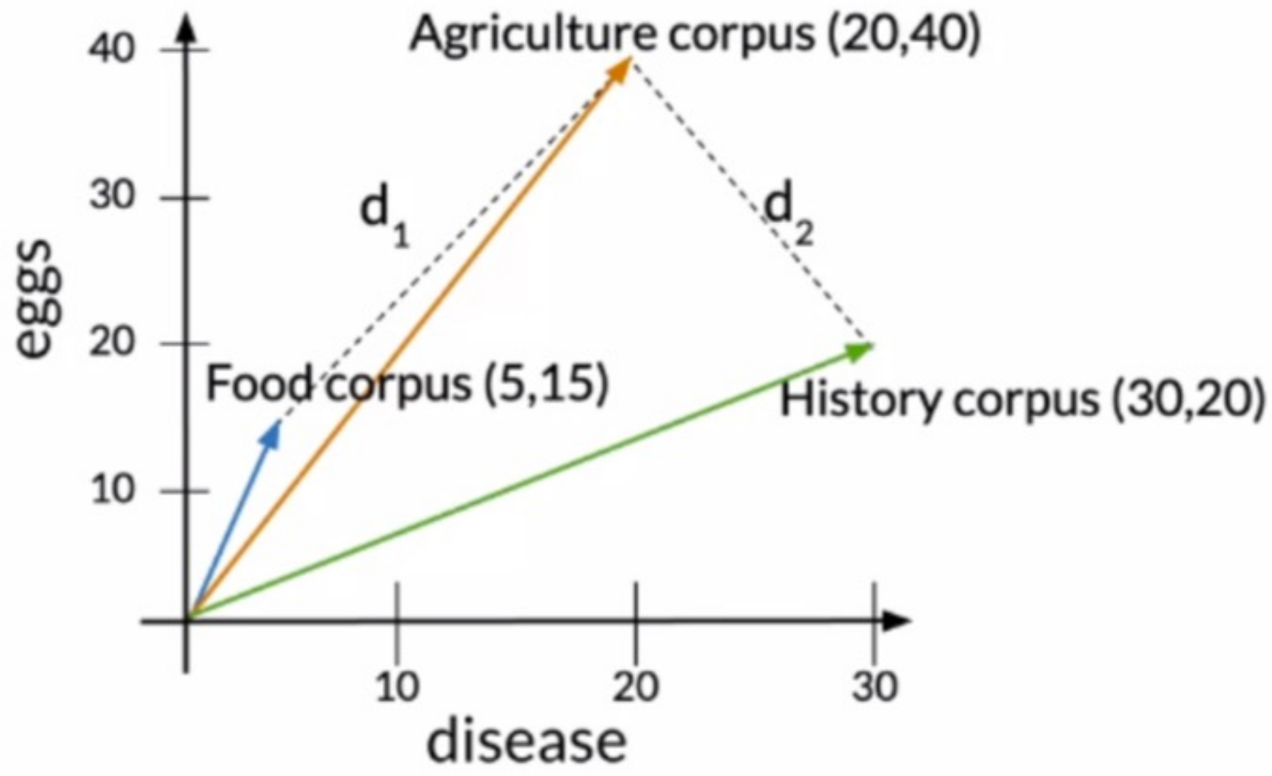
Cosine similarity

- Euclidian distance vs Cosine Similarity



Euclidian distance limitations

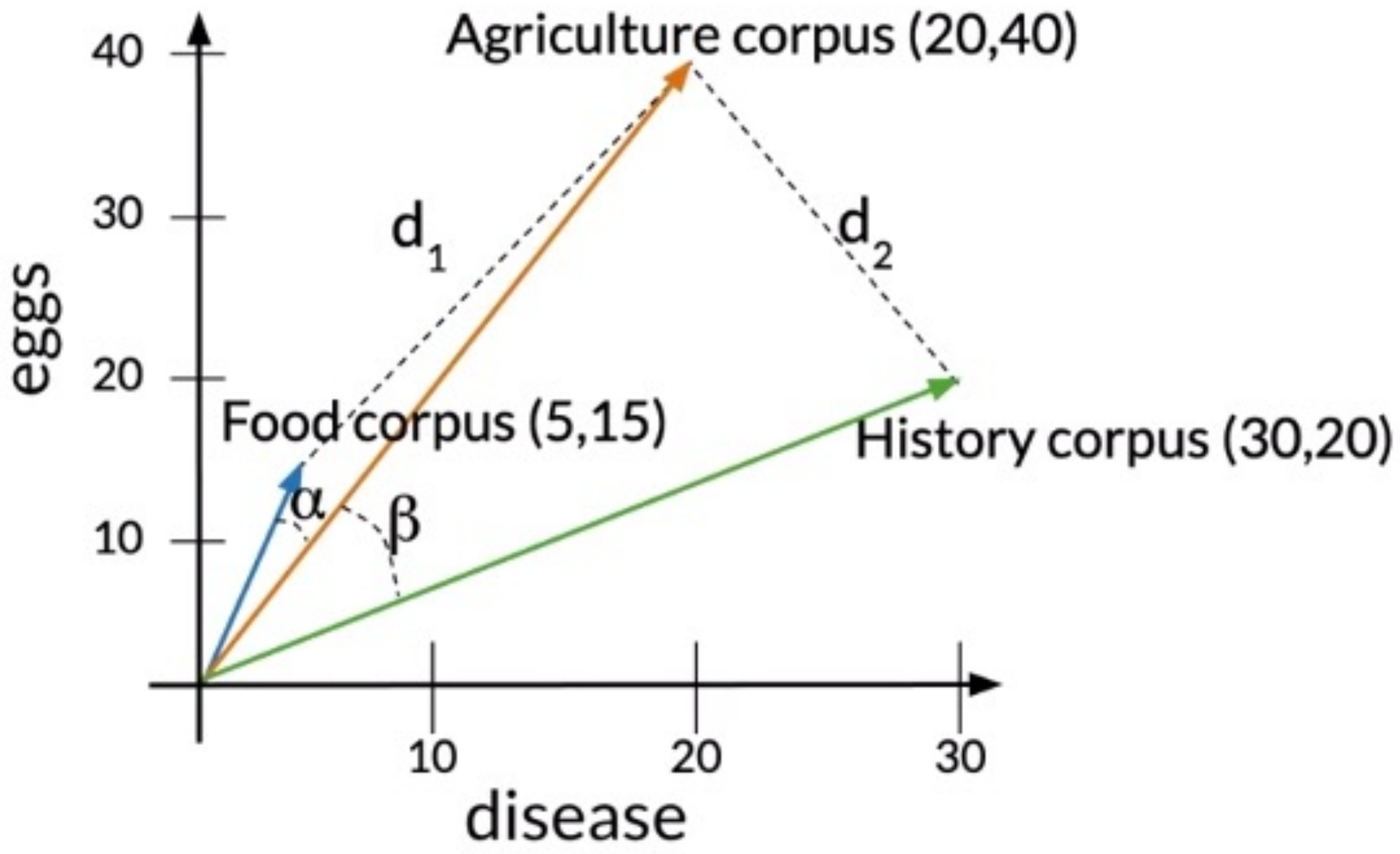
- In this case, measuring the euclidian distance suggests the agriculture and food corpora have less in common than the agriculture and history corpora…



Euclidean distance: $d_2 < d_1$

Cosine Similarity

- Another method for computing similarity is to compute the **cosine of the inner angle between 2 vectors**
- See if 2 vectors are pointing in the same direction
- $\beta > \alpha$
- This metric is not biased by the magnitude of the vector representations
- So this is a more adapted metric when the corpora are of different sizes



Law of Cosines and the Dot Product

- Now you know about vector norms, another way of expressing the dot product that we haven't seen is :

$$\mathbf{a} \cdot \boldsymbol{b} = ||\boldsymbol{a}||||\boldsymbol{b}||cos\theta$$

- This comes from the law of cosines

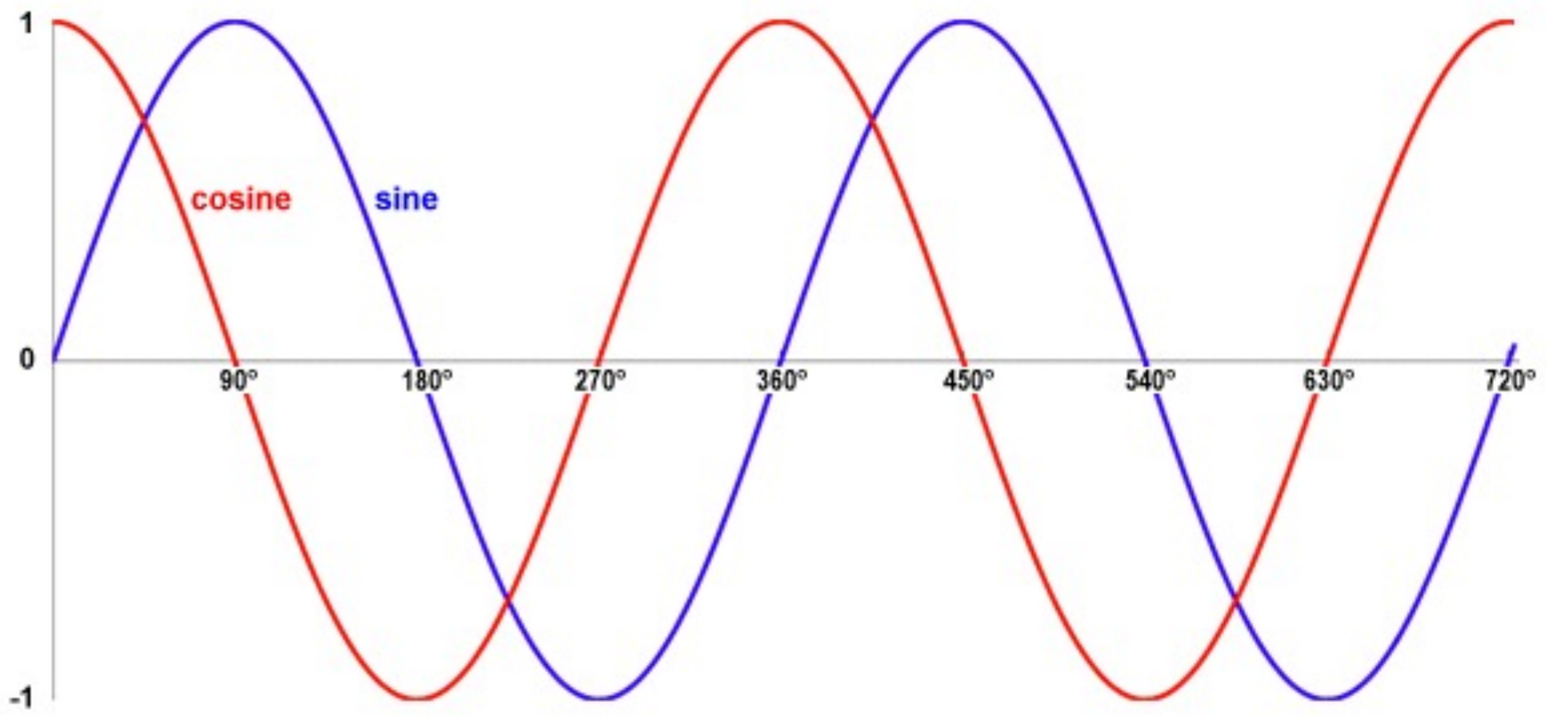
- (See here for proof)

New intuition about the dot product

- The result of the dot product is therefore impacted by
  - The magnitude of the vectors
  - their direction / the angle between them

- When $\theta < 90°$ dot product is positive
- When $\theta = 90°$ dot product = 0
- When $90° < \theta < 180°$ dot product is negative

Sine and Cosine functions



Cosine Similarity

- Remember

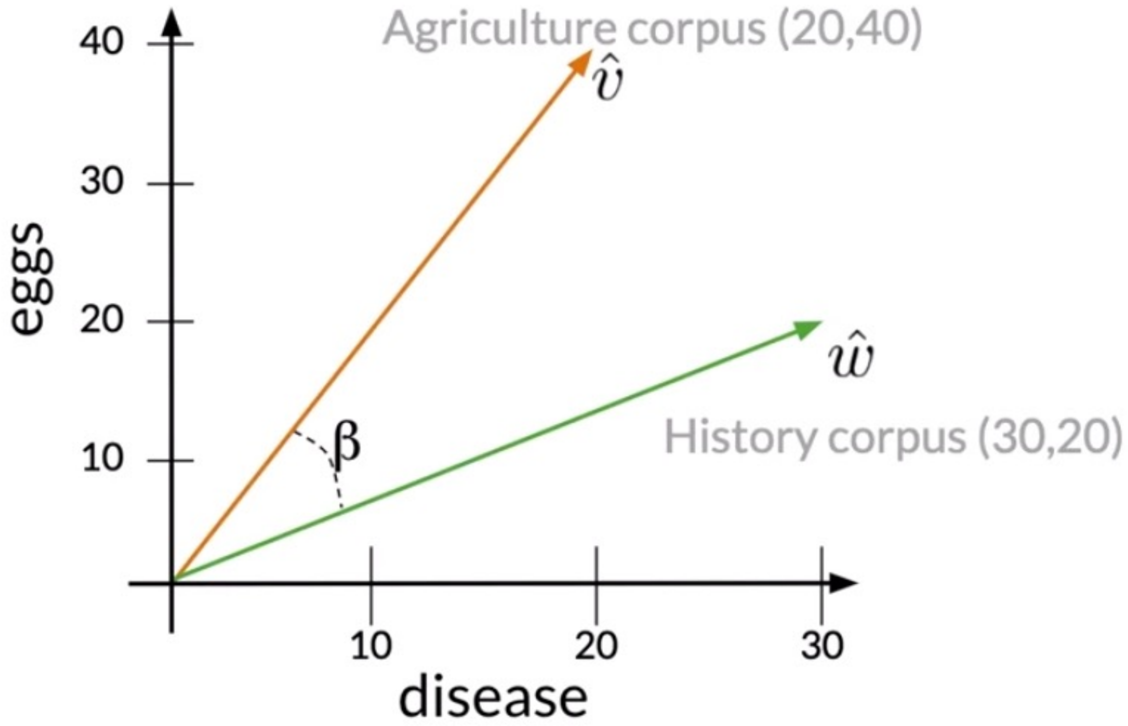
$$\mathbf{a} \cdot \boldsymbol{b} = \|\boldsymbol{a}\|\|\boldsymbol{b}\| cos\theta$$

- So

$$Similarity(\boldsymbol{a}, \boldsymbol{b}) = \cos\theta = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$$

Cosine similarity

In our example, because word counts are positive, the cosine similarity cannot be negative



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

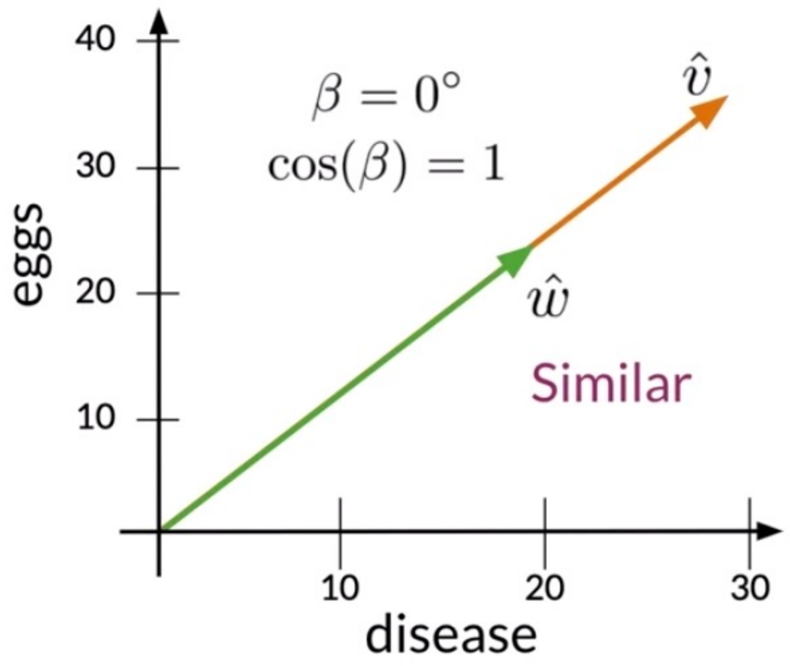
$$\boxed{\cos(\beta) = \frac{\hat{v} \cdot \hat{w}}{\|\hat{v}\| \|\hat{w}\|}}$$

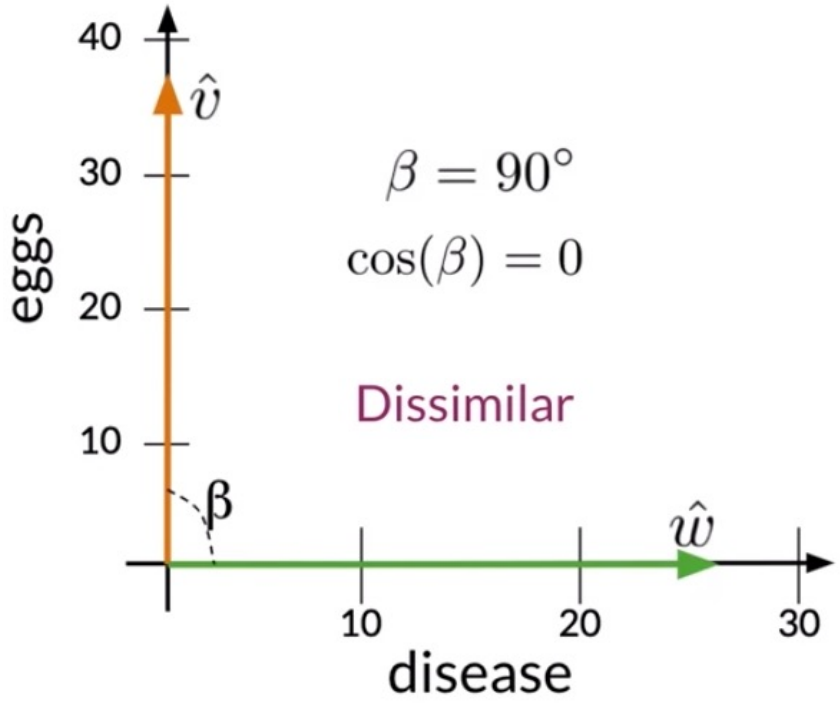
$$= \frac{(20 \times 30) + (40 \times 20)}{\sqrt{20^2 + 40^2} \times \sqrt{30^2 + 20^2}}$$

$$= 0.87$$

What does cosine similarity tell us about the similarity between 2 vectors ?

- Max angle is 90° for reasons explained previously





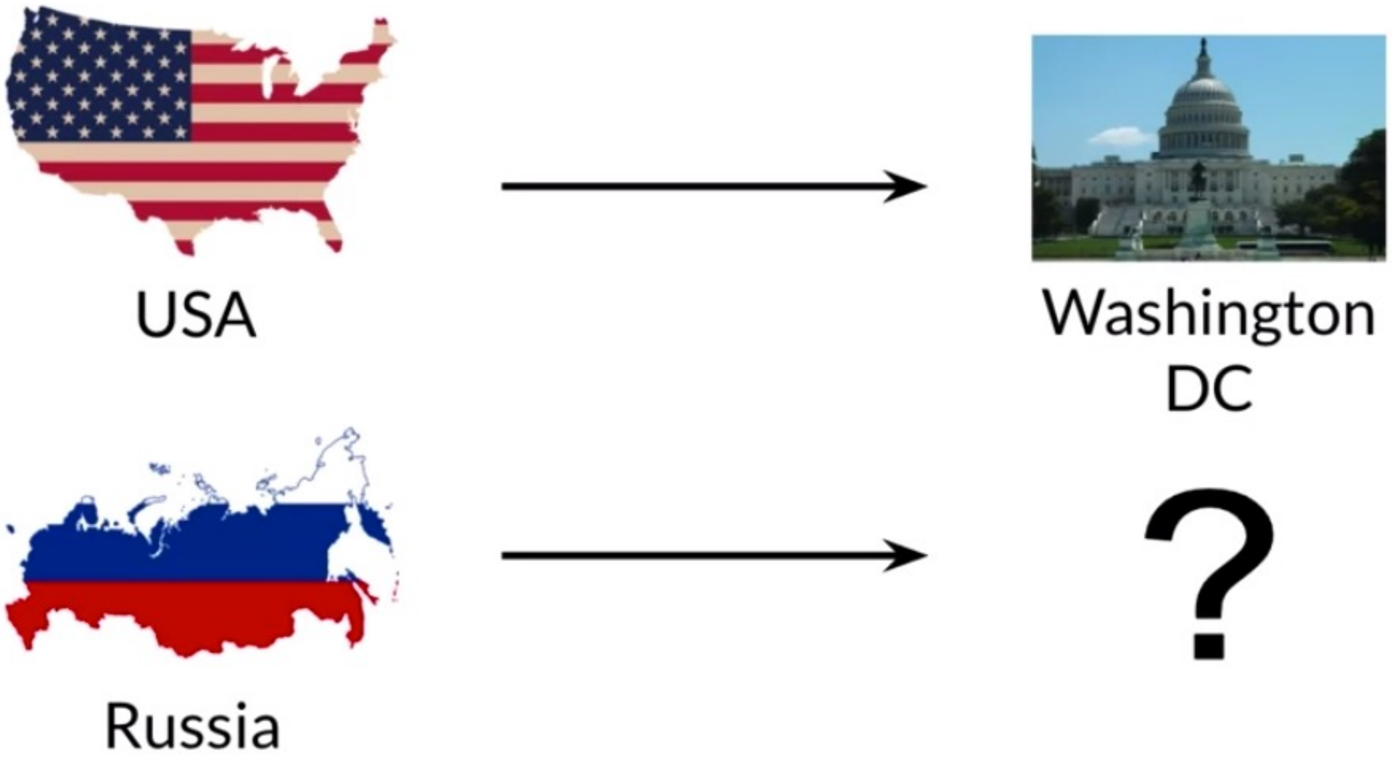
Cosine Similarity

- So cosine similarity is proportional ($\propto$) to the similarity between the directions of the vectors

- $0 < \text{simil} < 1$ for the vector space we’ve seen so far.

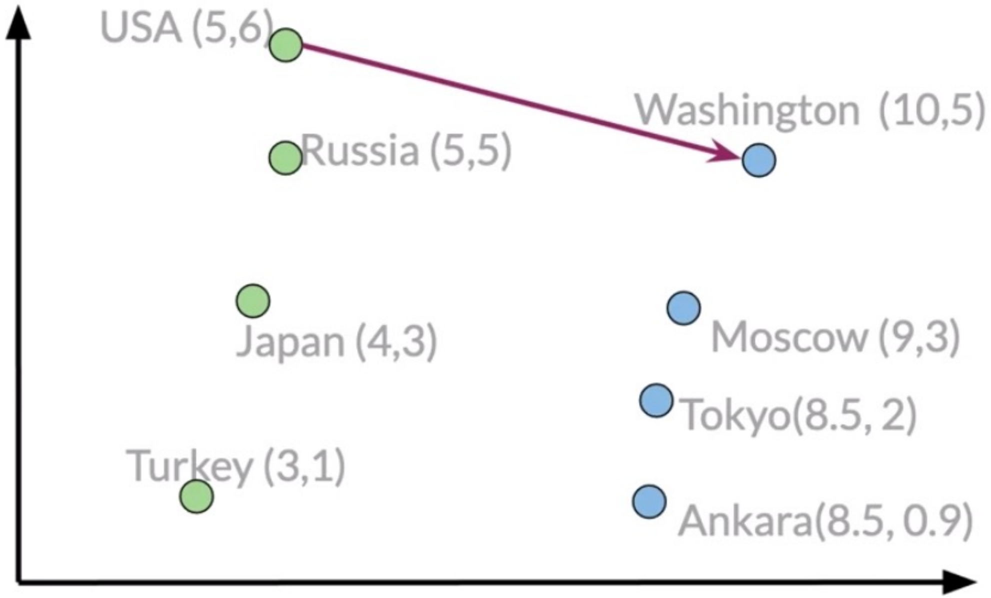
How to manipulate vector representations

- We can use them to infer unknown relations between words.
- For example you can use the relation between the USA and its capital to infer the capital of Russia !



Manipulating word vectors

- To find the relation between a country and its capital, you can use linear algebra.
  - Find the vector that leads you from a country to its capital (subtract one from the other)
  - This vector encodes the relationship « has capital »

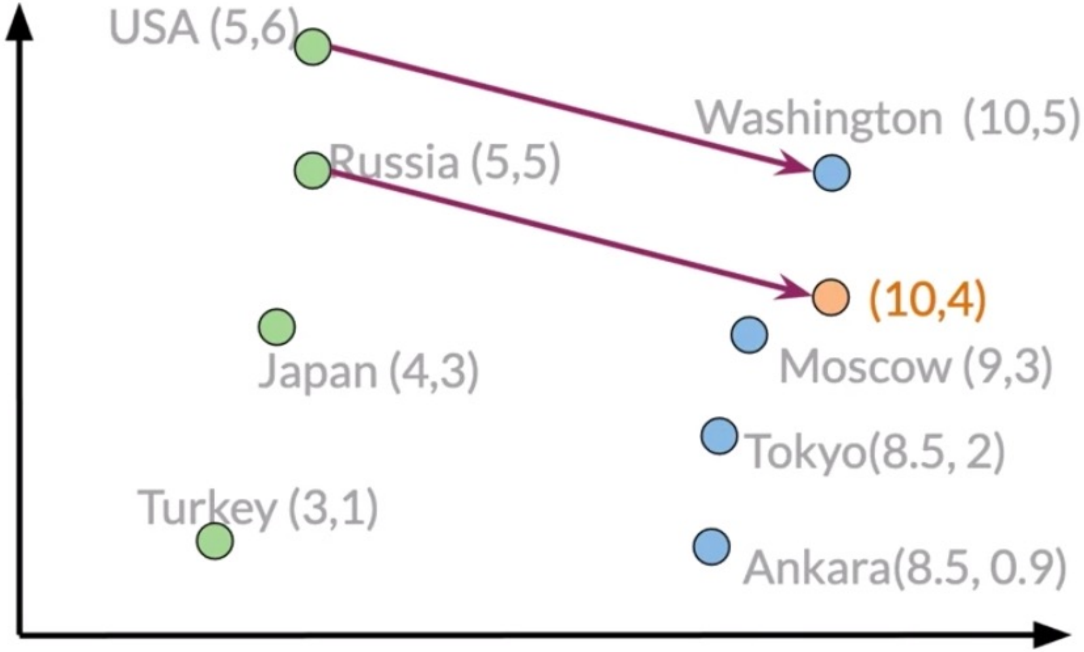


Washington - USA = $\begin{bmatrix} 5 & -1 \end{bmatrix}$

Remember a vector gives you directions, or how to move in space using its coordinates.

Making a prediction

- However the result of your computation may not land perfectly on the desired vector…
- So you need to use euclidian distance or cosine similarity to find the vector closest to your prediction



Washington - USA = $\begin{bmatrix} 5 & -1 \end{bmatrix}$

Russia + $\begin{bmatrix} 5 & -1 \end{bmatrix}$ = $\begin{bmatrix} 10 & 4 \end{bmatrix}$