

Hide menu

Lecture: Vector Space Models

- ✓ **Video:** Week Introduction
47 sec
- ✓ **Video:** Vector Space Models
2 min
- ✓ **Reading:** Vector Space Models
10 min
- ✓ **Video:** Word by Word and Word by Doc.
4 min
- ✓ **Reading:** Word by Word and Word by Doc.
10 min
- ✓ **Lab:** Linear algebra in Python with Numpy
1h
- ✓ **Video:** Euclidean Distance
3 min
- ✓ **Reading:** Euclidian Distance
10 min
- ✓ **Video:** Cosine Similarity: Intuition
2 min

📖 **Reading:** Cosine Similarity: Intuition
10 min

▶ **Video:** Cosine Similarity
3 min

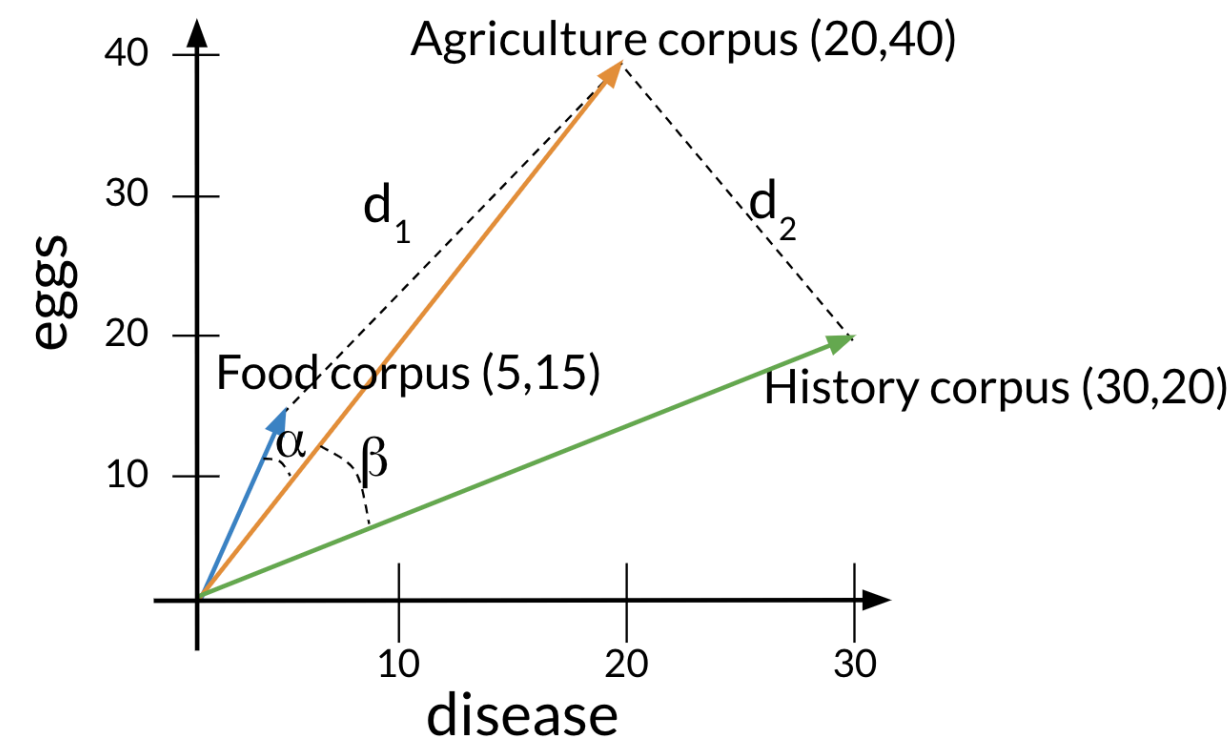
📖 **Reading:** Cosine Similarity
10 min

Week 3 > Cosine Similarity: Intuition

< Previous Next >

Cosine Similarity: Intuition

One of the issues with euclidean distance is that it is not always accurate and sometimes we are not looking for that type of similarity metric. For example, when comparing large documents to smaller ones with euclidean distance one could get an inaccurate result. Look at the diagram below:



Euclidean distance: $d_2 < d_1$

Angles comparison: $\beta > \alpha$

The cosine of the angle between the vectors

Normally the **food** corpus and the **agriculture** corpus are more similar because they have the same proportion of words. However the food corpus is much smaller than the agriculture corpus. To further clarify, although the history corpus and the agriculture corpus are different, they have a smaller euclidean distance. Hence $d_2 < d_1$.

To solve this problem, we look at the cosine between the vectors. This allows us to compare β and α .

