Text Similarity

- Text similarity is to measure how two documents are clos or far apart from each other.
- It can be used for plagiarism detection, finding similar questions in the question-and-answer websites, suggesting similar papers, recommending a product, etc.
- The vector representation of a document can be used to compare between texts. How? By measuring the distance between the document-representing vectors.
- Which metric distance to use?

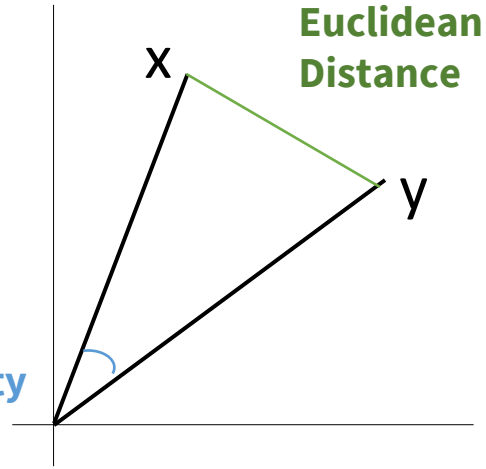
Text Similarity

Euclidean distance

$$d(x, y) = \|x - y\|$$

Cosine Similarity

$$k(x, y) = \frac{x^T y}{\|x\|\|y\|}$$



- Two documents can have different term frequencies or uneven lengths, but they can still share the same topic and be similar.
- Cosine similarity is a better measure that can capture similarity between two texts.
- Cosine similarity is defined in scikit-learn:

```
from sklearn.metrics.pairwise
import cosine_similarity
```