Solve the exercise on slide 10.

Calculate the cosine similarity between the symptoms of the reports.

- Report 1 → $r_1 = [1, 1, 1, 0, 0]$
- Report 2 → $r_2 = [1, 0, 0, 1, 1]$

Calculation:
$r_1 \cdot r_2 = 1\cdot1 + 1\cdot0 + 1\cdot0 + 0\cdot1 + 0\cdot1 = 1$
$\|r_1\| = \sqrt{(1+1+1)} = \sqrt{3}$
$\|r_2\| = \sqrt{(1+1+1)} = \sqrt{3}$

Cosine Similarity = $(r_1 \cdot r_2) / (\|r_1\|\|r_2\|) = 1 / (\sqrt{3}\cdot\sqrt{3}) = 1/3 \approx 0.333$

Is cosine similarity the same as the Euclidian distance when the vectors are normalized?

d_Euk(u, v) = $\sqrt{(2 - 2\cdot\cos(\theta))} = \sqrt{(2\cdot(1 - \text{CosineSimilarity}))}$

So higher cosine similarity ⇒ lower Euclidean distance.

Would the results profit from using SNOMED CT?

Yes. SNOMED CT links synonyms and related terms (e.g., “shortness of breath” = “dyspnea”), reducing wording differences and improving similarity accuracy.

**Make the code in Sentiment-NB.jpynb in the materials of week 3 run on your machine. The three data text files are in the folder data of week 3. Why is padding the text not
useful here?**

Padding isn’t needed because Naive Bayes uses bag-of-words features, not word order or sequence length. It just adds useless zeros. Padding is only useful for sequence models like RNNs or Transformers that need fixed-length inputs.