Introduction to LLM

Practice Session 2

CL/NLP Basics

The Focus of Practice Session

- We revisit the concepts from the lecture, but in a practical / hands-on way.
- For most concepts, we start with a tiny, simple implementation so you see how it works inside.
- Then we look at how to use the same idea with reliable library functions.
    - This is how you would actually do it in real code projects.
- This means you don’t only learn how to call it, but you also understand what it’s doing internally.
- In the practice session, the goal is:
    - to follow the logic and flow of the code, and
    - to be able to complete missing parts of the tasks in a reasonable, logical way.

Coding in Exam

- We focus on understanding, not memorization.
- You will not lose points for forgetting exact library function names.
- You will get points for writing the correct idea / logic.
- We care that you can show what the algorithm does, not just code syntax.
- Pseudo-code is OK if it shows the correct steps clearly.
- Syntax mistakes are not critical.
- Big logic mistakes are.

Correct implementation (valid code, logically correct)

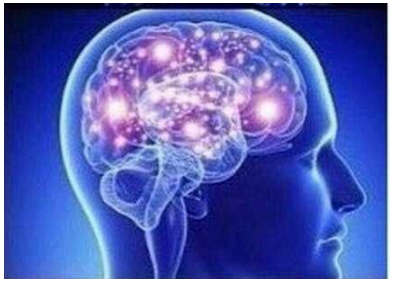
```
cosine_similarity(a, b):
    dot = 0.0
    for i from 0 to length(a)-1:
        dot = dot + a[i] * b[i]
    mag_a = 0.0
    mag_b = 0.0
    for i from 0 to length(a)-1:
        mag_a = mag_a + a[i] * a[i]
        mag_b = mag_b + b[i] * b[i]
    mag_a = sqrt(mag_a)
    mag_b = sqrt(mag_b)
    if mag_a == 0 OR mag_b == 0:
        return 0.0
    return dot / (mag_a * mag_b)
```

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Pseudo-code (not real code, but still logically correct)

cosine_similarity_pseudo(a, b):

```
dot  =  Σ over i of (a[i] * b[i])

len_a =  sqrt( Σ over i of (a[i]^2) )

len_b =  sqrt( Σ over i of (b[i]^2) )

if len_a == 0 OR len_b == 0:

    return 0

return dot / (len_a * len_b)
```

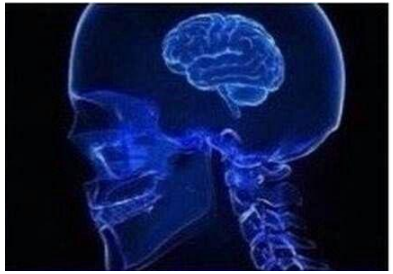


$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Wrong logic (looks like code, but is NOT cosine similarity)

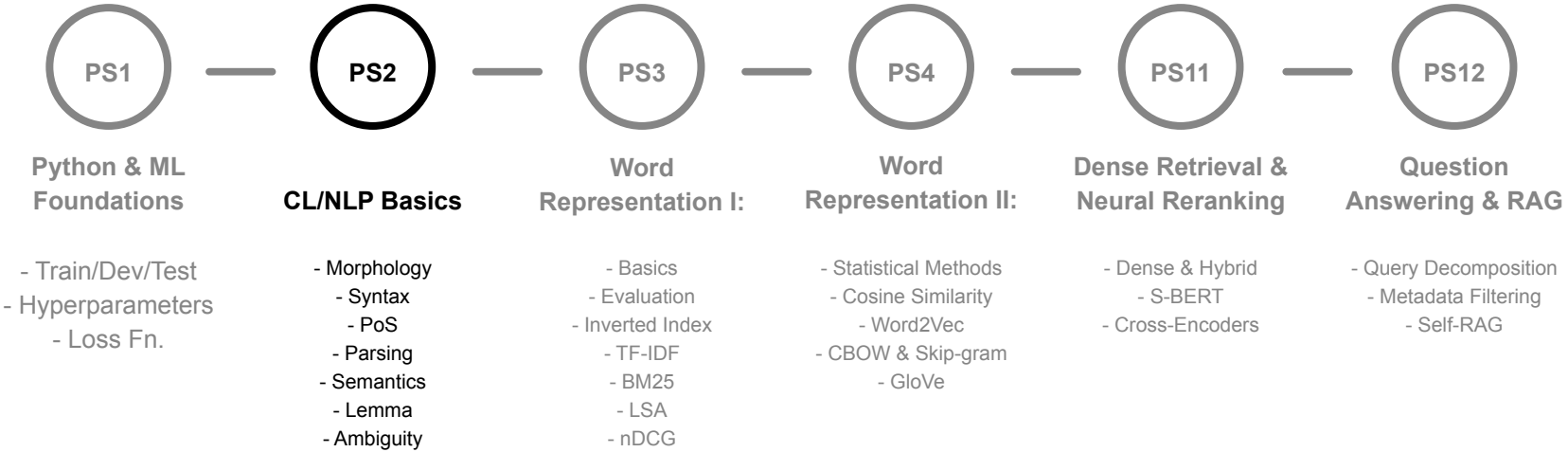
cosine_similarity_wrong(a, b):

```
dot = 0.0

sum_a = 0.0

sum_b = 0.0

for i from 0 to length(a)-1:

    dot   = dot   + a[i] * b[i]

    sum_a = sum_a + a[i]

    sum_b = sum_b + b[i]

return dot / (sum_a * sum_b)
```

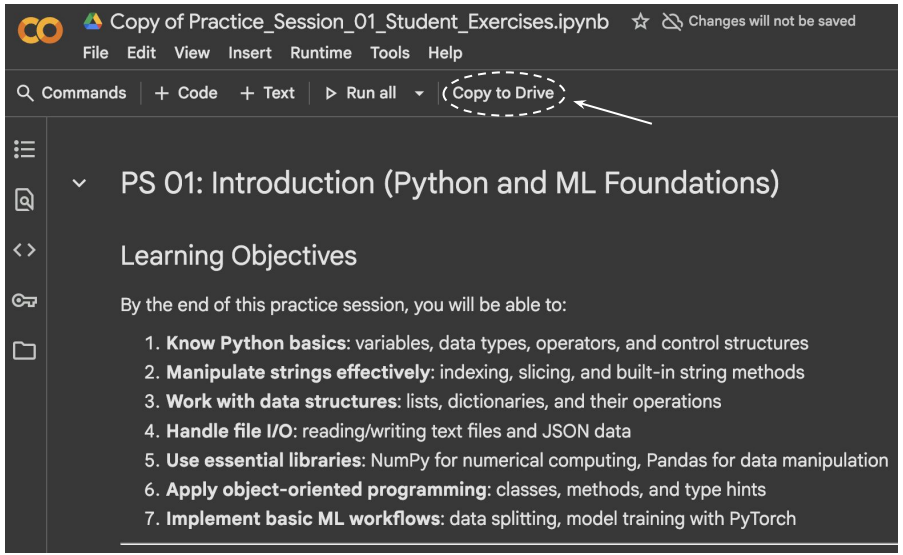


$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Timeline


PS1
PS2
PS3
PS4
PS11
PS12
Python & ML Foundations
- Train/Dev/Test
- Hyperparameters
- Loss Fn.
CL/NLP Basics
- Morphology
- Syntax
- PoS
- Parsing
- Semantics
- Lemma
- Ambiguity
Word Representation I:
- Basics
- Evaluation
- Inverted Index
- TF-IDF
- BM25
- LSA
- nDCG
Word Representation II:
- Statistical Methods
- Cosine Similarity
- Word2Vec
- CBOW & Skip-gram
- GloVe
Dense Retrieval & Neural Reranking
- Dense & Hybrid
- S-BERT
- Cross-Encoders
Question Answering & RAG
- Query Decomposition
- Metadata Filtering
- Self-RAG

PS2: Colab Notebook (Available on Moodle)



- https://colab.research.google.com/drive/1EFXK8CyUVjg2n7Bx1_QOCG_6b3Jbxqib