Assignment 2
Text Representation & Word Embeddings

Part A — Foundations

23 questions

★	**Must-know** — asked very frequently in GenAI / ML interviews (10 questions). Prepare these first.
[CODE]	**Hands-on task** — write and run the code (3 tasks).

1. What is a vector space model in NLP?
2. What is the Bag-of-Words model? What are its limitations? ★
3. What is a Bag of N-grams model? What problem does it solve over plain BoW?
4. What is TF-IDF? How do you calculate TF, IDF and TF-IDF? Work out a small example by hand. ★ · Asked: NLP interview rounds
5. What are word embeddings? How are they better than one-hot / BoW vectors? ★
6. Which algorithms are used to train word embeddings?
7. Explain Word2Vec and its two architectures (CBOW and Skip-gram). What is the difference? ★ · Asked: NLP interview rounds
8. What are the key hyperparameters of Word2Vec (vector_size, window, min_count, sg, negative) and their role?
9. Compare Word2Vec vs FastText vs GloVe vs Doc2Vec. · Asked: NLP interview rounds
10. What is the difference between static embeddings (Word2Vec) and contextual embeddings (BERT)? Give an example with the word "bank".
11. How do you handle out-of-vocabulary (OOV) words? ★
12. Explain cosine similarity and why it is important in NLP. ★
13. What similarity / distance metrics are used (cosine, Euclidean, dot product, Jaccard)?
14. Euclidean distance vs cosine similarity — when would you use which? ★
15. For measuring similarity, why do we use cosine and not sine or tan?
16. How do you check the similarity between two words or two documents?
17. Create vectors using CountVectorizer (unigrams and bigrams). [CODE]
18. Create vectors using TfidfVectorizer and print the top-weighted terms of a document. [CODE]

19. Train (1) Word2Vec, (2) FastText and (3) Doc2Vec on a small corpus and find similar words. [CODE]

20. Why can averaging Word2Vec vectors to make a sentence vector lose information? What is better? ★ · Asked: NLP interview rounds

21. What are Sentence Transformers (SBERT)? Why are they better for semantic search than averaged word vectors or BERT's raw [CLS] vector? ★ · Asked: NLP interview rounds

22. Pretrained embeddings vs training embeddings on your own domain data – which would you choose and when? · Asked: NLP interview rounds

23. Explain embeddings to a non-technical person without using the word "vector". ★ · Asked: Deloitte