Assignment 1 – Word Sense: Static Vs contextual embeddings

Task:

Discover how different families of embedding models treat the polysemous word **“bank.”**

You will compare **static word-level embeddings** (e.g., Word2Vec, GloVe) with **contextual sentence-level embeddings** (e.g., spaCy transformer) and analyze whether the neighbours you retrieve cluster around the *finance* sense, the *river-edge* sense, or both.

Take following 2 sentences and demonstrate

a) Embedding vectors for the Spacy models such as which use static lookups are the same
b) Models using contextual embeddings (such as) will produce different embedding vectors for the same word, based on context

"I sat on a river bank",

"I went to bank to deposit money"

Task:

1) Use a model such as Google News Word2Vec or Glove. Retrieve the 20 nearest neighbors of the token "bank" using cosine similarity.
2) Observe the outcomes – are most related to finance or to river bank?
3) Explain the outputs (as comments in code)
4) Find out which spacy models implement inference time embeddings generation and which one use static lookup
5) Implement code to demonstrate that the model that uses static lookup generates the same vector when processing the word “bank” in the 2 sentences
6) Implement code to demonstrate that the model that does inference time embedding generation is creating different embeddings for the same word “bank” in the context of the 2 sentences

Submit:

Link to completed notebook with the outputs in it

Assignment 2 – Discover themes in an Authoritative healthcare report

Task:

Use topic modelling on a WHO healthcare report to find the top 5 topics in the document.

1) **Create training Corpus:**
   Programtically do the following:
   Download the *Executive Summary* (pages 1-24) of the **“Global Tuberculosis Report 2024”** published by the World Health Organization in October 2024.
   Direct link (PDF, 5 MB): se
   e WHO publication page: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024?utm_source=chatgpt.com

   Programmatically download the PDF (e.g., requests), extract raw text with pdfminer.six or PyMuPDF, and print the first 500 characters.
2) **Clean and segment:**
   Remove tables/figures, lowercase, strip punctuation & stopwords (use NLTK’s English + WHO medical stop-list you create). Split into paragraph-level “documents”.
3) **Build baseline LDA**
   Vectorize with CountVectorizer(min_df=5, max_df=0.9, ngram_range=(1,2)).
   Train LDA for K = [5, 7, 9, 11]. Compute *c_v* coherence (gensim).
4) **Interpret topics:**
   For K* show top-10 words per topic. Give each topic a short label (“Funding-gap”, “Drug-resistance”, …) and write 1-sentence explanation.
5) **Visualize (optional)**
   Run pyLDAvis for K* and embed the interactive HTML (or take a static screenshot).

Submit:

Link to completed notebook with the outputs in it