**PROPOSED BLOCK DIAGRAM**

INPUT

Words

Phrases

Documents(Corpus)

EMBEDDINGS (n-dimensional)

Preferred model is text-embedding-ada-002

SIMILARITY EXTRACTION

Cosine Similarity  
Euclidean Similarity  
Jaccord Similarity  
Pearson/Sparman Correlation

VISUALIZATION

PREPROCESSING (***Chunking if required*)**

Tiktoken (we can also integrate regex/NLTK/ SpaCy)