Week 1 - AI Fundamentals Cheatsheet

1. TOKENS

- Smallest units of text (subwords, word pieces, or characters).
- Tokenization = breaking text into smaller pieces.
- Example: "Artificial Intelligence" -> ["Artificial", "Intelligence"].
- Libraries: transformers AutoTokenizer, tiktoken, nltk.

2. EMBEDDINGS

- Numerical vector representations of words, sentences, or documents.
- Capture semantic meaning: similar ideas are close in vector space.
- Example: "AI revolution" similar to "machine learning boom".
- Libraries: sentence-transformers, OpenAI embeddings, Hugging Face.

3. PIPELINES

- Simplify multi-step AI workflows.
- Example:

```
from transformers import pipeline
summarizer = pipeline("summarization", model="facebook/bart-large-cnn")
```

- Pipelines chain tokenization -> inference -> decoding -> output.
- In LangChain: used for embedding + retrieval + QA workflows.

4. SIMILARITY SEARCH

- Cosine Similarity = measure of semantic closeness (1 = identical, 0 = unrelated).
- Used in: document retrieval, semantic search, and question answering.
- Formula:

$$\text{CosSim}(A, B) = (A * B) / (||A|| * ||B||)$$

KEY TAKEAWAYS

- Tokenization breaks text into model-readable chunks.
- Embeddings translate meaning into numbers.
- Pipelines connect AI tasks seamlessly.

- Similarity search powers retrieval and context understanding.

Recommended Models & Tools

- Tokenizer: bert-base-uncased
- Embeddings: all-MiniLM-L6-v2
- Libraries: transformers, sentence-transformers, LangChain