## **🔹 Step 1: Understanding the Goal**

* **SemanticChunker** divides text into meaningful chunks **based on sentence similarity**.
* It **compares each sentence** to the previous one and **computes a similarity score**.
* If similarity **drops below a defined threshold**, a **new chunk** is created.

**🔸 Why?**

* It helps in **document summarization**, **text processing**, and **AI applications** like chatbots.

## **🔹 Step 2: Preparing the Input Text**

We use **plain text** as input, and it is **split into sentences**.

**Example text:**

plaintext

CopyEdit

Artificial intelligence (AI) is transforming industries worldwide.

Machine learning, a subset of AI, enables computers to learn from data and make predictions.

Deep learning, a more advanced form of machine learning, mimics the human brain using neural networks.

These technologies are widely used in healthcare, finance, and autonomous vehicles.

However, ethical concerns regarding AI, such as bias and job displacement, are growing.

Companies are working on AI regulations to ensure responsible usage.

Meanwhile, research in AI continues to evolve, leading to innovations in natural language processing.

Chatbots and virtual assistants are examples of AI applications that have become mainstream.

## **🔹 Step 3: Compute Sentence Similarity**

* We **convert each sentence into a vector representation** using a model like HuggingFaceEmbeddings().
* Then, we **calculate similarity scores** between consecutive sentences using **cosine similarity**.

| **Sentence Number** | **Sentence** | **Similarity Score (w.r.t. previous)** |
| --- | --- | --- |
| 1 | Artificial intelligence (AI) is transforming industries worldwide. | - |
| 2 | Machine learning, a subset of AI, enables computers to learn from data and make predictions. | **0.92** |
| 3 | Deep learning, a more advanced form of machine learning, mimics the human brain using neural networks. | **0.90** |
| 4 | These technologies are widely used in healthcare, finance, and autonomous vehicles. | **0.85** |
| 5 | However, ethical concerns regarding AI, such as bias and job displacement, are growing. | **0.65** |
| 6 | Companies are working on AI regulations to ensure responsible usage. | **0.72** |
| 7 | Meanwhile, research in AI continues to evolve, leading to innovations in natural language processing. | **0.78** |
| 8 | Chatbots and virtual assistants are examples of AI applications that have become mainstream. | **0.80** |

## **🔹 Step 4: Apply Different Chunking Methods**

Now, we use different **thresholding methods** to determine **breakpoints**.

### **1️⃣ Probability-Based Chunking (probability, amount=0.5)**

📌 **Rule**: If similarity drops **below 0.5**, create a new chunk.

**Threshold = 0.5** Since **the lowest similarity is 0.65 (above 0.5)**, **no chunking happens**.

* If we set **threshold = 0.75**, chunking happens **after Sentence 4**.

#### **Result:**

pgsql

CopyEdit

Chunk 1:

Artificial intelligence (AI) is transforming industries worldwide.

Machine learning, a subset of AI, enables computers to learn from data and make predictions.

Deep learning, a more advanced form of machine learning, mimics the human brain using neural networks.

These technologies are widely used in healthcare, finance, and autonomous vehicles.

Chunk 2:

However, ethical concerns regarding AI, such as bias and job displacement, are growing.

Companies are working on AI regulations to ensure responsible usage.

Meanwhile, research in AI continues to evolve, leading to innovations in natural language processing.

Chatbots and virtual assistants are examples of AI applications that have become mainstream.

**📝 Explanation:**

* The **first 4 sentences** are **one chunk** (high similarity).
* A **new chunk starts** after **Sentence 4** (similarity **drops to 0.65**).

### **2️⃣ Standard Deviation-Based Chunking (standard\_deviation, amount=1.25)**

📌 **Rule**: If similarity **drops below Mean - (1.25 × SD)**, create a new chunk.

#### **Step 1: Calculate Mean and Standard Deviation**

* **Mean Similarity** = **0.80**
* **Standard Deviation (SD)** = **0.10**
* **Threshold** = Mean - (1.25 × SD) = 0.80 - 0.125 = **0.675**

#### **Step 2: Compare Similarities**

* **Sentence 5** has **0.65** (< **0.675**) → **Chunk break occurs.**

#### **Result:**

pgsql

CopyEdit

Chunk 1:

Artificial intelligence (AI) is transforming industries worldwide.

Machine learning, a subset of AI, enables computers to learn from data and make predictions.

Deep learning, a more advanced form of machine learning, mimics the human brain using neural networks.

These technologies are widely used in healthcare, finance, and autonomous vehicles.

Chunk 2:

However, ethical concerns regarding AI, such as bias and job displacement, are growing.

Companies are working on AI regulations to ensure responsible usage.

Meanwhile, research in AI continues to evolve, leading to innovations in natural language processing.

Chatbots and virtual assistants are examples of AI applications that have become mainstream.

**📝 Explanation:**

* **Adaptive threshold** based on text distribution.
* **Chunking occurs** when similarity **drops below 0.675**.

### **3️⃣ IQR-Based Chunking (interquartile, amount=0.5)**

📌 **Rule**: If similarity **drops below Q1 - (0.5 × IQR)**, create a new chunk.

#### **Step 1: Calculate IQR**

* **Q1 (25th percentile) = 0.72**
* **Q3 (75th percentile) = 0.90**
* **IQR = Q3 - Q1 = 0.90 - 0.72 = 0.18**
* **Threshold** = Q1 - (0.5 × IQR) = 0.72 - 0.09 = **0.63**

#### **Step 2: Compare Similarities**

* Sentence 5 has **0.65** (> **0.63**) → **No break**.
* If we set **a stricter threshold (e.g., 0.70)**, a **chunk break would occur.**

#### **Result:**

pgsql

CopyEdit

Chunk 1:

Artificial intelligence (AI) is transforming industries worldwide.

Machine learning, a subset of AI, enables computers to learn from data and make predictions.

Deep learning, a more advanced form of machine learning, mimics the human brain using neural networks.

These technologies are widely used in healthcare, finance, and autonomous vehicles.

However, ethical concerns regarding AI, such as bias and job displacement, are growing.

Chunk 2:

Companies are working on AI regulations to ensure responsible usage.

Meanwhile, research in AI continues to evolve, leading to innovations in natural language processing.

Chatbots and virtual assistants are examples of AI applications that have become mainstream.

**📝 Explanation:**

* IQR-based **threshold adapts better to data distribution**.
* More **natural segmentation** than standard deviation.

## **🔹 Step 5: Final Comparison**

| **Method** | **Breakpoint Formula** | **Effect** |
| --- | --- | --- |
| **Probability** | If similarity < 0.5 | **Fixed threshold**, not flexible |
| **Standard Deviation** | If similarity < Mean - (1.25 × SD) | **Adaptive**, works with variable text |
| **IQR** | If similarity < Q1 - (0.5 × IQR) | **Data-driven**, best for topic shifts |

## **🔹 Summary**

* **Probability method** → Simple but rigid.
* **Standard Deviation** → Adaptive, works well with variation.
* **IQR method** → Best for **structured text segmentation**.

Would you like **Python code implementation** for this? 🚀

4o