Visualization of Your Search Engine Work (Stages 1 & 2)

Sample Documents (Your Dataset)

```
Doc0: apple banana apple
Doc1: banana mango
Doc2: apple mango mango
```

Step 1 — Lexicon (Vocabulary Dictionary)

What happens:

- Every unique word in the dataset gets a **term ID**.
- We also count **how many documents each word appears in** (Document Frequency, DF).

Term	Term ID	DF
apple	0	2
banana	1	2
mango	2	2

Visualization:

```
Lexicon:
+--------+--------+----+
| Term   | TermID | DF |
+--------+--------+----+
| apple  | 0      | 2  |
| banana | 1      | 2  |
| mango  | 2      | 2  |
+--------+--------+----+
```

Metaphor: Lexicon = **dictionary of all words** your engine knows.

Step 2 — Forward Index (Document → Terms)

What happens:

- Each document is represented **from the perspective of the document**.
- Stores which **terms appear, how often, and where**.

Forward Index Table

DocID	Term IDs	Term Freqs	Positions	Length
0	[0, 1, 0]	[2,1]	[0,1,2]	3
1	[1, 2]	[1,1]	[0,1]	2
2	[0, 2, 2]	[1,2]	[0,1,2]	3

Visualization (Metaphor):

```
Document 0: "apple banana apple"
doc_id = 0
term_ids = [apple:0, banana:1, apple:0]
term_freqs = {apple:2, banana:1}
positions = [0,1,2]
length = 3

Document 1: "banana mango"
doc_id = 1
term_ids = [banana:1, mango:2]
term_freqs = {banana:1, mango:1}
positions = [0,1]
length = 2
```

Diagram Style:

```
[Doc0] "apple banana apple"
   |--- apple (term_id:0) freq:2 pos:[0,2]
   |--- banana(term_id:1) freq:1 pos:[1]

[Doc1] "banana mango"
   |--- banana(term_id:1) freq:1 pos:[0]
   |--- mango (term_id:2) freq:1 pos:[1]

[Doc2] "apple mango mango"
   |--- apple (term_id:0) freq:1 pos:[0]
   |--- mango (term_id:2) freq:2 pos:[1,2]
```

Metaphor: Forward Index = **annotated notebook of each document**, showing **what words appear, how many times, and where**.

Step 2.3 — Segmentation

For large datasets:

```
forward_index_0.bin → Doc0-Doc999
forward_index_1.bin → Doc1000-Doc1999
...
```

Metaphor: Like breaking notebooks into chapters so you don't have to carry the whole library at once.

Combined View — Lexicon + Forward Index

```
Lexicon:                     Forward Index:

Term      TermID              DocID  TermIDs          TermFreqs
apple       0                 0      [0,1,0]        [2,1]
banana      1                 1      [1,2]          [1,1]
mango       2                 2      [0,2,2]        [1,2]
```

Flow:

```
User query → look up term_id in lexicon → get doc info from forward
index → use term frequencies/positions → rank documents
```

💡 Key Points of This Visualization

1. **Lexicon** is **term-centric** (what the engine knows globally).
2. **Forward Index** is **document-centric** (how each document contains terms).
3. **Segmented binary files** make it scalable.
4. Positions allow **phrase search**; term frequencies allow **BM25 ranking**.