# Assignment 1

**Question 1:**

1. How many documents are there in this corpus?
2. **84474**
3. Why different fields are treated with different kinds of java class? i.e., StringField and TextField are used for different fields in this example, why?
4. **StringField** :
   * It is a field that is indexed but not tokenized.
   * In the given example DOCNO is not needed to be tokenized because it is a numeric value tokenizing doesn’t make sense.
5. **TextField**:

- A field that is indexed and tokenized.

- The fields [TEXT],[BYLINE],[HEAD] consists of text, I thinks it’s better to tokenize these particular fields in index.

**Question 2:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analyzer** | **Tokenization applied?** | **How many tokens are there for this field?** | **Stemming applied?** | **Stop words removed?** | **How many terms are there in the dictionary?** |
| **Keyword Analyzer** | No | 84837 | No | No | 84043 |
| **Simple Analyzer** | Yes | 37330144 | No | No | 169981 |
| **Stop Analyzer** | Yes | 26216475 | No | Yes | 169948 |
| **Standard Analyzer** | Yes | 26649680 | No | Yes | 233384 |