|  |  |  |
| --- | --- | --- |
| **Topic** | **Simple Definition** | **Analogy / Real-World Example** |
| **Text Analytics** | Analyzing large amounts of text to find patterns and insights. | Like reading thousands of reviews and summarizing emotions and themes automatically. |
| **Corpus** | A collection of text documents used for analysis. | A bookshelf where each book is a document. |
| **Text Preprocessing** | Cleaning the text (removing punctuation, converting to lowercase, etc.) | Like washing and peeling vegetables before cooking. |
| **Tokenization** | Breaking text into smaller parts (words or sentences). | Like cutting a pizza into slices so it's easier to eat/analyze. |
| **Stop Words** | Common words like "the", "is", "and" that are usually removed from analysis. | Like skipping filler words when reading a book summary. |
| **Stemming** | Reducing words to their root form. | Grouping "running", "runner", and "ran" into "run". |
| **Bag of Words (BoW)** | Representing text as the count of words, ignoring grammar and order. | Dumping all words from a document into a bag and counting each one. |
| **TF-IDF** | Highlights rare and unique words in a document compared to others. | Everyone says “hello”, but only one says “antidisestablishmentarianism” — that word stands out! |
| **Word Cloud** | A visual where more frequent words appear larger. | Like a popularity chart — big names mean more mentions. |
| **Facebook Seamless Dataset** | A multilingual dataset for training language and speech models. | A giant global chat archive for building translation tools like Siri or Google Translate. |