LUCENE - ANALYSIS

http://www.tutorialspoint.com/lucene/lucene_analysis.htm

As we've seen in one of the previous chapter *Lucene - Indexing Process*, Lucene uses *IndexWriter* which analyzes the *Document*s using the *Analyzer* and then creates/open/edit indexes as required. In this chapter, we are going to discuss various types of Analyzer objects and other relevant objects which are used during analysis process. Understanding Analysis process and how analyzers work will give you great insight over how lucene indexes the documents.

Following is the list of objects that we'll discuss in due course.

Sr. No.	Class & Description
1	Token Token represents text or word in a document with relevant details like its metadata *position, startoffset, endoffset, tokentypeanditspositionincrement*.
2	TokenStream TokenStream is an output of analysis process and it comprises of series of tokens. It is an abstract class.
3	Analyzer This is abstract base class of for each and every type of Analyzer.
4	WhitespaceAnalyzer This analyzer spilts the text in a document based on whitespace.
5	SimpleAnalyzer This analyzer spilts the text in a document based on non-letter characters and then lowercase them.
6	StopAnalyzer This analyzer works similar to SimpleAnalyzer and remove the common words like 'a','an','the' etc.
7	StandardAnalyzer This is the most sofisticated analyzer and is capable of handling names, email address etc. It lowercases each token and removes common words and punctuation if any.