INDEXING with POI and LUCENE

As my previous post shows how to index PDF Documents with Lucene, I thought that it would be worth to post how to index Microsoft format files too because those file types are very commonly used. Lucene always requires a String in order to index the content and therefore we need to extract the text from the document before giving it to Lucene for indexing. To parse the document we can use Apache POI which provides a Java API for Microsoft format files.