Mit der Hilfe von Lucene indexierte ich eine Auswahl aus der privaten MP3-Sammlung. Um nicht die ganze Sammlung ständig kopieren zu müssen, nahm ich zur Realisierung nur eine kleine Testmenge aus der Sammlung heraus. Da die Indexierung bei beiden Mengen gleich funktioniert, konnte ich so die Dauer der Tests kürzen. In den Tests verglich ich die Dauer der Indexierung, die Dauer der Suche und natürlich die Ergebnisse, welche eine Suchanfrage zurückgab. Die verschiedenen Analyzer von Lucene zeigten dabei einige Unterschiede.

Die Analyzer

Lucene nutzt die sogenannten Analyzer, um Informationen aus den Dateien zu extrahieren. Im Normalfall werden Textdateien eingelesen und indexiert. Da dieses Projekt aber auf MP3-Dateien und deren ID3-Tags basiert, enthielten die entsprechenden Dateien keine grossen Texte. Für Lucene sind MP3s nicht die optimalsten Dateien. Die MP3s enthalten vermehrt Informationen wie Titel, Interpret, Album, usw. in den ID3-Tags. Die Tags werden extrahiert und mit Lucene in den Index geschrieben. In diesen Feldern sind meistens keine grossen Texte zu finden, daher wurden auch die Songtexte noch zusätzlich miteinbezogen.

Die Analyzer selbst teilen die Texte nach bestimmten Kriterien auf und ignorieren bestimmte Wörter. Die ignorierten Wörter werden auch als Stopp-Wörter bezeichnet. Verschiedene Methoden zur Worterkennung sind ebenfalls vorhanden. Die Arbeit konzentrierte sich auf vier Analyzer: Den Standard-Analyzer, den Stop-Analyzer, den Whitespace-Analyzer und den Simple-Analyzer. Der Standard-Analyzer erkannte Nicht-Buchstaben und konnte damit Kreationen wie X&Y aufteilen. Der Whitespace-Analyzer hingegen findet solche Konstellationen nicht. Für die Worterkennung nutzen alle die Leerzeichen als Trennung. Der Simple-Analyzer und Whitespace-Analyzer waren in den Tests die schnellsten im Indexieren. Hingegen war die Suchleistung teilweise schlecht und es wurden Ergebnisse nicht gefunden. Der Standard-Analyzer hatte sich als gute Option herausgestellt. Die Indexierung dauerte einerseits ein wenig länger, allerdings lieferte die Suche gute Ergebnisse in einer anständigen Zeit. Als Fazit muss man sagen, dass je nach Einsatzgebiet ein bestimmter Analyzer ausgewählt werden sollte. Für eine schnelle Indexierung würde ich den Whitespace-Analyzer bevorzugen, für schnelle Suchergebnisse wäre der Standard-Analyzer die erste Wahl.

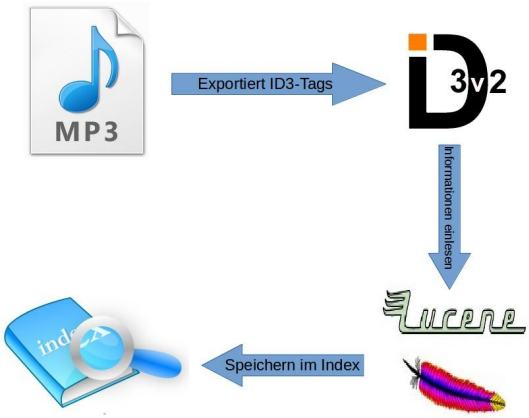

Abbildung 1: Ablauf des Informationsgewinns aus den MP3-Dateien

Suche nach Musikmustern

Die Königsdisziplin in der Suche mit Musik, ist wohl die Suche nach Musikmustern. Shazam ist ein Vorzeigebeispiel, wie eine solche Suchmaschine realisiert werden kann. Es reichen wenige Sekunden aus, um ein Lied korrekt zu erkennen. Dabei wird ein akkustischer Fingerabdruck erstellt und mit einem Eintrag in einer Datenbank verglichen. Der akkustische Fingerabdruck für ein Lied bleibt gleich, egal ob man nur wenige Sekunden oder das komplette Lied zur Verfügung hat. Der Fingerabdruck selbst basiert auf der spektralen Flachheit. Das heisst, dass harte rythmische Änderungen als eine 1 registriert werden. Gleichmässiger Singsang wird hingegen als 0 erkannt. So reichen wenige Sekunden aus, die spektrale Flachheit zu bestimmen und somit den Fingerabdruck zu berechnen.