**LUCENE IN ACTION**

# CHAPTER 5:

## Indexing a Microsoft Word document

Hiện nay bộ soạn thảo Microsoft Word vẫn là phần mềm được sử dụng nhiều nhất trong soạn thảo văn phòng. Vì vậy việc thực hiện indexing được MS Word là một việc thiết thực cần phải làm , cần phải biết cách parse tài liệu MS Word để có thể tìm kiếm với Lucene. Nhưng sự khó khăn nhất gặp phải đó là MS Word là chương trình có bản quyền vì vậy format của nó cũng được giấu kín. Nhưng may thay , có một vài project mã nguồn mở có thể khắc phục được vấn đề này . Trong phần này có giới thiệu 2 project mã nguồn mở Jakarta POI và TextMining.org dùng để rút trích text ra.