**Major features:** 將大量Html檔案切成小檔案進行爬蟲分析，不使用額外空間儲存資料而是邊爬蟲邊處理，達到我們期待的low storage。

**Major difficulties encountered:**

最大的困難在於無論使用Python 或 Java爬蟲都會遇到html tag只能抓到一個的問題，而雲端資料一個資料檔卻有多個Html檔案，所以必須自行寫一個Parser先取出一個Html檔案，再進行爬蟲，在設計Parser上遇到較多困難也花較多時間。

**Instructions for installation/compilation/ configuration/execution environment:**

1. 下載lucene Library。
2. 將 Library import 到Intellij IDE 裡。
3. 先呼叫library裡頭的method確認執行狀況。
4. 使用Intellij IDE進行軟體開發 完成作業輸出格式。

**Team members list:**

**107598063 劉冠志** 負責把多個Html切成單一Html方便爬蟲以及使用 index 資料結構。

**107598058 朱峻平** 負責使用Java Jsoup進行爬蟲資料分析以及切出有效Token。