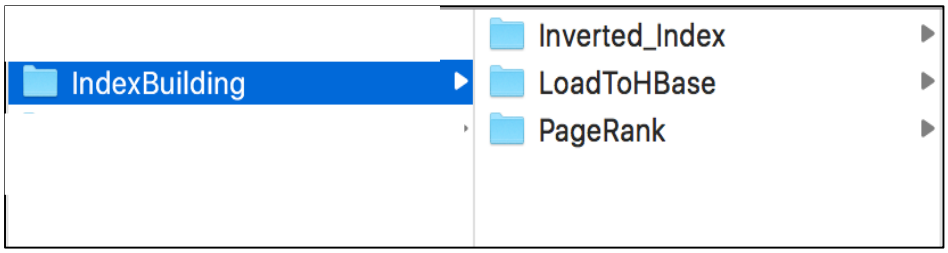
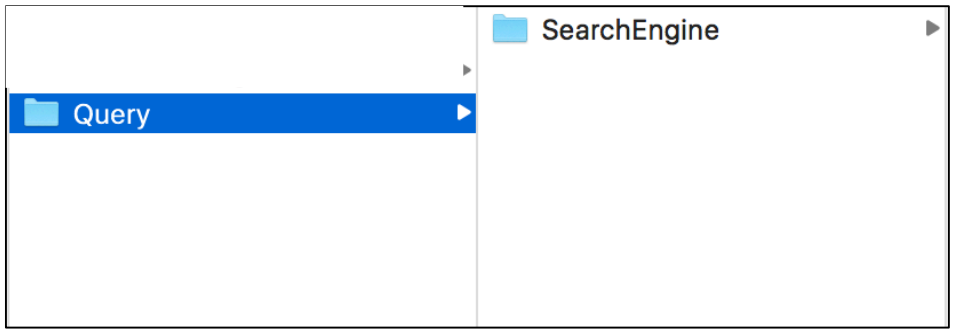
CloudProgramming HW3-Search Engine Report

學號姓名：101062231 林展逸

1. **File**

共有兩個部分，第一部分是 IndexBuilding，第二部分是 Query





2. **Instruction**

首先是 IndexBuilding

(1) Inverted_Index:

Compile: sh compile.sh

Run: sh run.sh

(從 executive.sh 中更改 input file)

(2) PageRank:

Run:

spark-submit --class PageRankSpark --num-executors 30 target/scala-2.10/page-rank-spark_2.10-1.0.jar input(ex:hdfs:///shared/HW2/sample-in/input-100M)

(3) LoadToHBase

Compile: sh compile.sh

Run: sh execute.sh

再來是 Query 的部分

(4) SearchEngine (change the input file in the code SearchEngine.java)

Compile: sh compile.sh

Run: sh run.sh

Search: 輸入 => search “XXX and XXX”

Ex: search “cat and dog”

3. Implementation

(1) 首先是 HBase 的 schema。

HBase 中共有兩個 Table，分是”s101062231:Inverted”,”s101062231:PageRank”

s101062231:Inverted

word	df	invertedInfo
gypsy	2	Growel's 101 1.0 [85920942]<maindiv>Wikipedia:WikiProject Spam/LinkReports/sydroger.blogspot.com 2.0 [66489813,66490285]
XXX	df	PageName1 tf [offset]<maindiv>PageName2 tf [offset]<maindiv>

s101062231:PageRank

page	pageRank
Page1	X.XXXXXXXX
Page2	X.XXXXXXXX

(2) 再來是 Query 的步驟

我們假設使用者搜尋”cat”、”dog”兩個字

先從 s101062231:Inverted 取出這兩個字的 **df** 及 **invertedInfo**。

再來先對照這兩個字的 pageList，找出同時包含”cat”、”dog”的 Page，同時算出他們的 TFIDF 值，如果同時包含”cat”、”dog”，則將 TFIDF 相加。

另外 offset 採 append 上去的方式。

接著取出 TFIDF 最高的 10 個 Page。這 10 個 Page 表示與”cat”、”dog”最具相關性。

接著從 s101062231:PageRank 中取出這 10 個 Page 的 PageRank。

再來針對 pageRank 在從高到低排序一次，從最 popular 到最不 popular。

(3) 最後從剛剛紀錄的 offset 中取前三個 offset，把 fragment 印出來。

```
**********
Search:
**********
search:Romani
search:gypsy
**********
Rank 1  Second Battle of Gaza   score = 8.231412596370928E-6
**********
[offest = 74953495]:
=>  rtillery decimated the attackers.  As a result of the EEF victories at the [[Battle of Romani]], [[Battle of Magdhaba]], an
d [[Battle of Rafa]] fought from August 1916 to January 1917, the EEF had pushed the defeated Ottoman Army eastwards. The EEF r
eoccupied the [[Egypt]]ian territory of the [[Sinai Peninsula]], and crossed over into the [[Ottoman Empire]] territory of sout
hern [[Palestine (region)|Palestine]]. However, the result of the First Battle of Gaza had been as close to a [[British Empire]
] victory as a defeat could get. In the three weeks between the two battles, the Gaza defences were strongly reinforced against
 a frontal attack. The strong entrenchments and fortifications proved unassailable during the disastrous frontal attacks, when
EEF casualties approached, and in some cases exceeded 50 per cent for slight gains.  ==Background== {{see also|Battle of Romani
|Battle of Magdhaba|Battle of Rafa|First Battle of Gaza}}  An 11 January [[War Cabinet]] decision to reduce large scale operati
ons in Palestine was reversed on the 26 February [[Supreme War Council|Anglo-French Congress]], and the [[Egyptian Expeditionar
y Force]] (EEF) was now required to capture the stronghold of Gaza as a first step towards Jerusalem.&lt;ref name=&quot;Woodwar
d68–9&quot;&gt;Woodward 2006, p. 68–9&lt;/ref&gt; Gaza was one of the most ancient cities in the world, being one of five city-
states mentioned in the Bible as ruled by the [[Philistines]], and had been fought over many times during its 4,000-year histor
y. The [[Egyptians]] and the [[Assyria]]ns had attacked Gaza, followed in 731 BC by the [[Greeks]],  with [[Alexander the Great
|Alexander]] conducting three attacks and the [[Siege of Gaza]] in 332 BC. The town was completely destroyed in 96 BC and rebui
lt slightly to the south of the original site. This Gaza was captured by [[Caliph Omar]] in 635 AD, by [[Saladin]] in 1187 AD,
and by [[Napoleon]] in 1799.&lt;ref&gt;Falls 1930 Vol. 1 p. 281&lt;/ref&gt; At Gaza there was a
[offest = 74954287]:
@
```