Information Retrieval Hw1 Team4

**Major features**

* High efficency Parser using own Regex algorithm
* Case folding Tokenization
* Low ram usage–

MapReduce Architecture for parsing document and building inverted index

10 mins building index

500 mb ram usage parsing document

* SPIMI-Invert algorithm adopt (without sorting terms)
* SPIMI: Compression - Dictionary-as-a-String
* Query processing: spell-correction

**Major difficulties encountered**

Inefficiency of parsing document lead to long time of index building

Over usage of memory lead to program termination

Doesn’t know how to build vector space model

**Team members list**

四電資四 99820318 陳科銘 main code, parser, index building, query

四電資四 99820323 謝宗廷 report, vector space model, map reduce

四電資三 100820315 高睿廷 report

四電資三 100820334 林純正 report