Inverted Index Construction

Technology/Concept:

- **Text Processing**: Building an inverted index is commonly associated with search engines and information retrieval systems. For large datasets, this process can be handled efficiently using **Apache Hadoop (MapReduce)** or **Apache Spark**.
- **Distributed File Systems (HDFS)**: If the number of documents is very large, these can be stored and processed using **HDFS** or similar distributed storage systems.
- **Indexing Systems**: Tools like **Apache Lucene** or **Elasticsearch** are typically used in real-world implementations to create and store inverted indexes.

Steps:

1. Read the text file and create RDD.
2. Parsing the content to get key value pairs (tuples within a tuple). This contains word and the corresponding document.
3. Generating frequency of each using mapper
4. Formatting the RDD to get desired result
5. Grouping the result by Key
6. Sorting the output by keys
7. Formatting the result to get the desired format

Output:

```
In [2]:
# Reading the given textfiles. As seen from the output we have got a list which has tuples.
textFile = sc.wholeTextFiles("C:/Users/ysrivastava/BDA/Assignment/InvertedIndex/*.txt")
textFile.collect()
```

```
Out[2]: [('file:/C:/Users/ysrivastava/BDA/Assignment/InvertedIndex/d1.txt',
  'China officially the People Republic of China is a sovereign state located in East Asia\nIt is the world most populous count
ry\nThe People Republic of China is a single party state governed by the Communist Party with its seat of government in the cap
ital city of Beijing exercises jurisdiction many provinces five autonomous regions four direct controlled municipalities Beijin
g Tianjin Shanghai and Chongqing and two mostly self governing special administrative regions Hong Kong and Macau\nChina also c
laims Taiwan which is controlled by the Republic of China a separate political entity as its extra province a claim which is co
ntroversial due to the complex political status of Taiwan'),
 ('file:/C:/Users/ysrivastava/BDA/Assignment/InvertedIndex/d2.txt',
  'India officially the Republic of India is a country in South Asia\nIt is the seventh largest country by area the second most
populous country\nThe most populous democracy in the world\nBounded by the Indian Ocean on the south the Arabian Sea on the sou
th west and the Bay of Bengal on the south east\nIt shares land borders with Pakistan to the west China Nepal and Bhutan to the
north east and Burma and Bangladesh to the east\nIn the Indian Ocean India is in the vicinity of Sri Lanka and the Maldives\nIn
addition India Andaman and Nicobar Islands share a maritime border with Thailand and Indonesia'),
 ('file:/C:/Users/ysrivastava/BDA/Assignment/InvertedIndex/d3.txt',
  'Relations between India and Pakistan have been strained by a number of historical and political issues and are defined by th
e violent partition of British India \nThe Kashmir dispute and the numerous military conflicts fought between the two nations\n
Consequently even though the two South Asian nations share historic cultural ethnic geographic and economic links their relatio
nship has been plagued by hostility and suspicion\nAfter the dissolution of the British Raj two new sovereign nations were form
ed the Union of India and the Dominion of Pakistan\nThe subsequent partition of the former British India displaced millions of
people with estimates of loss of life varying from several hundred thousand to a million\nIndia emerged as a secular nation wit
h a Hindu majority population and a large Muslim minority while Pakistan was established as an Islamic republic with an overwhe
lming Muslim majority population\nSoon after their independence India and Pakistan established diplomatic relations but the vio
lent partition and numerous territorial disputes would overshadow their relationship\nSince their independence the two countrie
s have fought three major wars one undeclared war and have been involved in numerous armed skirmishes and military standoffs\nT
he Kashmir dispute is the main centre point of all of these conflicts with the exception of the India Pakistan War and Banglade
sh Liberation War which resulted in the secession of East Pakistan now Bangladesh'),
 ('file:/C:/Users/ysrivastava/BDA/Assignment/InvertedIndex/d4.txt',
  'The India US China Pakistan strategic quadrilateral\nAlthough the disputed border between China and India is often highlight
ed as the major sticking point in Sino Indian relations in reality it has remained relatively peaceful since the end of a war t
he potential for overt military conflict in the region remains minimal\nOf much greater concern is the strategic quadrilateral
relationship in South Asia involving China India the United States and Pakistan\nIt has both regional and wider implications\nA
```

```
In [32]:
Kep_pair1 = textFile.flatMap(lambda x: map_out(x)).cache()
Kep_pair1.collect()
# From the output below, we are getting key value pairs(tuples within a tuple). This contains w
```

```
Out[32]: [(('china', 'd1'), 1),
 (('officially', 'd1'), 1),
 (('the', 'd1'), 1),
 (('people', 'd1'), 1),
 (('republic', 'd1'), 1),
 (('of', 'd1'), 1),
 (('china', 'd1'), 1),
 (('is', 'd1'), 1),
 (('a', 'd1'), 1),
 (('sovereign', 'd1'), 1),
 (('state', 'd1'), 1),
 (('located', 'd1'), 1),
 (('in', 'd1'), 1),
 (('east', 'd1'), 1),
 (('asia', 'd1'), 1),
 (('it', 'd1'), 1),
 (('is', 'd1'), 1),
 (('the', 'd1'), 1),
 (('world', 'd1'), 1),
 (('most', 'd1'), 1),
```

```
(('most', 'd1'), 1),
```

In [33]:

```
# Setting up our reducer - This will give us the frequecy of each word with respect to th
from operator import add
red = Kep_pair1.reduceByKey(add)
red.collect()
```

Out[33]:

```
[(('china', 'd1'), 5),
 (('officially', 'd1'), 1),
 (('the', 'd1'), 7),
 (('people', 'd1'), 2),
 (('republic', 'd1'), 3),
 (('of', 'd1'), 6),
 (('is', 'd1'), 5),
 (('a', 'd1'), 4),
 (('sovereign', 'd1'), 1),
 (('state', 'd1'), 2),
 (('located', 'd1'), 1),
 (('in', 'd1'), 2),
 (('east', 'd1'), 1),
 (('asia', 'd1'), 1),
 (('it', 'd1'), 1),
 (('world', 'd1'), 1),
 (('most', 'd1'), 1),
 (('populous', 'd1'), 1),
 (('country', 'd1'), 1),
```

In [34]:

```
new = red.map(lambda x: (x[0][0],(x[0][1], x[1])))
new.collect()
# Formating the output, which should keep the word as key and its corresponding information as value
```

Out[34]:

```
[('china', ('d1', 5)),
 ('officially', ('d1', 1)),
 ('the', ('d1', 7)),
 ('people', ('d1', 2)),
 ('republic', ('d1', 3)),
 ('of', ('d1', 6)),
 ('is', ('d1', 5)),
 ('a', ('d1', 4)),
 ('sovereign', ('d1', 1)),
 ('state', ('d1', 2)),
 ('located', ('d1', 1)),
 ('in', ('d1', 2)),
 ('east', ('d1', 1)),
 ('asia', ('d1', 1)),
 ('it', ('d1', 1)),
 ('world', ('d1', 1)),
 ('most', ('d1', 1)),
 ('populous', ('d1', 1)),
 ('country', ('d1', 1)),
 ('single', ('d1', 1)),
```

In [35]:

```
new2 = new.groupByKey().mapValues(lambda x: list(x))
new2.collect()
# just basic grouping using the words as key
```

Out[35]:

```
[('china', [('d1', 5), ('d2', 1), ('d4', 11), ('d5', 1)]),
 ('officially', [('d1', 1), ('d2', 1), ('d5', 1)]),
 ('the',
  [('d1', 7),
   ('d2', 17),
   ('d3', 18),
   ('d4', 26),
   ('d5', 9),
   ('d6', 16),
   ('d7', 15)]),
 ('people', [('d1', 2), ('d3', 1)]),
 ('republic', [('d1', 3), ('d2', 1), ('d3', 1), ('d5', 1), ('d6', 1)]),
 ('of',
  [('d1', 6),
```

In [36]:
```
new3 = new2.sortByKey()
new3.collect()
```

Out[36]:
```
[('a',
  [('d1', 4),
   ('d2', 2),
   ('d3', 5),
   ('d4', 5),
   ('d5', 3),
   ('d6', 3),
   ('d7', 6)]),
 ('about', [('d4', 1)]),
 ('across', [('d5', 2)]),
 ('actors', [('d4', 1)]),
 ('addition', [('d2', 1)]),
 ('administrative', [('d1', 1)]),
 ('affect', [('d4', 1)]),
 ('affects', [('d4', 1)]),
 ('afghanistan', [('d4', 1)]),
 ('after', [('d3', 2), ('d7', 1)]),
 ('agenda', [('d7', 1)]),
 ('agree', [('d7', 1)]),
```

In [37]:
```
new4 = new3.map(lambda x: new_data(x)).cache()
new4.collect()
```

Out[37]:
```
['a#d1:4;d2:2;d3:5;d4:5;d5:3;d6:3;d7:6;',
 'about#d4:1;',
 'across#d5:2;',
 'actors#d4:1;',
 'addition#d2:1;',
 'administrative#d1:1;',
 'affect#d4:1;',
 'affects#d4:1;',
 'afghanistan#d4:1;',
 'after#d3:2;d7:1;',
 'agenda#d7:1;',
 'agree#d7:1;',
 'air#d7:1;',
 'aircraft#d7:1;',
 'airlines#d7:1;',
 'alaska#d5:1;d6:1;',
 'all#d3:1;',
```