Introduction of Task

We use the map-reudce framework of Hadoop to calculate Jaccard Similarity of two given text datasets, and output the text pairs which the Jaccard Similarity is greater than the given threshold.

- What's Jaccard Similarity ?

  Generally Speaking, given set S1 and set S2, the Jaccard Similiarity is calculated as follow:

  $$Jaccard(S1, S2) = \frac{S1 \cap S2}{S1 \cup S2}$$

  Specifically, for text similarity, we should use N-gram to get above sets. If you have learned language model, you must be familiar with n-gram, which uses a window of width N to divide the text into many strings of equal length (i.e., N) from left to right. In addition, in order to avoid the situation that the length of text is less than N, we can add N-1 `'#'` as prefix and N-1 `'$'` as suffix.

  For example, the 2-gram representations of text `Gorbachev` and `Gorbechyov` are as follows:

  $$Gorbachev: \{\#\mathbf{G}, \mathbf{Go}, \mathbf{or}, \mathbf{rb}, ba, ac, \mathbf{ch}, he, ev, \mathbf{v\$}\}$$
  $$Gorbechyov: \{\#\mathbf{G}, \mathbf{Go}, \mathbf{or}, \mathbf{rb}, be, ec, \mathbf{ch}, hy, yo, ov, \mathbf{v\$}\}$$
  $$Jaccard\ \ Similarity = \frac{6}{10 + 11 - 6} = \frac{2}{5} = 0.4$$

Task: Given two text datasets, if the Jaccard Similarity of text1 from one dataset and text2 from another dataset is greater than the given threshold θ , we output the <text1, text2> as result. And the results are case insensitive.

Input: text dataset R, text dataset S, similarity threshold θ and N

Output: $T = \{< r, s > | r \in R, s \in S, Jaccard(r, s) >= \theta\}$

PS. The result is like `<Gorbachev,Gorbechyov> 0.4` format (we also give the jaccard similarity value).

How to Run the Code

Make sure you have configured your Hadoop environment correctly, and the version of Hadoop we used is 2.6.0 and the version of JAVA is 1.8.0.

Then, you can run it by following command:

```
hadoop jar JacSimCalc.jar <text dataset 1> <text dataset 2> <threshold> <N> <output>
```

PS. Totally five parameters need to be given. The <text dataset 1> and <text dataset 2> are the datasets need to be joined. The threshold is θ and N is used in N-gram.

For instance, if we want to join the `author.txt` with itself after we have put it into `/input/` path on hdfs. And we set threshold as 0.3, N as 3 and output results into `/output` path. So the command is as following:

```
hadoop jar JacSimCalc.jar /intput/author.txt /input/author.txt 0.3 3 /output
```

Then you can see the results at `/output` path.

In addition, you can use following command to put `author.txt` into `/input/` path:

```
hdfs dfs -put author.txt /input
```

And you can use following command to get the results from `/output` path:

```
hdfs dfs -get /output/part* ./results.txt
```

That's all. Welcome to ask questions or give advices. Thanks !