Wikipedia reference inverted index[1]

Summarization pattern (Challenge 3_b)

For this exercise you have to prepare a report including:

- Your data collections comment describing what they represent.
- Your source codes: a file precising in a comment the compiling and execution commands.
- Propose a test battery and report execution measures.
- Compress everything and sent a .zip or a .tar to genoveva.vargas@gmail.com

1.1 Problem statement

Recall that the purpose of an inverted index is to allow fast full text searches, at a cost of increased processing when a document is added to the database. It is an index data structure storing a mapping from content, such as words or numbers, to its locations in a database file, or in a document or a set of documents. For example, given the following three documents:

- `T[0] = "it is what it is"`
- `T[1] = "what is it"`
- `T[2] = "it is a banana”`

The corresponding inverted index is the following:

- `"a":      {2}`
- `"banana": {2}`
- `"is":     {0, 1, 2}`
- `"it":     {0, 1, 2}`
- `"what":   {0, 1}`

A term search for the terms "`what`", "`is`" and "`it`" would give the set:

- `{0,1,2} ∩ {0,1,2} ∩ {0,1} = {0,1}`

A full inverted index for these documents is the following:

- `"a":      {(2, 2)}`
- `"banana": {(2, 3)}`
- `"is":     {(0, 1), (0, 4), (1, 1), (2, 1)}`
- `"it":     {(0, 0), (0, 3), (1, 2), (2, 0)}`
- `"what":   {(0, 2), (1, 0)}`

If we run a phrase search for "`what is it`" we get hits for all the words in both document 0 and 1. But the terms occur consecutively only in document 1.

Suppose we want to add StackOverflow links to each Wikipedia page that is referenced in a StackOverflow comment. The following example analyzes each comment in Stack - Overflow to find hyperlinks to Wikipedia. If there is one, the link is output with the comment ID to generate the inverted index. When it comes to the reduce phase, all the comment IDs that reference the same hyperlink will be grouped together. These groups

[1] This challenge is an example proposed in the book MapReduce design patterns, pp. 25.

are then concatenated together into a white space delimited String and directly output to the file system. From here, this data file can be used to update the Wikipedia page with all the comments that reference it.

Definition: Given a set of user's comments, build an inverted index of Wikipedia URLs to a set of answer post IDs .

1.2 To Do

Look at page 35 of the book "Map Reduce design patterns" and see the proposed `Map` and `Reduce` codes. Prepare a data collection from StackOverflow and Wikipedia.

- Prepare collections of different sizes to run your tests trying to get to the limits of your solution.
- Propose and implement a combiner that can optimize your solution. Explain the principle and use examples for supporting your arguments it.
- Make comparisons. Do not hesitate to prepare graphics.