

StackExchange

Carlos Hernandez, Samuel Foster, Joseph Tuazon



173 Q&A Communities.

We studied duplicate questions.

We focused on six communities:

- 1. Math - 1M questions**
- 2. Gaming - 84K questions**
- 3. Software Engineering - 51K questions**
- 4. Chemistry - 28K questions**
- 5. Cooking - 20K questions**
- 6. Anime & Manga - 10K questions**

(6.57 GB total PostHistory data)

Data

Data

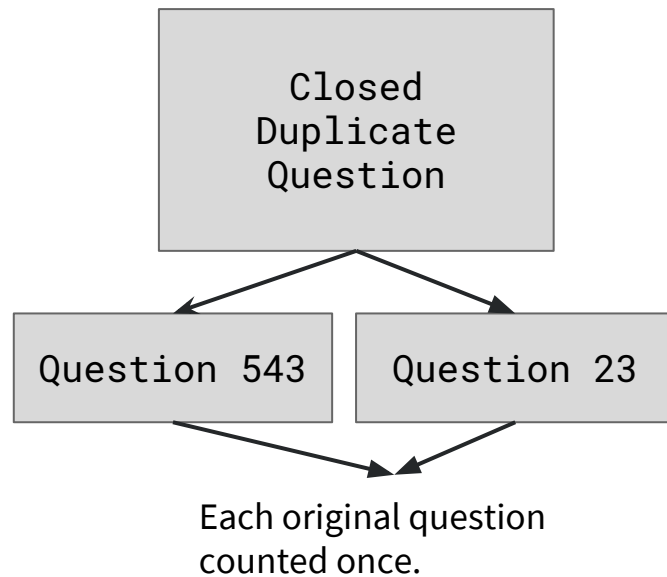
- Community data from <https://archive.org/details/stackexchange>
- Focused on **PostHistory.XML** of each
- If question is closed as duplicate, its type ID will be 10 and may list the question(s) it is a duplicate of
- Some posts are marked as duplicate, but do not point to any original question
- Trust is being placed in each community's moderators to accurately close and link to original questions

MapReduce

Mapping

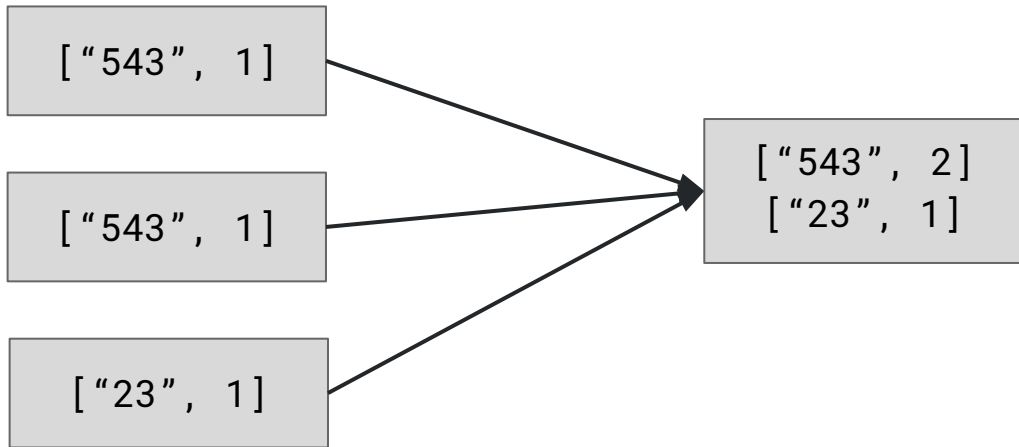
- Rows were read in from PostHistory.XML
- During Map process, we used a regular expression to grab the original Ids
 - Simple Row: <row PostHistoryTypeId="10" Text="{OriginalQuestionIds:[543, 23]}/>
 - Mapped to [543, 1] and [23, 1]

```
51 public static class DuplicateCountMapper extends Mapper<Object, Text, Text, IntWritable> {
52     private final static IntWritable one = new IntWritable(1);
53     private Text word = new Text();
54
55     public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
56         String text = value.toString();
57         if (text.contains("PostHistoryTypeId=\"10\"") && text.contains("OriginalQuestionIds")) {
58             text = text.split("[\\{\\}"])[1]; // Find the list of question Ids
59
60             for (String linkId : text.split(",")) {
61                 word.set(linkId.replaceAll("\\s", "")); // Remove whitespace
62                 context.write(word, one);
63             }
64         }
65     }
66 }
67 }
```



Reducing

- Similar IDs are combined
- Reduced to count of each ID
- Output pushed to file



```
/**
 * With the original Ids in place, we reduce them to count them. Output to
 * file.
 */
public static class DuplicateCountReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
        int count = 0;

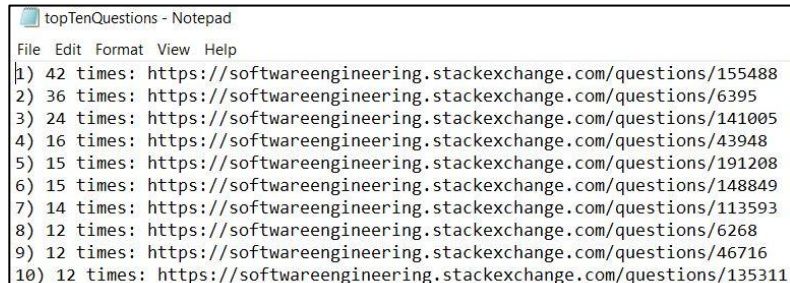
        for (IntWritable val : values) {
            count += val.get();
        }

        result.set(count);
        context.write(key, result);
    }
}
```


Filtering Output (Python)

- Sorted the IDs and appended to the URL of the respective community
- Printed the top ten to a file:

```
14 # Each line in the output
15 for line in doc:
16     # Find the ID and count as separate multi-digit groups
17     m = re.search('(\d*)\D*(\d*)', line)
18     if m:
19         # First group is ID (add link prefix)
20         id = linkPrefix + m.group(1)
21
22         # Second group is count
23         count = m.group(2)
24
25         # Add to list
26         idCountList.append([id, int(count)])
27
28 # Sort list in-place ascending
29 idCountList.sort(key=lambda tup: tup[1])
30
31 # Grab top 10 from end of list
32 topTenQuestions = idCountList[-10:]
33
34 # Put in descending order
35 topTenQuestions.reverse()
36
37 # Output the list of questions
38 index = 1
39
40 # Change this based on community being parse (SE, in this case)
41 outputFile = open("SE_topTenQuestions.txt", "w+")
42
43 for question in topTenQuestions:
44     outputFile.write(str(index) + ") " + str(question[1]) + " times: " + question[0] + "\n")
45     index = index + 1
```



topTenQuestions - Notepad

File Edit Format View Help

```
1) 42 times: https://softwareengineering.stackexchange.com/questions/155488
2) 36 times: https://softwareengineering.stackexchange.com/questions/6395
3) 24 times: https://softwareengineering.stackexchange.com/questions/141005
4) 16 times: https://softwareengineering.stackexchange.com/questions/43948
5) 15 times: https://softwareengineering.stackexchange.com/questions/191208
6) 15 times: https://softwareengineering.stackexchange.com/questions/148849
7) 14 times: https://softwareengineering.stackexchange.com/questions/113593
8) 12 times: https://softwareengineering.stackexchange.com/questions/6268
9) 12 times: https://softwareengineering.stackexchange.com/questions/46716
10) 12 times: https://softwareengineering.stackexchange.com/questions/135311
```

Procedure

Procedure

1. SSH into XSEDE Bridges
2. Start Hadoop container
3. Transfer MapReduce JAR → Local XSEDE filesystem
4. Transfer Community's PostHistory.XML → Local XSEDE filesystem
5. Create input directory in Hadoop filesystem
6. Transfer the PostHistory.XML to Hadoop filesystem
7. Call MapReduce JAR on PostHistory and set output folder
8. Copy output folder to XSEDE local filesystem
9. SFTP the output to true local filesystem
10. Run Python script on output file for filtering

XSEDE Bridges

- Software Eng. Community
- Splits: 5
- Map Time: 38.3 s
- Reduce Time: 18.2 s

```
18/12/10 14:25:36 INFO client.RMProxy: Connecting to ResourceManager at r426.opa.bridges.psc.edu/10.4.117.174:8032
18/12/10 14:25:36 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and
your application with ToolRunner to remedy this.
18/12/10 14:26:08 INFO input.FileInputFormat: Total input paths to process : 1
18/12/10 14:26:28 INFO mapreduce.JobSubmitter: number of splits:5
18/12/10 14:26:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1544469419530_0002
18/12/10 14:26:34 INFO impl.YarnClientImpl: Submitted application application_1544469419530_0002
18/12/10 14:26:34 INFO mapreduce.Job: The url to track the job: http://r426.opa.bridges.psc.edu:8088/proxy/application_1544469419530_0002/
18/12/10 14:26:45 INFO mapreduce.Job: job_1544469419530_0002
18/12/10 14:26:45 INFO mapreduce.Job: Job job_1544469419530_0002 running in uber mode : false
18/12/10 14:26:45 INFO mapreduce.Job: map 0% reduce 0%
18/12/10 14:26:55 INFO mapreduce.Job: map 100% reduce 0%
18/12/10 14:27:07 INFO mapreduce.Job: map 100% reduce 100%
18/12/10 14:27:50 INFO mapreduce.Job: Job job_1544469419530_0002 completed successfully
18/12/10 14:27:50 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=23135
    FILE: Number of bytes written=769875
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=597265003
    HDFS: Number of bytes written=13135
    HDFS: Number of read operations=18
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=5
    Launched reduce tasks=1
    Data-local map tasks=5
    Total time spent by all maps in occupied slots (ms)=6013100
    Total time spent by all reduces in occupied slots (ms)=2858028
    Total time spent by all map tasks (ms)=38300
    Total time spent by all reduce tasks (ms)=18204
    Total vcore-milliseconds taken by all map tasks=38300
    Total vcore-milliseconds taken by all reduce tasks=18204
    Total megabyte-milliseconds taken by all map tasks=191959600
    Total megabyte-milliseconds taken by all reduce tasks=91238448
  Map-Reduce Framework
    Map input records=534550
    Map output records=2496
    Map output bytes=26166
    Map output materialized bytes=23159
    Input split bytes=685
    Combine input records=2496
    Combine output records=1847
    Reduce input groups=1534
    Reduce shuffle bytes=23159
    Reduce input records=1847
    Reduce output records=1534
    Spilled Records=3694
    Shuffled Maps =5
    Failed Shuffles=0
    Merged Map outputs=5
    GC time elapsed (ms)=400
    CPU time spent (ms)=12220
    Physical memory (bytes) snapshot=4618141696
    Virtual memory (bytes) snapshot=38580457472
    Total committed heap usage (bytes)=9309257728
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=597264318
  File Output Format Counters
    Bytes Written=13135
```

Results



Anime & Manga

1. **What's the difference between the FMA and FMA Brotherhood series?** (5 times)
2. **Where is this picture from? How do I use Reverse Image Search to find the source of an anime/manga image?** (4 times)
3. **Who exactly is Archer from the Unlimited Blade Works movie?** (3 times)
4. **Does the Vampire Knight manga have additional story that the anime doesn't?** (3 times)
5. **Which episodes of the Naruto anime are core plot and which are filler?** (3 times)
6. **What are all the marriages and child relationships following the final chapter?** (2 times)
7. **Can I increase my lifespan by killing myself using the Death Note?** (2 times)
8. **Is it possible to kill people indirectly with the Death Note?** (2 times)
9. **Do we know how Senju Hashirama died?** (2 times)
10. **Can an animated show created outside Japan be called Anime?** (2 times)



Cooking

1. **How do I know if food left at room temperature is still safe to eat?** (123 times)
2. **How long can I store a food in the pantry, refrigerator, or freezer?** (85 times)
3. **Is it bad to leave the crock pot on “warm” (not low) all day?** (12 times)
4. **Rules for refreezing food** (10 times)
5. **Is there a problem with defrosting meat on the counter?** (9 times)
6. **How long can cooked food be safely stored at room/warm temperature?** (9 times)
7. **How dangerous is it to refreeze meat that has been thawed?** (9 times)
8. **Why is it dangerous to eat meat which has been left out and then cooked?** (7 times)
9. **How to fix food that got extra salty?** (6 times)
10. **What is a substitute for red or white wine in a recipe?** (6 times)



Chemistry

1. **Resources for learning Chemistry** (20 times)
2. **Why do elements in columns 6 and 11 assume 'abnormal' electron configurations?** (14 times)
3. **How can I predict if a reaction will occur between any two (or more) substances?** (9 times)
4. **Fundamental forces behind covalent bonding** (7 times)
5. **Is negative pH level physically possible?** (6 times)
6. **How do I figure out the hybridization of a particular atom in a molecule?** (6 times)
7. **What makes C=O more stable than C(OH)₂** (6 times)
8. **Why does the 3rd electron shell start filling up with scandium?** (5 times)
9. **Ortho-effect in substituted aromatic acids and bases** (5 times)
10. **Is pure water very corrosive?** (5 times)



Gaming

1. **Is there a list of error codes for Minecraft?** (38 times)
2. **How do I build a house for my NPCs?** (33 times)
3. **Why can't I destroy or place blocks?** (32 times)
4. **How do I find my follower if and when they leave me?** (32 times)
5. **I can't play Pokemon GO! What's wrong?** (29 times)
6. **Why isn't my Minecraft LAN server working?** (27 times)
7. **Minecraft crashes on launch with EXCEPTION_ACCESS_VIOLATION, Problematic frame: ig4dev32.dll or ig4dev64.dll or ig4icd32.dll or ig4icd64.dll** (22 times)
8. **How do I recover my old Clash of Clans (COC) game save & base?** (22 times)
9. **I'm stuck in a teleporting loop. What can I do?** (21 times)
10. **How can I tell how long (more or less) it will take me to complete a game?** (18 times)



Software Engineering

1. **I've inherited 200K lines of spaghetti code — what now?** (42 times)
2. **How do you dive into large code bases?** (36 times)
3. **How would you know if you've written readable and easily maintainable code?** (24 times)
4. **How can I convince management to deal with technical debt?** (16 times)
5. **Approaches to checking multiple conditions?** (15 times)
6. **Style for control flow with validation checks** (15 times)
7. **How can I tactfully suggest improvements to others' badly designed code during review?** (14 times)
8. **When is a BIG Rewrite the answer?** (12 times)
9. **What technical details should a programmer of a web application consider before making the site public?** (12 times)
10. **What is the most effective way to add functionality to unfamiliar, structurally unsound code?** (12 times)



Mathematics

1. How can I evaluate $\sum_{n=0}^{\infty} (n+1)x^n$? (102 times)
2. Why $\sqrt{-1 \times -1} \neq \sqrt{-1}^2$? (66 times)
3. **How do I compute $a^b \bmod c$ by hand?** (59 times)
4. **How do we sum up sin and cos series when the angles are in arithmetic progression?** (58 times)
5. Value of $\sum x^n$ (57 times)
6. **Why does $1 + 2 + 3 + \dots = -1/12$?** (54 times)
7. **Is it true that $0.999999999 \dots = 1$?** (54 times)
8. **Proving $1^3 + 2^3 + \dots + n^3 = (n(n+1)/2)^2$?** (49 times)
9. **If $AB = I$ then $BA = I$** (46 times)
10. **Expected time to roll all 1 through 6 on a die** (43 times)

Conclusion

StackExchange Duplicate Analysis

- Although setup required with XSEDE Bridges, overall runtime is low
- Not useful on tiny communities (or less actively moderated ones)
- **Provides useful insight on the questions being asked by a large community about a subject**
- **Output can influence new curriculum and better understanding**

Information & Data

- GitHub Repository: <https://github.com/FosterSamuel/cs4650-capstone>
- Output Files (Post-Parsing):
<https://github.com/FosterSamuel/cs4650-capstone/tree/master/outputAfterParsing>
- StackExchange Data: <https://archive.org/details/stackexchange>

Any questions?