Data preparation

Java code to implement preprocessing

```
// remove all invalid fieles
public void delete() throws FileNotFoundException {

      for(File file: dir.listFiles()){
            FileInputStream fis = new FileInputStream(file);
            int count=0;
            sc = new Scanner(fis);
            while(sc.hasNext()) {
                  sc.next();
                  count++;
            }
            if(count < 50) {
                  file.delete();
            }
      }
    }
```

```
// remove non-word characters
public void filt_out(int id) throws IOException {
      BufferedReader bin;

      for(File file: dir.listFiles()) {
            FileInputStream fis = new FileInputStream(file);
            bin = new BufferedReader(new InputStreamReader(fis));
            String line = null;
            StringBuffer sb = new StringBuffer();
            // add hw number to the beginning of the review
            sb.append(id + "\t");
            while((line = bin.readLine()) != null) {
                  line = line.replaceAll("\\W", " ");
                  sb.append(line);
            }

            sb.append("\n");

            try{
                  FileWriter fw = new FileWriter(file);
                  fw.write(sb.toString().toLowerCase());
                  fw.close();
            }
            catch(IOException e) {
                  e.printStackTrace();
            }
      }
      }
```

```
// take in review and remove stop words from it
public Text evaluate(String input) throws IOException {

      FileInputStream fstream = new FileInputStream(stopwords);
      BufferedReader br = new BufferedReader(new
InputStreamReader(fstream));

      //Read File Line By Line
      String line;
      while ((line = br.readLine()) != null) {
            stopWordsSet.add(line);
      }
      br.close();

      // add non stop words into list
      String[] words = input.split("\\s");
      for(String word : words)
         {
             if(!stopWordsSet.contains(word))
             {
                 wordsList.add(word + " ");
             }
         }

      result.set(wordsList.toString());
      return result;
    }
```

```
--use udf to remove stop words
hive> ADD JAR removestopwords.jar;
hive> CREATE TEMPORARY FUNCTION remove AS 'default1.Removestopwords';
hive> CREATE TABLE reviews (hw_num INT, review STRING)
    > AS
    > SELECT hw_num, remove(review) FROM reviews;
```

Result of preprocessing

(*take hw1_review_5.txt as example*)

Original review	**After preprocessing**
Overall, this homework assignment was easy and straightforward for me. There was good variety in the questions and I enjoyed reading the required article. The second problem was frustrating for a while because I arrived at a correct answer but didn't know it due to rounding error. It didn't take too long and was overall an enjoyable process	1 homework assignment easy straightforward good variety questions enjoyed reading required article problem frustrating arrived correct answer didn due rounding error didn long enjoyable process

Basic Analysis

homework 1 receives 255 positive words and 199 negative words based on 81 review texts
homework 2 receives 189 positive words and 167 negative words based on 74 review texts
homework 3 receives 197 positive words and 171 negative words based on 74 review texts
homework 4 receives 169 positive words and 234 negative words based on 80 review texts
homework 5 receives 207 positive words and 192 negative words based on 80 review texts
homework 6 receives 199 positive words and 217 negative words based on 75 review texts
homework 7 receives 213 positive words and 248 negative words based on 74 review texts
homework 8 receives 166 positive words and 254 negative words based on 74 review texts

Overall, homework 1 was perceived most positive on average across all students, and homework 8 was perceived most negative on average across all students.

positive words:
good 142
work 139
interesting 105
easy 90
pretty 81

negative words:
problem 350
pig 166
problems 137
hard 121
difficult 110

Most of these results make sense. However, there are some words that do not make sense. First: pig. Because we learned pig as an academic term. Most students used pig in the review of homework is to express the meaning of pig as academic term instead of a negative word.

Also, we get the result for negative word "problem(s)" which occurs 350 + 137 times. This also is a little misunderstanding because some of the reviews used the word "problem" or "problems" to show that which question is discussing under the review text file.

To fix those problems, we need to combine the result of n-grams to do further analysis.

N-grams

Approach:

- Load all homework reviews into table reviews
- Do 2-grams and 3-grams over the entire homework reviews
- Retrieve the 3 most positive and negative 2-grams and 3-grams by comparing with positive.txt and negative.txt

Code script

```
$hive> SET n=2;
$hive> SELECT EXPLODE(NGRAMS(SENTENCES(LOWER(review)),${hiveconf:n},25))
       AS bigrams
       FROM reviews;
$hive> SET n=3;
```

```
$hive> SELECT EXPLODE(NGRAMS(SENTENCES(LOWER(review)),${hiveconf:n},40))
       AS bigrams
       FROM reviews;
```

Result of top 25 2-grams of all hw reviews:


```
{"ngram":["ha","ha"],"estfrequency":305.0}
{"ngram":["problem","2"],"estfrequency":62.0}
{"ngram":["lot","time"],"estfrequency":54.0}
{"ngram":["problem","3"],"estfrequency":50.0}
{"ngram":["problem","1"],"estfrequency":49.0}
{"ngram":["homework","assignment"],"estfrequency":47.0}
{"ngram":["long","time"],"estfrequency":39.0}
{"ngram":["command","line"],"estfrequency":36.0}
{"ngram":["big","data"],"estfrequency":33.0}
{"ngram":["learned","lot"],"estfrequency":31.0}
{"ngram":["map","reduce"],"estfrequency":30.0}
{"ngram":["question","1"],"estfrequency":26.0}
{"ngram":["time","finish"],"estfrequency":26.0}
{"ngram":["homework","good"],"estfrequency":26.0}
{"ngram":["homework","pretty"],"estfrequency":24.0}
{"ngram":["mapper","reducer"],"estfrequency":24.0}
{"ngram":["time","consuming"],"estfrequency":23.0}
{"ngram":["thought","homework"],"estfrequency":22.0}
{"ngram":["word","count"],"estfrequency":20.0}
{"ngram":["real","world"],"estfrequency":19.0}
{"ngram":["spring","break"],"estfrequency":17.0}
{"ngram":["homework","homework"],"estfrequency":17.0}
{"ngram":["la","la"],"estfrequency":17.0}
{"ngram":["spend","time"],"estfrequency":16.0}
{"ngram":["time","homework"],"estfrequency":16.0}
{"ngram":["mapreduce","program"],"estfrequency":16.0}
{"ngram":["found","homework"],"estfrequency":16.0}
{"ngram":["distributed","cache"],"estfrequency":15.0}
{"ngram":["homework","hard"],"estfrequency":15.0}
{"ngram":["learned","class"],"estfrequency":15.0}
{"ngram":["homework","lot"],"estfrequency":15.0}
{"ngram":["pig","latin"],"estfrequency":15.0}
{"ngram":["homework","difficult"],"estfrequency":15.0}
{"ngram":["good","homework"],"estfrequency":15.0}
{"ngram":["pretty","easy"],"estfrequency":14.0}
```

Result of top 40 3-grams of all hw reviews:

{"ngram":["ha","ha","ha"],"estfrequency":304.0}
{"ngram":["la","la","la"],"estfrequency":15.0}
{"ngram":["lot","time","finish"],"estfrequency":12.0}
{"ngram":["spent","lot","time"],"estfrequency":9.0}
{"ngram":["assignment","hard","finish"],"estfrequency":8.0}
{"ngram":["homework","problem","1"],"estfrequency":7.0}
{"ngram":["positive","introduction","actual"],"estfrequency":7.0}
{"ngram":["homework","time","consuming"],"estfrequency":7.0}
{"ngram":["iwill","spend","time"],"estfrequency":7.0}
{"ngram":["sessionmakes","sense","process"],"estfrequency":7.0}
{"ngram":["application","exciting","apply"],"estfrequency":6.0}
{"ngram":["problem","2","bit"],"estfrequency":6.0}
{"ngram":["slides","lab","review"],"estfrequency":6.0}
{"ngram":["lab","review","sessionmakes"],"estfrequency":6.0}
{"ngram":["homework","assignment","hard"],"estfrequency":6.0}
{"ngram":["finishedreviewing","information","book"],"estfrequency":6.0}
{"ngram":["review","sessionmakes","sense"],"estfrequency":6.0}
{"ngram":["driver","mapper","reducer"],"estfrequency":6.0}
{"ngram":["finish","question","1"],"estfrequency":6.0}
{"ngram":["hard","finish","finishedreviewing"],"estfrequency":6.0}
{"ngram":["hope","iwill","spend"],"estfrequency":6.0}
{"ngram":["finish","finishedreviewing","information"],"estfrequency":6.0}
{"ngram":["book","slides","lab"],"estfrequency":6.0}
{"ngram":["blood","blood","blood"],"estfrequency":6.0}
{"ngram":["information","book","slides"],"estfrequency":6.0}
{"ngram":["problem","1","difficult"],"estfrequency":6.0}
{"ngram":["thought","research","problem"],"estfrequency":6.0}
{"ngram":["long","time","finish"],"estfrequency":6.0}
{"ngram":["long","time","complete"],"estfrequency":6.0}
{"ngram":["thelab","lot","time"],"estfrequency":6.0}
{"ngram":["homework","learned","lot"],"estfrequency":6.0}
{"ngram":["3","hours","finish"],"estfrequency":5.0}
{"ngram":["assignment","spent","approximately"],"estfrequency":5.0}
{"ngram":["homework","good","job"],"estfrequency":5.0}
{"ngram":["average","word","length"],"estfrequency":5.0}
{"ngram":["learnt","corporate","partner"],"estfrequency":5.0}
{"ngram":["time","search","internet"],"estfrequency":5.0}
{"ngram":["hadoop","file","system"],"estfrequency":5.0}
{"ngram":["hours","finish","question"],"estfrequency":5.0}
{"ngram":["introduction","actual","usage"],"estfrequency":4.0}
{"ngram":["homework","positive","introduction"],"estfrequency":4.0}

There are some bias in the results, for instance, ["ha","ha"], ["ha","ha","ha"] and ["la","la","la"]， ["blood","blood","blood"], etc. There 2-grams or 3-grams do not make any sense. So, we ignore them, and search other grams according to the positive and negative txt file.

From the results we find that the 3 most positive 2-grams are

```
{"ngram":["homework","good"],"estfrequency":26.0},
{"ngram":["homework","pretty"],"estfrequency":24.0},
{"ngram":["pretty","easy"],"estfrequency":14.0}.
```

The 3 most positive 3-grams:

```
{"ngram":["application","exciting","apply"],"estfrequency":6.0}
{"ngram":["homework","good","job"],"estfrequency":5.0}
{"ngram":["homework","positive","introduction"],"estfrequency":4.0}
```

By comparing to the positive and negative file, we finally find the 3 most positive and negative 2-grams and 3-grams as above. These 2-grams and 3-grams make sense.

The Favorite Topics

Approaches:

- Create 2 tables to store the first half (hw1-hw6) reviews and the second half (hw7-hw8) reviews. Create 2 other tables to store hw1-hw2 reviews and hw3-hw6 reviews.
- Load data into the corresponding tables
- Do N-grams over records stored in these tables.
- Join the n-gram results with positive.txt and negative.txt. Then counter the number of records.

Hive Script

```
CREATE TABLE one_six(
hw_num INT,
review STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t';

$hive> SET n=2;
$hive> SELECT EXPLODE(NGRAMS(SENTENCES(LOWER(review)),${hiveconf:n},25))
        AS bigrams
        FROM reviews;
```

Results of 2-grmas and 3-grams

Among the top 100 2-grams of hw1 to hw6, the positive 2-grams are as following	**Among the top 100 2-grams of hw7 to hw8, the positive 2-grams are as following**
{"ngram":["homework","good"],"estfrequency":22.0} {"ngram":["good","job"],"estfrequency":10.0} {"ngram":["homework","easier"],"estfrequency":10.0} {"ngram":["good","homework"],"estfrequency":10.0} {"ngram":["homework","interesting"],"estfrequency":9.0} {"ngram":["pretty","easy"],"estfrequency":9.0} {"ngram":["enjoyed","homework"],"estfrequency":8.0} {"ngram":["bonferroni","principle"],"estfrequency":6.0} {"ngram":["good","practice"],"estfrequency":6.0}	{"ngram":["pretty","easy"],"estfrequency":5.0} {"ngram":["homework","easy"],"estfrequency":5.0} {"ngram":["good","homework"],"estfrequency":5.0} {"ngram":["homework","pretty"],"estfrequency":4.0} {"ngram":["homework","good"],"estfrequency":4.0} {"ngram":["interesting","homework"],"estfrequency":3.0} {"ngram":["enjoyed","homework"],"estfrequency":3.0} {"ngram":["easy","understand"],"estfrequency":3.0} {"ngram":["good","learn"],"estfrequency":3.0} {"ngram":["enjoyed","working"],"estfrequency":3.0} {"ngram":["exciting","apply"],"estfrequency":2.0}

Among the top 100 3-grams of hw1 to hw6, the positive 3-grams are as following	**Among the top 100 3-grams of hw7 to hw8, the positive 3-grams are as following**
{"ngram":["homework","good","job"],"estfrequency":5.0} {"ngram":["positive","introduction","actual"],"estfrequency":5.0} {"ngram":["application","exciting","apply"],"estfrequency":4.0} {"ngram":["homework","assignment","good"],"estfrequency":4.0} {"ngram":["exciting","apply","mapreduce"],"estfrequency":4.0} {"ngram":["hovering","compareto","clearer"],"estfrequency":2.0} {"ngram":["interesting","academic","discussion"],"estfrequency":2.0}	{"ngram":["good","homework","time"],"estfrequency":2.0} {"ngram":["internet","watch","pig"],"estfrequency":2.0} {"ngram":["good","learn","pig"],"estfrequency":2.0} {"ngram":["important","techniques","mapreduce"],"estfrequency":2.0} {"ngram":["positive","introduction","actual"],"estfrequency":2.0} {"ngram":["pig","slightly","advanced"],"estfrequency":2.0} {"ngram":["practical","feel","good"],"estfrequency":2.0}

{"ngram":["feels","good","points"],"estfrequency":2.0} {"ngram":["understanding","prefer","true"],"estfrequency":2.0} {"ngram":["reasonable","step","step"],"estfrequency":2.0} {"ngram":["feel","satisfied","pleased"],"estfrequency":2.0} {"ngram":["configuration","convenient","lot"],"estfrequency":2.0}	{"ngram":["feel","satisfied","pleased"],"estfrequency":2.0}

From the results, we can find that the frequency of positive 2-grams in hw1 to hw6 is more than that in hw7 to hw8, which means hw1 to hw6 are easier or more positive than hw7 to hw8.

Among the top 50 2-grams of hw1 to hw2, the positive 2-grams are as following	**Among the top 50 2-grams of hw3 to hw6, the positive 2-grams are as following**
{"ngram":["homework","good"],"estfrequency":11.0} {"ngram":["homework","pretty"],"estfrequency":8.0} {"ngram":["bonferroni","principle"],"estfrequency":6.0} {"ngram":["good","job"],"estfrequency":6.0} {"ngram":["homework","interesting"],"estfrequency":5.0} {"ngram":["homework","straightforward"],"estfrequency":5.0} {"ngram":["good","homework"],"estfrequency":5.0} {"ngram":["easy","homework"],"estfrequency":4.0} {"ngram":["pretty","easy"],"estfrequency":4.0}	{"ngram":["homework","pretty"],"estfrequency":12.0} {"ngram":["homework","good"],"estfrequency":11.0} {"ngram":["homework","easier"],"estfrequency":9.0} {"ngram":["homework","fun"],"estfrequency":6.0}

Among the top 50 3-grams of hw1 to hw2, the positive 3-grams are as following	**Among the top 50 3-grams of hw3 to hw6, the positive 3-grams are as following**
{"ngram":["homework","good","job"],"estfrequency":4.0} {"ngram":["problem","1","open"],"estfrequency":3.0} {"ngram":["optimized","answer","wrote"],"estfrequency":2.0} {"ngram":["made","interesting","academic"],"estfrequency":2.0} {"ngram":["homework","assignment","good"],"estfrequency":2.0}	{"ngram":["problem","2","good"],"estfrequency":4.0} {"ngram":["positive","introduction","actual"],"estfrequency":4.0} {"ngram":["good","job","teaching"],"estfrequency":3.0} {"ngram":["homework","good","practice"],"estfrequency":3.0}

{"ngram":["big","data","important"],"estfrequency":2.0} {"ngram":["classmate","feels","good"],"estfrequency":2.0} {"ngram":["thought","homework","good"],"estfrequency":2.0} {"ngram":["found","homework","interesting"],"estfrequency":2.0} {"ngram":["model","made","interesting"],"estfrequency":2.0} {"ngram":["true","problem","easily"],"estfrequency":2.0} {"ngram":["interesting","academic","discussion"],"estfrequency":2.0} {"ngram":["feels","good","points"],"estfrequency":2.0} {"ngram":["hints","kind","confusion"],"estfrequency":2.0} {"ngram":["understanding","prefer","true"],"estfrequency":2.0}	

From the results, we can find that the frequency of positive 3-grams in hw1 to hw2 is more than that in hw3 to hw6, which means hw1 to hw6 are easier or more positive than hw3 to hw6. And hw3 to hw6 is better than hw7 to hw8.

We run the program over reviews of our team members. The following are the result of 2-grams
{"ngram":["homework","easier"],"estfrequency":2.0}
{"ngram":["pretty","easy"],"estfrequency":2.0}
We found that the first half of homeworks (hw1 to hw 4) are much easier than the following homework.