

```
1 import components.set.Set;
2 import components.set.Set1L;
3 import components.simplereader.SimpleReader;
4 import components.simplereader.SimpleReader1L;
5 import components.simplewriter.SimpleWriter;
6 import components.simplewriter.SimpleWriter1L;
7
8 /**
9  * Utility class to support string reassembly from fragments.
10  *
11  * @author Isaac Frank
12  *
13  * @mathdefinitions <pre>
14  *
15  * OVERLAPS (
16  *   s1: string of character,
17  *   s2: string of character,
18  *   k: integer
19  * ) : boolean is
20  *  $0 \leq k$  and  $k \leq |s1|$  and  $k \leq |s2|$  and
21  *  $s1[|s1|-k, |s1|) = s2[0, k)$ 
22  *
23  * SUBSTRINGS (
24  *   strSet: finite set of string of character,
25  *   s: string of character
26  * ) : finite set of string of character is
27  * {t: string of character
28  *   where (t is in strSet and t is substring of s)
29  *   (t)}
30  *
31  * SUPERSTRINGS (
32  *   strSet: finite set of string of character,
33  *   s: string of character
34  * ) : finite set of string of character is
35  * {t: string of character
36  *   where (t is in strSet and s is substring of t)
37  *   (t)}
38  *
39  * CONTAINS_NO_SUBSTRING_PAIRS (
40  *   strSet: finite set of string of character
41  * ) : boolean is
42  * for all t: string of character
43  *   where (t is in strSet)
44  *   (SUBSTRINGS(strSet \ {t}, t) = {})
```

```

45 *
46 * ALL_SUPERSTRINGS (
47 *   strSet: finite set of string of character
48 * ) : set of string of character is
49 * {t: string of character
50 *   where (SUBSTRINGS(strSet, t) = strSet)
51 *   (t)}
52 *
53 * CONTAINS_NO_OVERLAPPING_PAIRS (
54 *   strSet: finite set of string of character
55 * ) : boolean is
56 * for all t1, t2: string of character, k: integer
57 *   where (t1 != t2 and t1 is in strSet and t2 is in strSet
and
58 *         1 <= k and k <= |s1| and k <= |s2|)
59 *   (not OVERLAPS(s1, s2, k))
60 *
61 * </pre>
62 */
63 public final class StringReassembly {
64
65     /**
66      * Private no-argument constructor to prevent instantiation of
this utility
67      * class.
68      */
69     private StringReassembly() {
70     }
71
72     /**
73      * Reports the maximum length of a common suffix of {@code
str1} and prefix
74      * of {@code str2}.
75      *
76      * @param str1
77      *           first string
78      * @param str2
79      *           second string
80      * @return maximum overlap between right end of {@code str1}
and left end of
81      *           {@code str2}
82      * @requires <pre>
83      *   str1 is not substring of str2 and
84      *   str2 is not substring of str1

```

```

85     * </pre>
86     * @ensures <pre>
87     * OVERLAPS(str1, str2, overlap) and
88     * for all k: integer
89     *     where (overlap < k and k <= |str1| and k <= |str2|)
90     * (not OVERLAPS(str1, str2, k))
91     * </pre>
92     */
93     public static int overlap(String str1, String str2) {
94         assert str1 != null : "Violation of: str1 is not null";
95         assert str2 != null : "Violation of: str2 is not null";
96         /*
97          * Start with maximum possible overlap and work down until
a match is
98          * found; think about it and try it on some examples to
see why
99          * iterating in the other direction doesn't work
100         */
101         int maxOverlap = str2.length() - 1;
102         while (!str1.regionMatches(str1.length() - maxOverlap,
str2, 0,
103                                     maxOverlap)) {
104             maxOverlap--;
105         }
106         return maxOverlap;
107     }
108
109     /**
110      * Returns concatenation of {@code str1} and {@code str2} from
which one of
111      * the two "copies" of the common string of {@code overlap}
characters at
112      * the end of {@code str1} and the beginning of {@code str2}
has been
113      * removed.
114      *
115      * @param str1
116      *     first string
117      * @param str2
118      *     second string
119      * @param overlap
120      *     amount of overlap
121      * @return combination with one "copy" of overlap removed
122      * @requires OVERLAPS(str1, str2, overlap)

```

```

123     * @ensures combination = str1[0, |str1|-overlap) * str2
124     */
125     public static String combination String str1, String str2, int
overlap) {
126         assert str1 != null : "Violation of: str1 is not null";
127         assert str2 != null : "Violation of: str2 is not null";
128         assert 0 <= overlap && overlap <= str1.length()
129             && overlap <= str2.length()
130             && str1.regionMatches(str1.length() - overlap,
str2, 0,
131                                     overlap) : ""
132         + "Violation of: OVERLAPS(str1,
str2, overlap)";
133
134         // return the combination of str1 and str2
135         return str1 + str2.substring(overlap);
136     }
137
138     /**
139     * Adds {@code str} to {@code strSet} if and only if it is not
a substring
140     * of any string already in {@code strSet}; and if it is
added, also removes
141     * from {@code strSet} any string already in {@code strSet}
that is a
142     * substring of {@code str}.
143     *
144     * @param strSet
145     *         set to consider adding to
146     * @param str
147     *         string to consider adding
148     * @updates strSet
149     * @requires CONTAINS_NO_SUBSTRING_PAIRS(strSet)
150     * @ensures <pre>
151     * if SUPERSTRINGS(#strSet, str) = {}
152     * then strSet = #strSet union {str} \ SUBSTRINGS(#strSet,
str)
153     * else strSet = #strSet
154     * </pre>
155     */
156     public static void addToSetAvoidingSubstrings(Set<String>
strSet,
157                                     String str) {
158         assert strSet != null : "Violation of: strSet is not

```

```
    null";
159     assert str != null : "Violation of: str is not null";
160     /*
161      * Note: Precondition not checked!
162      */
163
164     // initializing vars
165     boolean addStr = true;
166     String strToRemove = new String();
167
168     // iterating through strSet
169     for (String x : strSet) {
170         if (addStr) {
171             // if str is a substring of any element in strSet,
do not add str
172             if (x.contains(str)) {
173                 addStr = false;
174                 // if x is substring of str, remove x and
later add str
175             } else if (!str.equals(x) && str.contains(x)) {
176                 strToRemove = x;
177             }
178         }
179     }
180
181     // modifying str based on info found in the loop
182     if (!strToRemove.equals("")) {
183         strSet.remove(strToRemove);
184     }
185     if (addStr) {
186         strSet.add(str);
187     }
188 }
189
190 /**
191  * Returns the set of all individual lines read from {@code
input}, except
192  * that any line that is a substring of another is not in the
returned set.
193  *
194  * @param input
195  *         source of strings, one per line
196  * @return set of lines read from {@code input}
197  * @requires input.is_open
```

```
198     * @ensures <pre>
199     * input.is_open and input.content = <> and
200     * linesFromInput = [maximal set of lines from #input.content
    such that
201     *
    CONTAINS_NO_SUBSTRING_PAIRS(linesFromInput)]
202     * </pre>
203     */
204     public static Set<String> linesFromInput(SimpleReader input) {
205         assert input != null : "Violation of: input is not null";
206         assert input.isOpen() : "Violation of: input.is_open";
207
208         // Empty set to add to
209         Set<String> lines = new Set1L<>();
210
211         // Iterate through until input is at the end
212         while (!input.atEOS()) {
213             addToSetAvoidingSubstrings(lines, input.nextLine());
214         }
215
216         return lines;
217     }
218
219     /**
220     * Returns the longest overlap between the suffix of one
    string and the
221     * prefix of another string in {@code strSet}, and identifies
    the two
222     * strings that achieve that overlap.
223     *
224     * @param strSet
225     *         the set of strings examined
226     * @param bestTwo
227     *         an array containing (upon return) the two
    strings with the
228     *         largest such overlap between the suffix of
    {@code bestTwo[0]}
229     *         and the prefix of {@code bestTwo[1]}
230     * @return the amount of overlap between those two strings
231     * @replaces bestTwo[0], bestTwo[1]
232     * @requires <pre>
233     * CONTAINS_NO_SUBSTRING_PAIRS(strSet) and
234     * bestTwo.length >= 2
235     * </pre>
```

```
236     * @ensures <pre>
237     * bestTwo[0] is in strSet and
238     * bestTwo[1] is in strSet and
239     * OVERLAPS(bestTwo[0], bestTwo[1], bestOverlap) and
240     * for all str1, str2: string of character, overlap: integer
241     *     where (str1 is in strSet and str2 is in strSet and
242     *             OVERLAPS(str1, str2, overlap))
243     *     (overlap <= bestOverlap)
244     * </pre>
245     */
246     private static int bestOverlap(Set<String> strSet, String[]
bestTwo) {
247         assert strSet != null : "Violation of: strSet is not
null";
248         assert bestTwo != null : "Violation of: bestTwo is not
null";
249         assert bestTwo.length >= 2 : "Violation of: bestTwo.length
>= 2";
250         /*
251         * Note: Rest of precondition not checked!
252         */
253         int bestOverlap = 0;
254         Set<String> processed = strSet.newInstance();
255         while (strSet.size() > 0) {
256             /*
257             * Remove one string from strSet to check against all
others
258             */
259             String str0 = strSet.removeAny();
260             for (String str1 : strSet) {
261                 /*
262                 * Check str0 and str1 for overlap first in one
order...
263                 */
264                 int overlapFrom0To1 = overlap(str0, str1);
265                 if (overlapFrom0To1 > bestOverlap) {
266                     /*
267                     * Update best overlap found so far, and the
two strings
268                     * that produced it
269                     */
270                     bestOverlap = overlapFrom0To1;
271                     bestTwo[0] = str0;
272                     bestTwo[1] = str1;
```

```
273     }
274     /*
275     * ... and then in the other order
276     */
277     int overlapFrom1To0 = overlap(str1, str0);
278     if (overlapFrom1To0 > bestOverlap) {
279         /*
280         * Update best overlap found so far, and the
281         two strings
282         * that produced it
283         */
284         bestOverlap = overlapFrom1To0;
285         bestTwo[0] = str1;
286         bestTwo[1] = str0;
287     }
288     /*
289     * Record that str0 has been checked against every
290     other string in
291     * strSet
292     */
293     processed.add(str0);
294     /*
295     * Restore strSet and return best overlap
296     */
297     strSet.transferFrom(processed);
298     return bestOverlap;
299 }
300
301 /**
302  * Combines strings in {@code strSet} as much as possible,
303  leaving in it
304  * only strings that have no overlap between a suffix of one
305  string and a
306  * prefix of another. Note: uses a "greedy approach" to
307  assembly, hence may
308  * not result in {@code strSet} being as small a set as
309  possible at the end.
310  *
311  * @param strSet
312  *        set of strings
313  * @updates strSet
314  * @requires CONTAINS_NO_SUBSTRING_PAIRS(strSet)
```



```

311     * @ensures <pre>
312     * ALL_SUPERSTRINGS(strSet) is subset of
    ALL_SUPERSTRINGS(#strSet) and
313     * |strSet| <= |#strSet| and
314     * CONTAINS_NO_SUBSTRING_PAIRS(strSet) and
315     * CONTAINS_NO_OVERLAPPING_PAIRS(strSet)
316     * </pre>
317     */
318     public static void assemble(Set<String> strSet) {
319         assert strSet != null : "Violation of: strSet is not
    null";
320         /*
321         * Note: Precondition not checked!
322         */
323         /*
324         * Combine strings as much possible, being greedy
325         */
326         boolean done = false;
327         while ((strSet.size() > 1) && !done) {
328             String[] bestTwo = new String[2];
329             int bestOverlap = bestOverlap(strSet, bestTwo);
330             if (bestOverlap == 0) {
331                 /*
332                 * No overlapping strings remain; can't do any
    more
333                 */
334                 done = true;
335             } else {
336                 /*
337                 * Replace the two most-overlapping strings with
    their
338                 * combination; this can be done with add rather
    than
339                 * addToSetAvoidingSubstrings because the latter
    would do the
340                 * same thing (this claim requires justification)
341                 */
342                 strSet.remove(bestTwo[0]);
343                 strSet.remove(bestTwo[1]);
344                 String overlapped = combination(bestTwo[0],
    bestTwo[1],
345                     bestOverlap);
346                 strSet.add(overlapped);
347             }

```

```
348     }
349 }
350
351 /**
352  * Prints the string {@code text} to {@code out}, replacing
    each '~' with a
353  * line separator.
354  *
355  * @param text
356  *         string to be output
357  * @param out
358  *         output stream
359  * @updates out
360  * @requires out.is_open
361  * @ensures <pre>
362  * out.is_open  and
363  * out.content = #out.content *
364  *   [text with each '~' replaced by line separator]
365  * </pre>
366  */
367 public static void printWithLineSeparators(String text,
SimpleWriter out) {
368     assert text != null : "Violation of: text is not null";
369     assert out != null : "Violation of: out is not null";
370     assert out.isOpen() : "Violation of: out.is_open";
371
372     // Iterates through, replacing all ~ with new lines
373     for (int i = 0; i < text.length(); i++) {
374         if (text.charAt(i) == '~') {
375             out.println();
376         } else {
377             out.print(text.charAt(i));
378         }
379     }
380 }
381
382 /**
383  * Given a file name (relative to the path where the
    application is running)
384  * that contains fragments of a single original source text,
    one fragment
385  * per line, outputs to stdout the result of trying to
    reassemble the
386  * original text from those fragments using a "greedy
```

```
    assembler". The
387     * result, if reassembly is complete, might be the original
    text; but this
388     * might not happen because a greedy assembler can make a
    mistake and end up
389     * predicting the fragments were from a string other than the
    true original
390     * source text. It can also end up with two or more fragments
    that are
391     * mutually non-overlapping, in which case it outputs the
    remaining
392     * fragments, appropriately labelled.
393     *
394     * @param args
395     *         Command-line arguments: not used
396     */
397     public static void main(String[] args) {
398         SimpleReader in = new SimpleReader1L();
399         SimpleWriter out = new SimpleWriter1L();
400         /*
401          * Get input file name
402          */
403         out.print("Input file (with fragments): ");
404         String inputFileName = in.nextLine();
405         SimpleReader inFile = new SimpleReader1L(inputFileName);
406         /*
407          * Get initial fragments from input file
408          */
409         Set<String> fragments = linesFromInput(inFile);
410         /*
411          * Close inFile; we're done with it
412          */
413         inFile.close();
414         /*
415          * Assemble fragments as far as possible
416          */
417         assemble(fragments);
418         /*
419          * Output fully assembled text or remaining fragments
420          */
421         if (fragments.size() == 1) {
422             out.println();
423             String text = fragments.removeAny();
424             printWithLineSeparators(text, out);

```

```
425     } else {
426         int fragmentNumber = 0;
427         for (String str : fragments) {
428             fragmentNumber++;
429             out.println();
430             out.println("-----");
431             out.println("  -- Fragment #" + fragmentNumber +
":  --");
432             out.println("-----");
433             printWithLineSeparators(str, out);
434         }
435     }
436     /*
437     * Close input and output streams
438     */
439     in.close();
440     out.close();
441 }
442
443 }
444
```