Saxon - Bug #3902

Regex non-greedy matching failure

2018-09-12 15:52 - Michael Kay

Status:	Closed	**Start date:**	2018-09-12
Priority:	Normal	**Due date:**	
Assignee:	Michael Kay	**% Done:**	100%
Category:	XSLT conformance	**Estimated time:**	0:00 hour
Sprint/Milestone:		**Spent time:**	0:00 hour
Legacy ID:		**Fix Committed on Branch:**	9.8, trunk
Applies to branch:	9.8, trunk	**Fixed in Maintenance Release:**	9.9.0.1, 9.8.0.15

Description

See https://saxonica.plan.io/boards/3/topics/7300

Made into XSLT3 test case analyze-string-099

History

#1 - 2018-09-12 16:07 - Michael Kay

The path taken by the non-greedy code for OpRepeat at Operation#884 is quite bewildering. Can't see the logic to it at all. Will need to study how this works in some examples where we actually get the right answer.

Note also in passing, the optimization that recognizes that a match must start with a particular character and fast-forwards to that character (at REMatcher#439) doesn't seem to be working as well as it should - although it recognizes that there's a fixed prefix it still seems to look for a match at every character position.

#2 - 2018-09-12 16:38 - Michael Kay

Not many tests exercise this path. In fact, only the following six tests do so:

In XSLT3: regex-syntax-0861, -0955

In QT3: fn-matches-50, re00974, -5, -6

And all of these are concerned with giving a boolean match/no-match result, not in extracting the matching substring.

#3 - 2018-09-12 17:47 - Michael Kay

Concerning the easier problem of the poorly-performing prefix comparison, the issue is that when there is a mismatch on the last character of the prefix (which in this case is the only character) we're taking the path as if the prefix matched, which negates the optimization in the case where the prefix has length 1.

Fixed this on the 9.9 branch.

#4 - 2018-09-13 10:36 - Michael Kay

Bug #3787 is in the same general area. In analyzing that bug, although I found an empirical resolution, I also remarked that the logic behind the existing code was impenetrable. It doesn't appear that I added a new test reflecting bug #3787 and it might be worth doing so.

#5 - 2018-09-13 15:15 - Michael Kay

I have replaced the impenetrable code (for non-greedy OpRepeat with a variable length match) with a greatly simplified version. This is working for the test case in question, but it doesn't yet handle min/max cardinality.

With this simple implementation, one of the tests in fn-matches-50 (the Perl regex test suite) fails with infinite backtracking: the regex in question is matches('abcde', '(?:r?)*?r|(.{2,4})') (a classic one that allows an infinite number of zero-length matches).

Solved this the traditional way of adding a ForceProgressIterator.

With this change all tests are passing. So given that we haven't implemented min and max, we clearly need more tests!

#6 - 2018-09-13 17:52 - Michael Kay

- Status changed from New to Resolved

- Fix Committed on Branch 9.8, trunk added

I have added a number of tests to QT3 fn-matches to handle non-greedy matching with min/max constraints.

All XSLT3 / QT3 tests now passing.

#7 - 2018-09-27 16:57 - O'Neil Delpratt

- % Done changed from 0 to 100

- Fixed in Maintenance Release 9.9.0.1 added

Bug fix applied in the Saxon 9.9.0.1 major release.

Leave open until fix applied in the next Saxon 9.8 maintenance release.

#8 - 2018-11-06 18:28 - O'Neil Delpratt

- Status changed from Resolved to Closed

- Fixed in Maintenance Release 9.8.0.15 added

Bug fix applied in the Saxon 9.8.0.15 maintenance release.