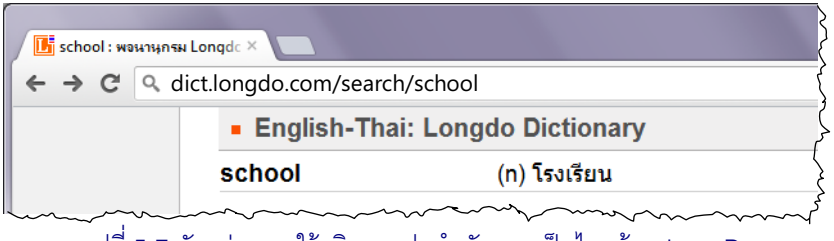
โปรแกรมแปลคำอังกฤษเป็นไทย

บริการแปลคำอังกฤษเป็นไทย หรือคำไทยเป็นอังกฤษมีอยู่หลายที่ในอินเทอร์เน็ต (ผู้อ่านสามารถค้นวลี "English Thai Dictionary" ด้วยกูเกิลดู) ในหัวข้อนี้จะขอใช้บริการของเว็บไซต์ LONGDO Dict [2] โดยจะเขียนโปรแกรมที่รับคำอังกฤษจากผู้ใช้ทางแป้นพิมพ์ จากนั้นตัวโปรแกรมจะเรียกใช้บริการที่ http://dict.longdo.com ได้หน้าเว็บคำแปลเป็นภาษาไทยกลับคืนมา ค้นหาบริเวณที่เก็บคำแปลในหน้าเว็บ แล้วนำคำแปลที่พบแสดงทางจอภาพ คำถามที่ตามมาคือ จะเขียนโปรแกรมต่อกับเว็บไซต์ได้อย่างไร จะส่งคำที่ต้องการหาคำแปลได้อย่างไร และผลที่ได้กลับมาเป็นอย่างไร คำแปลที่ได้อยู่ที่ไหน ?

เริ่มที่สองคำถามแรกก่อน ถ้าอยากทราบว่า "school" แปลว่าอะไร ก็ให้เปิดเว็บบราวเซอร์แล้วไปที่ http://dict.longdo.com/search/school จะได้ผลดังรูปที่ 5-7 หากต้องการคำแปลของ "scrape" ต้องไปที่ http://dict.longdo.com/search/scrape ดูสองตัวอย่างนี้แล้ว ผู้อ่านเดาได้หรือยังว่า ถ้าต้องการหาคำอื่นจะไปที่ใด ปัญหาคือจะให้โปรแกรมของเราไปที่เว็บเพจที่ต้องการได้อย่างไร ? ไม่ยาก ใช้ `Scanner` เช่นเคย แต่แทนที่จะส่ง `System.in` เพื่ออ่านจากแป้นพิมพ์ หรือส่ง `new File(...)` เพื่ออ่านจากแฟ้ม คราวนี้ส่ง `new URLStream(url)` เพื่ออ่านจากเว็บที่ตำแหน่ง `url` แทน [3] ดังตัวอย่างในรหัสที่ 5-18 เป็นการเชื่อมเพื่ออ่านเว็บเพจคำแปลของคำอังกฤษที่เก็บในตัวแปร `word` เมื่อสร้างตัวอ่านเว็บชื่อ `web` เสร็จ ก็สามารถใช้เมท็อด `web.nextLine()` ของ `Scanner` อ่านออกมาทีละบรรทัดเช่นที่เคยทำมา



รูปที่ 5-7 ตัวอย่างการใช้บริการแปลคำอังกฤษเป็นไทยด้วย LONGDO

```
String site = " http://dict.longdo.com/search/";
String url = site + word;
Scanner web = new Scanner(new URLStream(url), "UTF-8");
...
```

รหัสที่ 5-18 ส่วนของโปรแกรมอ่านเว็บเพจ

---

[2] เหตุผลที่เลือกเว็บไซต์นี้ไม่มีอะไรนอกจากความง่ายในการเขียนโปรแกรมเพื่อส่งคำไป และตีความคำแปลที่ได้กลับมา และต้องขอเน้นว่าเนื่องจากโปรแกรมนี้เรียกใช้เว็บไซต์ภายนอกเครื่องที่สั่งทำงาน ดังนั้นจึงต้องเชื่อมต่อกับอินเทอร์เน็ตก่อนใช้งาน และก็ไม่แน่เสมอไปว่าจะทำงานได้ เพราะเว็บไซต์อาจไม่พร้อมให้บริการก็เป็นได้

[3] `URLStream` เป็นคลาสพิเศษของ JLab มีชื่อเต็มว่า `jlab.URLStream`

คำถามถัดมาคือ ข้อมูลที่อ่านได้จากตัวอ่านเว็บมีรูปแบบเช่นไร รูปที่ 5-8 แสดงบรรทัดต้น ๆ ของข้อมูลในเว็บเพจที่อ่านได้ เมื่อไปหาคำแปลของคำว่า “school” (เครื่องหมาย … ในรูปแทนข้อมูลมากมายที่ขอไม่แสดงให้ครบ) ข้อมูลเหล่านี้อยู่ในรูปแบบที่เรียกว่า HTML (HyperText Markup Language) [4] ประกอบด้วยเนื้อความของเอกสาร การกำกับเนื้อความเพื่อบอกความหมาย ลักษณะการนำเสนอในเว็บบราวเซอร์ และอื่น ๆ อีกมากมาย

```
<html xmlns:og="http://opengraphprotocol.org/schema/"
xmlns:fb="http://www.facebook.com/2008/fbml">
<head>
  <meta property="og:title" content="คำศัพท์ school: พจนานุกรม Longdo ... />
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
  ...
</head>
...
HREF="search/school"><b>school</b></A></td><td>[N] โรงเรียน , <b>See also
...
```

ต้องหาบรรทัดที่มีข้อความนี้ก่อน

คำแปลอยู่ตรงนี้

รูปที่ 5-8 ตัวอย่างข้อความที่อ่านได้เมื่อต่อไปยังเว็บเพจของ longdo

และก็มาถึงคำถามสุดท้ายคือ แล้วคำแปลอยู่ที่ใด ? สำหรับเว็บไซต์ LONGDO Dict ผู้เขียนได้ศึกษาข้อมูลที่อ่านกลับมา พบว่า คำแปลจะอยู่ในบรรทัดที่มีข้อความ `HREF="search/` ตามด้วยคำศัพท์ (ดูบรรทัดก่อนสุดท้ายในรูปที่ 5-8) โดยคำแปลจะพบที่ตำแหน่งหลัง `"<td>"` และจบที่ตำแหน่งก่อนเครื่องหมายจุลภาค , (เนื่องจากอาจมีคำแปลหลายแบบ [5] เพื่อความง่าย จึงขอค้นและแสดงคำแปลแรกที่พบก็พอ) ภาระส่วนใหญ่ของการเขียนโปรแกรมนี้ก็คือ การหาคำแปลที่ปรากฏในบรรทัดที่มีรูปแบบที่ได้อธิบายมานี่เอง [6]

รหัสที่ 5-19 แสดงโปรแกรมสมบูรณ์เพื่อแปลคำอังกฤษเป็นคำไทย เริ่มด้วยการรับคำจากผู้ใช้ จากนั้น (บรรทัดที่ 9) ใช้ `trim` เพื่อขจัดช่องว่างซ้ายและขวา ตามด้วยการเปลี่ยนเป็นตัวเล็ก บรรทัดที่ 10 สร้างชื่อของเว็บเพจที่มีคำแปล แล้วสร้างตัวอ่านเว็บ ขอให้สังเกตที่ท้ายคำสั่ง มีการส่งสตริง `"UTF-8"` ไปด้วย ตรงนี้หมายความว่า เว็บเพจที่จะอ่านนี้มีการเข้ารหัสแบบที่เรียกว่า UTF-8 (ถ้าดูที่ปลายบรรทัดที่ 5 ของรูปที่ 5-8 จะพบข้อความ `charset=utf-8` เป็นการระบุว่าเอกสารนี้เข้ารหัสข้อความแบบ UTF-8) ก่อนเข้าวงวนค้นคำแปลในเอกสาร มีการเตรียมตัวแปร `thai` ให้มีค่าเริ่มต้นเป็น `""` และตัวแปร `pattern` เก็บข้อความที่เราต้องการค้นในเอกสาร

---

[4] หลังจากใช้บราวเซอร์ไปที่เว็บเพจแล้ว หากต้องการดูข้อมูล HTML ที่บรรยายหน้านั้น ให้คลิกเมาส์ปุ่มขวาที่หน้าเอกสาร แล้วเลือกเมนู View Source

[5] รูปที่ 5-7 ไม่ได้แสดงบราวเซอร์เต็มจอภาพ แต่ถ้าผู้อ่านลองต่อเว็บไปดูที่หน้านี้ จะพบว่า มีรายละเอียดคำแปลมากกว่าหนึ่งความหมายในหน้านี้

[6] เนื่องจากเว็บเพจถูกออกแบบมาเพื่อนำเสนอเนื้อหาให้คนดู แต่นี่เราต้องการเขียนโปรแกรมให้ “เลือกดู” ส่วนที่สนใจในเว็บเพจ ซึ่งต้องอาศัยการค้นหาข้อมูลในเว็บเพจ เรียกกลวิธีแบบนี้ว่า *การขูดเว็บ* (web scraping)

การค้นดำเนินไปเรื่อยๆ ตราบเท่าที่ยังไม่พบบรรทัดที่มีคำแปล (thai ยังเป็น "") และตัวอ่านเว็บยังอ่านไม่หมด (web.hasNext() เป็นจริง) เมื่ออ่านบรรทัดใหม่เข้ามา ก็ค้นด้วย indexOf ว่ามี pattern หรือไม่ (บรรทัดที่ 18) ถ้ามี ก็ค้น "<td>" ต่อ (บรรทัดที่ 21) โดยเริ่มค้นต่อจากที่พบ pattern และถ้าพบ "<td>" ก็ค้นเครื่องหมาย , ต่ออีกในบรรทัดที่ 24 ถ้าพบอีก คำแปลก็คือสตริงย่อยที่อยู่หลัง "<td>" ไปจนถึงก่อน , ที่พบ (บรรทัดที่ 25) หากไม่พบในบรรทัดที่เพิ่งอ่านมา ก็วนกลับไปอ่านบรรทัดถัดไป โดยที่ thai ยังเป็น "" เมื่อหลุดจากวงวนโดย thai ไม่ใช่ "" แสดงว่า thai เก็บคำแปล ก็แสดงคำแปลนั้นทางจอภาพ (บรรทัดที่ 30) ไม่เช่นนั้นให้แจ้งให้ผู้ใช้ทราบว่า ไม่พบคำแปล

```
01 import java.util.Scanner;
02 import jlab.URLStream;
03 // โปรแกรมแปลคำอังกฤษเป็นไทย
04 public class English2Thai {
05   public static void main(String[] args) {
06     Scanner kb = new Scanner(System.in);
07     System.out.print("คำ = ");
08     String eng = kb.nextLine();
09     eng = eng.trim().toLowerCase();
10     String url = "http://dict.longdo.com/search/" + eng;
11     Scanner web = new Scanner(new URLStream(url), "UTF-8");
12     String pattern = ("href=\"search/" + eng).toLowerCase();
13     String thai = "";
14     while (thai.equals("") && web.hasNext()) {
15       // href="search/target ... <td>[N] จุดมุ่งหมาย,
16       String t = web.nextLine();
17       t = t.toLowerCase();
18       int j = t.indexOf(pattern, 0);
19       if (j >= 0) {
20         j += pattern.length();
21         j = t.indexOf("<td>", j);
22         if (j >= 0) {
23           j += "<td>".length();
24           int k = t.indexOf(",", j);
25           if (k >= 0) thai = t.substring(j, k);
26         }
27       }
28     }
29     if (thai != null) {
30       System.out.println("แปลว่า " + thai.trim());
31     } else {
32       System.out.println("ไม่พบคำแปล");
33     }
34   }
35 }
```

รหัสที่ 5-19 โปรแกรมแปลคำอังกฤษเป็นไทย