© ECTI-CARD 2009, May'09, Bangkok, Thailand. ISBN: 978-974-8285-62-7

การจัดอันดับความนิยมของศัพท์ภาษาอังกฤษในเว็บ

ทศพล ธนะทิพานนท์
วิทยากรและนักเขียนอิสระ
thanatipanonda@gmail.com

วรเศรษฐ สุวรรณิก
ภาควิชาวิทยาการคอมพิวเตอร์
มหาวิทยาลัยเกษตรศาสตร์
worasait.suwannik@gmail.com

บทคัดย่อ

บทความนี้เสนอวิธีการจัดอันดับความนิยมของศัพท์ภาษาอังกฤษ แนวคิดในการจัดอันดับก็คือดูจากจำนวนเว็บเพจที่มีคำศัพท์ที่สนใจปรากฏอยู่ โดย Google สามารถประมาณจำนวนเว็บเพจได้ เราถือว่าคำที่มีผลการค้นหาของ Google มากกว่าเป็นคำที่ได้รับความนิยมมากกว่า ผลการจัดอันดับก็คือคำที่ได้รับความนิยมมากที่สุดในเว็บคือ www และเราได้นำคำที่นิยมมากที่สุดประมาณสามพันคำไปจัดพิมพ์เป็นหนังสือ

Abstract

This paper proposes a method to rank English word popularity. The idea is to look at the number of web pages that contains the word. Google can estimate the number. We assume that a word with more Google search result is more popular. The result is "www" is the most popular word in the web. Finally, we took about three thousand words that are most popular and published as a book.

คำสำคัญ

ศัพท์, ภาษาอังกฤษ, Google, จัดอันดับความนิยม

1. บทนำ

ภาษาอังกฤษเป็นภาษาที่สำคัญมากเป็นอันดับต้นๆของภาษาทั้งหลายในโลก ข้อมูลข่าวสารจำนวนมากเขียนเป็นภาษาอังกฤษ เช่น ในเดือนธันวาคม 2551 เว็บ wikipedia.org มีบทความภาษาอังกฤษอยู่ประมาณ 2.7 ล้านบทความ และในปัจจุบันมีชาวต่างชาติเดินทางเข้ามาที่ประเทศไทยมากขึ้น โอกาสที่เราต้องติดต่อสื่อสารกับคนเหล่านั้นก็มีมากขึ้นตามไปด้วย

เครื่องมือสำคัญอันหนึ่งในการเรียนภาษาอังกฤษก็คือพจนานุกรมอังกฤษ-ไทยหรืออังกฤษ-อังกฤษ พจนานุกรมเป็นหนังสืออ้างอิงซึ่งเรียงคำศัพท์ตามลำดับตัวอักษรจาก a-z ถึงแม้จะมีประโยชน์มากในการเรียนภาษา แต่พจนานุกรมไม่ได้แนะนำว่าควรท่องศัพท์อะไรก่อนหรือหลัง ศัพท์ไหนมีความสำคัญมากหรือน้อย ใช้บ่อยหรือว่าใช้น้อย ทำให้ผู้เรียนภาษาไม่ทราบว่าควรจะเลือกท่องศัพท์ใดก่อนหรือหลัง

อย่างไรก็ดี มีพจนานุกรมบางเล่มแสดงความสำคัญของคำศัพท์ เช่น พจนานุกรม Se-ed's New Compact English-Thai Dictionary ได้ทำเครื่องหมายรูปดาวที่หน้าคำศัพท์ที่ควรรู้และควรจำ และพจนานุกรม Collins COBUILD Advanced Learner's English Dictionary บอกระดับความถี่ของการใช้งานศัพท์ แต่ถึงแม้จะมีการแสดงความสำคัญของศัพท์แล้วก็ตาม แต่การจัดอันดับศัพท์เหล่านั้นก็ยังไม่ละเอียดมาก เช่น อาจจะมีระดับความถี่ของการใช้งานเพียง 5 ระดับ เป็นต้น

บทความนี้เสนอวิธีการจัดอันดับความสำคัญของคำศัพท์ภาษาอังกฤษโดยดูจากความนิยมในการใช้คำในอินเทอร์เน็ท ประโยชน์ของการจัดอันดับคำศัพท์ภาษาอังกฤษคือ ผู้เริ่มต้นศึกษาภาษาอังกฤษเลือกศัพท์มาท่องได้อย่างมีเป้าหมายมากขึ้น, คุณครูสามารถเลือกศัพท์ให้นักเรียนท่องได้อย่างเหมาะสมตามระดับของนักเรียน, คุณพ่อคุณแม่สามารถเลือกศัพท์ให้คุณลูกท่องได้อย่างเหมาะสม, ใช้ประกอบการทบทวนศัพท์ภาษาอังกฤษในระดับเบื้องต้น, และทำให้รู้จักศัพท์ที่นิยมมากที่สุดจำนวนประมาณหนึ่งร้อยคำของทุกหมวดหมู่รวมกัน

2. จัดอันดับความนิยม

เราจัดอันดับความนิยมในการใช้ศัพท์ภาษาอังกฤษโดยดูจากจำนวนผลการค้นหาของ Google

ตัวอย่างเช่น ถ้าเราอยากจะรู้ว่าคำว่า ant ได้รับความนิยมเพียงใด ก็ไปที่ Google, พิมพ์คีย์เวิร์ดว่า ant, และคลิกค้นหา ซึ่ง Google ก็จะแสดงผลการค้นหามาให้ และบอกด้วยว่าผลการค้นหามีประมาณกี่เว็บเพจ เช่น

ผลการค้นหา 1 - 10 รายการจากประมาณ 93,200,000 สำหรับคำว่า ant (0.18 วินาที)

เราถือว่าคำที่มีผลการค้นหามากกว่า จะเป็นคำที่นิยมมากกว่า และควรท่องศัพท์ที่คนนิยมใช้ ก่อนที่จะท่องศัพท์ที่ไม่ได้รับความนิยม

3. ระดับความสามารถ

เราเชื่อว่าคนที่เก่งภาษาอังกฤษมากกว่าอีกคนอยู่หนึ่งระดับขั้น จะต้องรู้ศัพท์มากกว่าอีกคนหนึ่งเกินหนึ่งเท่าตัว ไม่เช่นนั้นก็จะแสดงความสามารถที่สูงกว่าออกมาให้เห็นอย่างชัดเจนไม่ได้ ด้วยเหตุนี้ ระดับการรู้ศัพท์จึงถูกแบ่งตามจำนวนคำศัพท์ที่รู้ ดังแสดงในตารางที่ 1

ตารางที่ 1 จำนวนคำที่ต้องรู้ในแต่ละระดับ

ระดับ	จำนวนคำที่ต้องรู้
1	1
2	3
3	7
4	20
5	55
6	148
7	403
8	1,097
9	2,981
10	8,103
11	22,026
12	59,874

เรากำหนดให้จำนวนคำที่ต้องรู้ในแต่ละระดับแสดงในคอลัมน์ที่ 2 ของตารางที่ 1 ค่าเหล่านี้เกิดจากการนำค่า e หรือฐานของล็อกธรรมชาติ (natural logarithm) ซึ่งมีค่าประมาณ 2.72 มายกกำลัง ในระดับที่ 1 มีศัพท์ต้องรู้เท่ากับ e^0 หรือ 1 คำ, ในระดับที่ 2 มีศัพท์ต้องรู้ e^1 คำ ซึ่งเมื่อปัดเป็นจำนวนเต็มจะได้ 3 คำ, ในระดับที่ 3 มีศัพท์ต้องรู้ e^2 คำ ซึ่งเมื่อปัดเป็นจำนวนเต็มจะได้ 7 คำ, เช่นนี้เรื่อยไป

สรุปคือจำนวนศัพท์ต้องรู้ในแต่ละระดับ ก็คือ e^{l-1} โดยที่ l คือระดับความสามารถ (level)

4. รายละเอียดการพัฒนา

4.1 ภาพรวมของระบบ

กระบวนการจัดอันดับคำศัพท์ภาษาอังกฤษยอดนิยมเป็นดังนี้

1. รวบรวมคำศัพท์ เริ่มต้นจากการดึงข้อมูลคำศัพท์ภาษาอังกฤษจากฐานข้อมูล Lexitron ซึ่งดาวน์โหลดได้ที่ http://lexitron.nectec.or.th

2. ตรวจสอบความนิยม นำคำศัพท์มาค้นด้วย Google API โดยสนใจเฉพาะจำนวนผล (result) ที่ Google ประมาณมาให้

3. แบ่งชนิดคำ นำคำศัพท์มาแบ่งชนิดตามที่ระบุไว้ในฐานข้อมูล Lexitron เช่น คำนาม, กริยา, คุณศัพท์, ฯลฯ โดยตัดชนิดตัวย่อ, อุปสรรค (prefix), ปัจจัย (suffix) ออกไป และเรียงลำดับศัพท์แต่ละชนิดตามความนิยม

4. ใส่คำแปล จากฐานข้อมูล Lexitron

4.2 การออกแบบและพัฒนาระบบ

4.2.1 ฐานข้อมูล Lexitron

ฐานข้อมูล Lexitron พัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ฐานข้อมูลนี้เป็น text file ที่มีรูปแบบคล้าย XML ฐานข้อมูลนี้ประกอบด้วยข้อมูลพจนานุกรมอังกฤษ-ไทย และ ไทย-อังกฤษ และมีข้อมูลเกี่ยวกับชนิดของคำและความหมาย บทความนี้ใช้เฉพาะฐานข้อมูลพจนานุกรมอังกฤษ-ไทย

ชนิดของคำในฐานข้อมูลแสดงในตารางที่ 2 เรายุบคำในหมวด INTER ไปรวมกับ INT เพราะว่าคำในหมวด INTER มีเพียงคำเดียว และเราไม่นำคำในหมวดคำย่อ, อุปสรรค (prefix), และปัจจัย (suffix) มาจัดอันดับ

ตารางที่ 2 ชนิดของคำในฐานข้อมูล Lexitron

คำย่อ	คำเต็ม	ชื่อไทย
ABBR	abbreviation	คำย่อ
ADJ	adjective	คุณศัพท์
ADV	adverb	วิเศษณ์
ART	article	อาร์ทิเคิล
AUX	auxililary verb	กริยานุเคราะห์
CONJ	conjunction	สันธาน

DET	determiner	คำบ่งชี้
IDM	idiom	สำนวน
INT	interjection	คำอุทาน
INTER	interjection	คำอุทาน
N	noun	นาม
PHRV	phrasal verb	กริยาวลี
PREP	preposition	บุพบท
PRF	prefix	อุปสรรค
PRON	pronoun	สรรพนาม
SL	slang	สแลง
SUF	suffix	ปัจจัย
VI	intransitive verb	อกรรมกริยา
VT	transitive verb	สกรรมกริยา

4.2.2 ตรวจสอบความนิยมของศัพท์

เราใช้ Google เพื่อตรวจสอบความนิยมของศัพท์ โดยส่งคำร้อง (request) ไปที่ http://ajax.googleapis.com/ajax/services/search/web [1] พร้อมกับส่งคำศัพท์ที่เราต้องการค้นหาไปทาง query string พารามิเตอร์ของ query string ที่เราใช้มีดังนี้ [2, 3]

- v คือ version ในที่นี้เราส่งค่า 1.0 ให้
- lr คือภาษา ในที่นี้เราส่งค่า lang_en เพื่อให้ค้นหาเฉพาะเว็บเพจภาษาอังกฤษ
- q คือคำที่ต้องการค้นหา เราจะใส่เครื่องหมายคำพูดล้อมรอบคำ ในกรณีที่คำศัพท์ที่ต้องการค้นหาประกอบด้วยคำสองคำ ดังเช่นคำในหมวดกริยาวลี

เราเขียนโปรแกรมภาษาจาวาเพื่อให้การตรวจสอบความนิยมเป็นไปโดยอัตโนมัติ เราใช้คลาส URL เพื่อส่งคำร้องไปยังเว็บของ Google คำสั่งที่ใช้คือ

```
URL url = new URL("http://ajax.googleapis.com/"
        + "ajax/services/search/web"
        + "?v=1.0"
        + "&lr=lang_en"
        + "&q=%22" + word + "%22");
```

โดยที่ word เป็นตัวแปรที่เก็บคำศัพท์ที่ต้องการค้นหา

หลังจากเชื่อมต่อไปยังเว็บของ Google แล้ว เราก็ใช้ stream เพื่ออ่านข้อมูลจาก URL ผลการค้นหาเป็นแบบ JSON (JavaScript Object Notation) [4] มีลักษณะทำนองนี้ (ข้อมูลบางส่วนถูกแทนด้วยจุดสามจุด ... และเราจัดรูปแบบการเยื้องของข้อมูลใหม่เพื่อให้ผู้อ่านเข้าใจง่ายขึ้น)

```
{
  "responseData":{
    "results":[...],
    "cursor":{
      "pages":[...],
      "estimatedResultCount":"6730000000",...
    }
  },
  "responseDetails": null,
  "responseStatus": 200
}
```

เราใช้คลาส JSONObject [5] เพื่อดึงค่าของ estimatedResultCount ด้วยคำสั่งต่อไปนี้

```
JSONObject json;
json = new JSONObject(stringBuilder.toString());
return json.getJSONObject("responseData")
          .getJSONObject("cursor")
          .getLong("estimatedResultCount");
```

โดยที่ stringBuilder เป็นตัวแปรที่เก็บค่าที่อ่านจาก stream

เพื่อความรวดเร็วในการทำงาน โปรแกรมที่เขียนขึ้นเป็นแบบหลายเธรด (multi-thread) โดยแบ่งเป็น 50 เธรด ใช้เวลา 25 นาที ในการค้นคำ 60,994 คำ

4.2.3 จัดรูปแบบข้อความ

ผลการจัดอันดับคำถูกเก็บไว้ในไฟล์ CSV (comma separated value) สำนักพิมพ์วรรณิกได้นำผลการจัดอันดับ

คำไปจัดพิมพ์เป็นหนังสือชื่อ “ศัพท์ภาษาอังกฤษยอดนิยม” ISBN 978-974-603-457-9 โดยได้จัดอันดับศัพท์ประมาณสามพันคำที่นิยมใช้ที่สุดในอินเทอร์เน็ท ศัพท์ในระดับ 1 ถึง 6 จะเรียงกันตามความนิยม จากนิยมมากไปหาน้อย โดยไม่มีการแยกชนิดของคำ แต่ในระดับ 7 ถึง 9 เราจัดกลุ่มศัพท์ตามชนิดของคำ (เช่น คำนาม, กริยา, ฯลฯ) และเรียงลำดับคำในแต่ละชนิดตามความนิยม

เราเตรียมต้นฉบับหนังสือด้วย Microsoft Word 2003 เนื่องจากหนังสือเล่มนี้มีศัพท์อยู่เป็นจำนวนมาก เราจึงเขียนแมโครด้วยภาษา VBA (Visual Basic for Application) เพื่อจัดรูปแบบอย่างรวดเร็ว โดยแมโครที่เขียนขึ้นสามารถทำให้ศัพท์ภาษาอังกฤษบางส่วนเป็นตัวหนาโดยอัตโนมัติ

5. ผลการทดสอบและการวิจารณ์ผล

เราเห็นว่าผลที่น่าสนใจมากที่สุดสำหรับการรวบรวมศัพท์ยอดนิยมครั้งนี้ ก็คือ ทราบว่าคำที่ปรากฏมากที่สุดในอินเทอร์เน็ทคือคำว่า www แต่น่าเสียดายที่คำดังกล่าวเป็นคำย่อ ซึ่งเราไม่นำมาจัดอันดับความนิยม โชคดีที่คำที่นิยมเป็นอันดับถัดมาไม่ใช่คำย่อ

ศัพท์ที่ได้รับความนิยม 100 อันดับแรก คือ and, by, the, of, a, to, is, with, from, at, or, all, other, one, may, time, about, do, can, has, was, but, up, will, if, information, when, us, so, out, site, your, find, re, back, search, name, go, date, top, day, post, best, web, get, am, just, how, my, there, use, me, have, pm, id, click, list, profile, show, free, online, which, old, business, type, good, forum, here, in, who, video, next, more, this, for, like, last, love, help, home, not, contact, our, new, be, south, out of, are, view, more than, you, no, food, France, that, no, now, I, law, an

สังเกตว่าในศัพท์ยอดนิยม 100 คำแรก มีศัพท์ที่เกี่ยวข้องกับคอมพิวเตอร์อยู่มากมาย เชน information, site, search, post, web, click, profile, online, forum, video, contact

6. บทสรุป

เรานำคำศัพท์ภาษาอังกฤษจาก Lexitron มาจัดอันดับความนิยม เครื่องมือที่เราใช้จัดอันดับความนิยมก็คือเว็บของ Google และเรานำผลการจัดอันดับความนิยมมาจัดพิมพ์เป็นหนังสือ กระบวนการเกือบทั้งหมดเป็นไปโดยอัตโนมัติ ทำให้สามารถเรียบเรียงเป็นหนังสือได้ภายในเวลา 7 วัน

6.1 แนวทางการพัฒนาต่อ

เพื่อให้การจัดอันดับศัพท์มีความถูกต้องแม่นยำขึ้น ผู้ที่สนใจพัฒนาต่ออาจจะนำความถี่ของจำนวนคำที่ปรากฏในแต่ละเว็บเพจเข้ามาจัดอันดับด้วย ไม่ใช่ดูแค่จำนวนเว็บเพจที่มีคำนั้นเพียงอย่างเดียว

นอกจากนั้นอาจจะพัฒนาการจัดอันดับโดยดูจากชนิดของคำ เพราะคำบางคำนิยมใช้ในชนิดหนึ่งมากกว่าอีกชนิดหนึ่ง เช่น คำว่า can มีหลายความหมาย ถ้าเป็นคำนามแปลว่ากระป๋อง แต่ถ้าเป็นกริยานุเคราะห์แปลว่าสามารถ ถ้าจัดอันดับตามชนิดของคำแล้ว คำว่า can ที่เป็นกริยานุเคราะห์น่าจะใช้บ่อยกว่าคำนาม

คำบางคำได้อันดับเพิ่มขึ้นเพราะมีตัวย่อมาช่วยเพิ่มจำนวนผลการค้นหา เช่น คำว่า us ได้คำย่อ US (United States) มาช่วย และคำว่า la ได้คำย่อ LA (Los Angeles) มาช่วย ดังนั้นการพัฒนาต่อไปอาจจะต้องหาวิธีตัดการช่วยเหลือจากตัวย่อ

สำหรับงานถัดไปอาจจะเป็นการจัดอันดับคำยอดนิยมภาษาไทยและภาษาอื่นๆ

7. กิตติกรรมประกาศ

เราขอบคุณศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ที่ได้จัดทำฐานข้อมูล Lexitron และขอบคุณ Google ที่ให้บริการ API สำหรับการค้นหา

8. เอกสารอ้างอิง

[1] Google AJAX Search API: Flash and other Non-Javascript Environments, http://code.google.com/intl/th/apis/ajaxsearch/documentation/#fonje

[2] Google AJAX Search API: Standard URL Arguments, http://code.google.com/intl/th/apis/ajaxsearch/documentation/reference.html#_intro_fonje

[3] Google AJAX Search API: Language Collection Values, http://www.google.com/coop/docs/cse/resultsxml.html#languageCollections

[4] JavaScript Object Notation, http://www.json.org/

[5] JSON Java API, http://www.json.org/javadoc/

 ISBN: 978-974-8285-62-7