ชื่อ _____ นามสกุล _____ รหัสนักศึกษา_____

Lab Sheet 3: Road to Google Part 1

เคล็ดลับในการค้นหาเว็บที่ต้องการจากคำสำคัญที่ Google ใช้นั้น เริ่มต้นมาจากการแบ่งเนื้อหาในเว็บออกเป็นส่วนย่อย ๆ หรือคำย่อย ๆ ถ้าคำใดปรากฏเยอะแสดงว่าเว็บนั้นน่าจะมีเนื้อหาที่เกี่ยวข้องกับคำนั้นเยอะ ในแลปนี้เราจะเรียนรู้พื้นฐานที่จำเป็นสำหรับการสร้าง Google

1. เราจะกำหนดให้ตัวอักษรที่อยู่ในข้อความหรือเนื้อหาต้องเป็นตัวอักษรเหล่านี้เท่านั้นเพื่อความสะดวกในการคำนวณ

   A–Z a–z ช่องว่าง 0–9 . , ? ! / % ( ) - + = ‘ & : ;
2. ให้นักศึกษาเขียนฟังก์ชัน **word_prob** ฟังก์ชันนี้จะทำการคำนวณความน่าจะเป็นที่คำที่กำหนดให้จะปรากฏในข้อความหลัก โดยรับ Argument สองตัวคือข้อความหลักและคำที่กำหนดให้ ฟังก์ชันนี้ให้ส่งค่าความน่าจะเป็นของคำที่ต้องการค้นหากลับมา (ดูตัวอย่างในหน้าถัดไป)
3. ให้ save ไฟล์เป็นชื่อ lab03_5xxxxxxxx.php โดย 5xxxxxxxx คือรหัสนักศึกษาของตัวเอง จากนั้นให้อัพโหลดไฟล์ไปยังเว็บส่งการบ้าน http://hw.cs.science.cmu.ac.th/CS_HW/p204202.html

ข้อควรระวัง

1. ตรวจสอบชื่อของฟังก์ชันว่าสะกดถูกต้องหรือไม่ ตัวพิมพ์เล็กพิมพ์ใหญ่มีความสำคัญ หากตั้งชื่อฟังก์ชันผิดจะทำให้เสียคะแนนได้
2. ตรวจสอบชื่อไฟล์ที่จะส่ง ตัวพิมพ์เล็กพิมพ์ใหญ่มีความสำคัญ หากตั้งชื่อไฟล์ผิดจะไม่ได้รับการตรวจ ให้ส่งไฟล์มาใหม่โดยไม่ลบไฟล์เก่าเพื่อเป็นหลักฐานว่าไม่ได้ส่งช้าเกินกำหนด
3. งานชิ้นนี้ส่งภายในวันที่ 9 **กันยายน 2558 เวลา 23:59**

ตัวอย่างผลลัพธ์ของการใช้ฟังก์ชัน word_prob (สีเป็นตัวบ่งบอกว่าคำนี้อยู่ตรงไหนของข้อความ)

```
$str = "New editions? This book aims to give readers a clear
understanding of the statistical methods that are widely used in
epidemiological analysis without depending on advanced mathematical
or statistical theory. It contains 13 chapters. Like any new
editions, some old material in this book is removed and new material
is added. The section on rarely used crossover designs is removed and
new sections on frequently used contingency table analysis are added.
The chapter on matched-data analysis (Chapter 10) contains new
sections on the analysis of multilevel categorical data. Seven
appendixes (description of the WCGS data set, binomial and Poisson
probability distributions, the odds ratio and its properties,
partitioning the chi-square statistic, maximum likelihood estimation
and likelihood functions, and problems) are also included. Problems
in Appendix F are designed to help readers gain a deeper
understanding of statistical principles.";

$words = array("section", "editions", "New", "new", "are");

foreach ($words as $word)
     echo "The probability that $word appears is " .
round(word_prob($str,$word),2) . "%<br>";
```

The probability that section appears is 0.74%
The probability that editions appears is 1.47%
The probability that New appears is 0.74%
The probability that new appears is 2.94%
The probability that are appears is 2.94%

ข้อสังเกต

1. คำที่กำหนดให้จะต้องเหมือนกับคำที่อยู่ในข้อความ ไม่ว่าจะเป็นการสะกด ตัวพิมพ์เล็กพิมพ์ใหญ่ เติม s หรือไม่เติม s ถึงจะนำมาคำนวณ
2. เราจะสังเกตว่าคำว่า are เป็นส่วนหนึ่งของคำว่า chi-square แต่เราจะไม่นำคำว่า chi-square มานับรวมกับคำว่า are
3. ความน่าจะเป็นของคำที่กำหนดคือ จำนวนคำที่กำหนดหารด้วยจำนวนคำทั้งหมดในข้อความ