

Homework 6

ข้อแนะนำ

คำสั่งด้านล่างเป็นคำสั่งที่นิติตอาจต้องใช้สำหรับการบ้าน หากลืมวิธีใช้ไปแล้ว หรือ อาจารย์ยังไม่ได้สอน นิติตมี man pages และ Google เป็นเพื่อนเสมอ

- | | | | |
|-------|--------|-------|--------|
| - wc | - uniq | - tr | - awk |
| - cat | - sort | - sed | - grep |

Webages ด้านล่างนี้อาจมีประโยชน์ ด้วย

- sed : <https://www.grymoire.com/Unix/Sed.html>
- awk : <https://www.grymoire.com/Unix/Awk.html>
- grep : <https://opensourceforu.com/2012/06/beginners-guide-gnu-grep-basics/>
- curl : <https://curl.haxx.se/libcurl/>

โจทย์

1) ไม่ว่าพวกเพนกวินจะชอบเกมเท่าใดก็ตาม มันก็จะหยุดเล่นเพื่ออ่านหนังสือทุกเมื่อหากมีโอกาas เรื่องมีอยู่ว่า หลังจากอ่านหนังสือเรื่อง "Alice in Wonderland" (<https://www.gutenberg.org/files/11/11-0.txt>) เจ้าเพนกวินน้อยตัวหนึ่งนามว่า Patricia เกิดอยากรู้ว่า แต่ละคำในหนังสือถูกใช้ซ้ำไปกี่ครั้ง เรามีเขียน Script เพื่อช่วย Patricia นับจำนวนคำซ้ำในหนังสือกันเถอะ

เราสามารถเริ่มต้นโดย

1. Download file 11.txt มาไว้ที่ Directory ก่อน
2. จากนั้นเปลี่ยนตัวอักษรตัวใหญ่ ให้เป็นตัวเล็กให้หมด
3. ลบ เลข วรรคตอน หรือ สัญลักษณ์ใด ๆ ที่ไม่ใช่ตัวหนังสือออก
4. เขียน Script ชื่อ count_uses.sh ที่จะอ่าน 11.txt เข้ามา และ แสดงรายการคำแต่ละคำที่อยู่ในไฟล์พร้อม ๆ กับ จำนวนครั้งที่คำนั้นปรากฏซ้ำใน File
 - a. โดย Script นี้จะถูกรันในรูปแบบดังนี้ ./count_uses.sh 11.txt
 - b. Output ของ Script ควรถูกเก็บในไฟล์ชื่อ output.txt (ข้อความด้านล่างแสดงตัวอย่างข้อความของใน output.txt)

```
hello 156
i 45
like 34
bananagrams 22
...
```

ข้อแนะนำ : ใช้ keywords ด้านล่าง หาใน Google

- sed remove numbers from file
- count occurrences of unique words in a file

ข้อจำกัด:

- อย่าใช้อาเรย์ – Script ที่นิสิตจะส่งควรใช้ประโยชน์จาก คำสั่ง ที่อยู่ในรายการด้านบน
- แต่ละคำใน output.txt ควรจะเรียงตามลำดับความถี่จากมากไปน้อย (ค้นคำและตัวเลขด้วยช่องว่างช่องเดียว) คำที่มีจำนวนครั้งที่ซ้ำเหมือนกันให้เรียงลำดับตามตัวอักษร

2) ในขณะที่ Patricia กำลังอ่าน Alice in Wonderland, เพื่อนของเธอ Charlie อยากจะอ่าน Macbeth ของ Shakespeare มากกว่า เพราะเขาจะต้องซ้อมบทละครในเรื่องนี้ เขารู้ URL ที่จะไป Download บทมาได้ แต่ต้องการความช่วยเหลือจากพวกเราในการประมวลบทละคร กล่าวคือ แม้เขาไม่ได้เป็นเพนกวิ้นเรื่องมาก จะให้เขาเล่นเป็นตัวละครอะไรก็ได้ แต่เขาขอเสียงตัวละครที่มีบรรทัดพูดมากที่สุด

เพื่อที่จะ Download บทละคร, นิสิตสามารถใช้คำสั่ง curl (อ่าน man page ของ curl เพื่อทำความเข้าใจด้วยนะ)

```
curl -o ./raw.html http://shakespeare.mit.edu/macbeth/full.html
```

ให้เขียน Script ชื่อ play_parser.sh (โดยมีคำสั่ง awk และ sed ด้านใน) โดย Script จะแสดงผลการประมวลบทละครออกมาหนึ่งบรรทัด ดังตัวอย่าง

58 george

ซึ่ง George ก็คือ ตัวละครที่มีบทพูดนับเป็นจำนวนบรรทัดมากที่สุด (ให้ทำชื่อตัวละครเป็นตัวพิมพ์เล็กทั้งหมดด้วย)

สำหรับโจทย์ข้อนี้, เราจะนับบทละครว่าเป็นหนึ่งบรรทัด หากมันเป็นบทพูดของตัวละครหนึ่ง และ บรรทัดนั้นมี Tag <A> ดังตัวอย่าง

```
<A NAME=speech15><b>MACBETH</b></a>
<blockquote>
<A NAME=3.4.43>Sweet remembrancer!</A><br>
<A NAME=3.4.44>Now, good digestion wait on appetite,</A><br>
<A NAME=3.4.45>And health on both!</A><br>
</blockquote>

<A NAME=speech16><b>LENNOX</b></a>
<blockquote>
<A NAME=3.4.46>May't please your highness sit.</A><br>
<p><i>The GHOST OF BANQUO enters, and sits in MACBETH's place</i></p>
```

</blockquote>

MACBETH

<blockquote>

Here had we now our country's honour roof'd,

Were the graced person of our Banquo present;

Who may I rather challenge for unkindness

Than pity for mischance!

</blockquote>

ตัวอย่างด้านบน มีบทพูดของ Macbeth 7 บรรทัด และ ของ Lennox 1 บรรทัด

เพื่อให้แบบฝึกหัดง่ายขึ้น, หากมีบรรทัดไหนถูกพูดด้วยตัวละครหลาย ๆ คนพร้อมกัน (เช่น ถูกพูดโดย “all,” “lords,” “both murderers,” or “soldiers”) ให้สมมติว่า บรรทัดนั้นถูกพูดโดยตัวละครอีกตัวหนึ่งไปเลย (ไม่ต้องไปหาว่า “all” นี้หมายถึงใครในฉากนี้บ้าง ในสมมติไปเลยว่า บรรทัดนั้นถูกพูดโดยตัวละครที่ชื่อ all)