ตัวตอบคำถามภาษาไทยโดยใช้คลังข้อมูลจากวิกิพีเดียภาษาไทยด้วยวิธีการ regular expression

ที่มาและความสำคัญ

เครื่องมือตอบคำถามเป็นภาษาไทย มีจำนวนน้อย และค่อนข้างที่จะเฉพาะทาง ปกติเมื่อทำการค้นหา จะค้นหาโดยใช้คำสำคัญในการค้นหา และจะได้ข้อมูลตอบกลับเป็นเอกสารหนึ่งๆ ซึ่งผู้ใช้งานจะต้องอ่านเอกสารเหล่านั้น จึงจะได้คำตอบที่ต้องการ ในขณะที่ปกติจะคิดคำถามขึ้นมาเป็นประโยคคำถาม และต้องการคำตอบที่ต้องการโดยไม่ต้องเสียเวลาอ่าน และวิกิภาษาไทยเป็นคลังข้อมูลภาษาไทยขนาดใหญ่ที่ทุกคนสามารถเข้าถึงได้

จุดประสงค์หลัก

เลือกเอกสารจากคำถามเพื่อค้นหาคำตอบได้ถูกต้อง และสามารถหาคำตอบได้ เป็นคำตอบสั้นๆ โดยคำถามจะต้องเป็นคำถามที่ลงท้ายหรือขึ้นต้นด้วยกลุ่มคำเหล่านี้ [‘ใคร’ ,‘ใด’, ‘ไหน’, ‘อะไร’,’เมื่อไหร่’]

วิธีการดำเนินการ

1. โหลดข้อมูลสำหรับเป็นฐานข้อมูลในการค้นหา https://dumps.wikimedia.org/thwiki/20181201/
2. ทำการ clean data ให้อยู่ในรูปแบบที่จะใช้งาน
   2.1 ดึงหัวข้อและชื่อไฟล์ทั้งหมด จัดเก็บเพื่อใช้งาน ในรูปแบบ json
   2.2 จัดทำข้อมูลความถี่ของแต่ละหัวข้อจากข้อมูลทั้งหมด เพื่อพิจารณาความสำคัญของคำ (โดยคาดว่า คำที่มีความถี่สูง จะมีความสำคัญน้อยกว่า)
3. ทำการเลือกหัวข้อเอกสาร จากคำถาม
   3.1 ใช้ pythainlp ในการตัดคำถามเบื้องต้น
   3.2 ตรวจสอบการตัดคำและปรับคำให้เหมาะสม โดยใช้หัวเรื่อง(article) ในการตรวจสอบ ตัวอย่าง เช่น คำว่า ‘เฮโจเกียว’ เมื่อใช้ pythainlp จะตัดได้เป็น [‘เฮ’,’โจเกียว’]
   3.3 คัดกรองเฉพาะคำที่เป็นหัวข้อ โดยเลือกคำที่มีความยาวมากที่สุดที่สามารถเป็นหัวข้อได้
   3.4 กรองและเรียงลำดับความน่าจะเป็น(ความสำคัญ)ที่จะมีคำตอบอยู่ในเอกสารหัวข้อนั้น โดยใช้ข้อมูลความถี่ของหัวข้อ ที่ได้จาก 2.2
   3.5 หาคะแนนของแต่ละหัวข้อ โดยคะแนนขึ้นอยู่กับจำนวนคำในคำถามปรากฏในเอกสารหัวข้อนั้นๆทั้งหมดกี่คำ (ไม่นับคำซ้ำ) และเรียงลำดับหัวข้อตามคะแนน
4. หาคำตอบจากเอกสาร
   4.1 ใช้วิธีการค้นหาคำตอบโดยใช้ regular expression สร้าง pattern ในการค้นหาคำตอบ
      4.1.1 นำหัวข้อเอกสารที่เลือกแทนที่ไปในคำถาม เพื่อให้การตัดคำออกมาได้โดยหัวข้อยังคงเดิม
      4.1.2 ทำการตัดคำถาม

4.1.3 นำรูปแบบที่จะใช้สกัดคำตอบ แทนที่คำที่เป็นคำถาม [‘ใคร’ ,’อะไร’,’ไหน’,’ใด’,’ชนิดใด’,’กี่’,’ เมื่อไหร่’]

4.1.4 นำคำที่ตัดมาสร้างเป็น pattern ทั้งหมดที่เป็นไปได้ เช่น ติดกันทุกคำ มีช่องว่างระหว่างทุกคำ มีคำอื่นๆนอกจากคำที่เป็นคำในคำถาม แทรกระหว่างคำ เป็นต้น

ผลการดำเนินการ

-สามารถตอบคำถามได้อย่างถูกต้องเฉพาะคำถามที่อยู่ในรูปแบบ ขึ้นต้นด้วยชื่อเฉพาะและลงท้ายด้วยคำที่เป็นคำถาม และมีชื่อหัวข้อเอกสารปรากฏอยู่ในคำถาม ค่อนข้างจำเพาะรูปแบบคำถามที่สามารถหาคำตอบได้

ตัวอย่างคำถามที่สามารถตอบได้อย่างถูกต้อง

1. คำถาม เฮโจเกียวเป็นเมืองหลวงของประเทศใด

   คำตอบ ญี่ปุ่น

2. คำถาม มหาวิทยาลัยคอร์เนลเป็นมหาวิทยาลัยเอกชนในไอวีลีกตั้งอยู่ที่เมืองใด

   คำตอบ อิทาคา

3. คำถาม คาบสมุทรจัตแลนด์เป็นคาบสมุทรในทวีปอะไร

   คำตอบ ยุโรป

ตัวอย่าง คำถามที่เลือกเอกสารถูกต้อง แต่คำตอบอาจไม่ถูกต้อง

1. คำถาม จีนกันส์ฉบับปฐมฤกษ์เปิดตัวเมื่อใด
   คำตอบ เมื่อเดือนกันยายน
2. คำถาม สถานีมารีนาเซาท์เพียร์ตั้งอยู่ที่ไหน
   คำตอบ ในพื้นที่ที่ถมทะเล
3. คำถาม วัดเขาจีนแล อยู่ในจังหวัดอะไร
   คำตอบ ลพบุรีประมาณ

อุปสรรค

1. จำนวนข้อมูลที่มีปริมาณมาก ใช้เวลาในการจัดการข้อมูลเป็นระยะเวลานาน
   ตัวอย่างเช่น การนับจำนวนคำที่ปรากฏในเอกสารทั้งหมด เนื่องจากใช้ระยะเวลาดำเนินการเกิน 72 ชั่วโมง จึงทำให้จำเป็นต้องใช้จากการนับในเอกสารทั้งหมด 5569 เอกสาร อาจจะมีปัญหาในเรื่องของความถูกต้องของการดำเนินการ
2. การจัดการข้อมูล หรือ โหลดข้อมูลทำได้ไม่ครบถ้วน เนื่องจากรูปแบบข้อมูลของ dump wiki มีขนาดใหญ่และดู format ทั้งหมดได้ยาก มีข้อมูลที่ตกหล่นบ้าง
3. รูปแบบคำและโครงสร้างภาไทยค่อนข้างซับซ้อน รวมถึงความหมายที่หลากหลาย ทำให้ไม่สามารถหาคำตอบที่กระจายตัวจากรูปแบบคำถามได้ เช่น คำถาม “ใครคือลูกของต๊ก ศิริพร”
   โดยเอกสารคำตอบ ที่มีคำตอบอยู่ คือ “มีเพลงที่ได้รับความนิยม เช่น *ฉันไม่ใช่นางเอก*, *ธุรกิจเธอ*, *หมดห่วง*, *ไม่มีฝีมือ* ต่อมาจนถึงปัจจุบันได้หันมาเปลี่ยนบทบาทด้วยการแสดงแนวตลก ชีวิตส่วนตัวได้สมรสกับนุ้ย เชิญยิ้ม (ชูเกียรติ เอี่ยมสุข) นักแสดงตลก ทั้งคู่มีบุตรชายด้วยกันคนเดียวชื่อภูสิษฐ์ เอี่ยมสุข (ชื่อเล่น ภู)” จะเห็นได้ว่า รูปแบบในเอกสารจะคำตอบจะอยู่ท้ายประโยค ขณะที่บางเอกสาร คำตอบจะอยู่ต้นประโยค
4. การตัดคำที่เป็นชื่อบุคคล หรือชื่อเฉพาะอื่นๆ ที่ไม่ได้เป็นชื่อหัวข้อจะทำได้ยาก pythainlp ยังไม่สามารถระบุ NER ได้ครบ หรือข้อมูลสำหรับการทำ NER มีไม่มากคลอบคลุม และการจัดเก็บหมวดหมู่ของ Wikipedia ไม่เป็นระเบียบ มีชุดข้อมูลสำหรับแต่ละหมวดหมู่หลายชุด (เช่น ชุดข้อมูลสถานที่ จะมี หมวดหมู่:สถานที่ , สถานที่ เป็นต้น)

สรุปผล

สามารถหาคำตอบสั้นๆจากบางคำถามได้ ที่มีรูปแบบตามเนื้อหาเอกสาร ได้ประยุกต์การนำ regular expression มาช่วยในการค้นหาคำตอบ วิธีการใช้ json ในการเก็บข้อมูลที่ทำการเตรียมไว้มาใช้ได้ในรอบถัดไป ได้ใช้คำสั่ง unix ในการอ่านและเขียนไฟล์เพื่อความสะดวกรวดเร็ว รวมถึงแนวคิดในการจัดการคำ รูปแบบต่างๆ