

ATJPROJECT

การวิเคราะห์ข้อมูลจาก แพลตฟอร์มทวิตเตอร์

DATA ANALYSIS FROM TWITTER
PLATFORM



Outline

- บทที่ 1 **บทนำ**
- บทที่ 2 **ทฤษฎีและงานวิจัยที่เกี่ยวข้อง**
- บทที่ 3 **วิธีการดำเนินงาน**
- บทที่ 4 **ผลการดำเนินงาน**
- บทที่ 5 **สรุปผลและข้อเสนอแนะ**



ATJPROJECT

บทที่ 1



ที่มาและความสำคัญของ ปัญหา

ปัจจุบันข้อมูลมีความสำคัญที่สามารถนำมาพัฒนาและทำให้องค์กรสามารถเติบโตไปข้างหน้าได้อย่างก้าวไกล สามารถสร้างกำไรและใช้เป็นส่วนหนึ่งที่จะช่วยให้การตัดสินใจในการทำกิจกรรมต่างๆ เป็นไปได้อย่างราบรื่นและช่วยลดต้นทุนการใช้ทรัพยากรได้ค่อนข้างสูง

ซึ่งปัญหาที่ผู้จัดทำมองเห็นคือเราจะได้ข้อมูลจำนวนมากได้ที่ไหน และเราจะทำยังไงให้ข้อมูลที่ได้นั้นสามารถแสดงออกมาให้ผู้ใช้งานนั้นสามารถอ่านเข้าใจได้ง่าย

โดยเราได้นำข้อมูลการโพสต์ของผู้ใช้งานบนแพลตฟอร์มที่มีจำนวนผู้ใช้งานที่ค่อนข้างสูงมาเก็บไว้เป็นชุดข้อมูลและนำมาจัดกลุ่มคำก่อนนำมาแสดงเป็นกลุ่มคำและกราฟแท่ง



วัตถุประสงค์โครงการ

สำรวจและเก็บข้อมูลตัวของผู้ใช้งานบนทวิตเตอร์ที่ได้โพสต์ไว้ใน
หัวข้อของ แฮชแท็ก (Hashtag) และเก็บไว้ในชุดข้อมูล
วิเคราะห์ข้อมูลตัวอักษรภายในโพสต์ของผู้ใช้งานโดยการนำคำพูดที่
อยู่ภายในโพสต์มาจัดเรียงเพื่อหาคำหรือประโยคที่มีการพูดถึงมาก
ที่สุด จัดกลุ่มคำและแสดงผลลัพท์
วิเคราะห์ข้อมูลจำนวนบัญชีผู้ใช้งานที่มีการกด รีทวิต(Retweet)
ในแต่ละหัวข้อ ของแฮชแท็ก(Hashtag) และนำผลการวิเคราะห์
มาจัดเรียงกันในรูปแบบกราฟ



ขอบเขตโครงการ

ขอบเขตของการเก็บรวบรวมข้อมูล

เก็บรวบรวมข้อมูลการโพสต์ของผู้ใช้งานผ่าน
หัวข้อ แฮชแท็ก(Hashtag) ซึ่งมีจำนวนการเก็บข้อมูล
อย่างจำกัดเนื่องจากผู้ทำโครงการไม่ได้ใช้งาน API ใน
ระดับ Enterprise

ข้อมูลที่ได้มาจากการสำรวจโพสต์ของผู้ใช้งาน
ประกอบไปด้วย วันที่และเวลาโพสต์เนื้อหาภายใน
โพสต์ ,หัวข้อที่เกี่ยวข้อง,จำนวนคนที่บอกรัก ,จำนวน
คนที่กดชอบโพสต์

ขอบเขตของการนำข้อมูล มาใช้ในการวิเคราะห์

นำข้อมูลที่ได้จากการสำรวจมาวิเคราะห์หาคำที่มีความ
นิยมจากผู้ใช้งานมากที่สุดน้อยที่สุดจัดเรียงตามลำดับ
การกรองคำหรือประโยคที่ไม่สามารถใช้งานได้และ
โพสต์ที่มีข้อมูลที่ซ้ำกันทุกคำภายในโพสต์
ข้อมูลเหล่านี้ไม่มีประโยชน์ในการนำมาวิเคราะห์และแสดง
ผล

ขอบเขตของการนำข้อมูลที่ได้จากการวิเคราะห์ ข้อมูลจากทวิตเตอร์มาแสดงผลลัพท์

การแสดงผลข้อมูลและผลลัพธ์จากการวิเคราะห์คือ
การนำคำหรือประโยคมาแสดงในรูปแบบของ
Word cloud คือการแสดงคำที่มีการใช้มากที่สุดจัด
ลำดับไปหาคำที่มีการกล่าวถึงน้อยที่สุด

ข้อมูลที่แสดงผลออกมาต้องเป็นในรูปแบบคำ
ประโยคหรือตัวอักษรที่มนุษย์สามารถอ่านได้

นำข้อมูลที่วิเคราะห์แล้วมาจัดเรียงแสดงผลเป็นรูป
แบบกราฟในช่วงเวลา แฮชแท็ก(Hashtag) แต่ละหัวข้อมี
ระดับความนิยมแตกต่างกันมากน้อยเพียงใด

ประโยชน์ที่คาดว่าจะได้รับ

- 1.เพื่อให้ได้ข้อมูลความนิยมของกลุ่มประชากรที่ใช้งาน ทวิตเตอร์มีการพูดคุยหรือสนใจสิ่งใดในขณะนั้น
- 2.เพื่อเป็นเครื่องมือช่วยตัดสินใจในการทำกิจกรรมหรือธุรกิจต่างๆ
- 3.สามารถเปรียบเทียบข้อมูลความแตกต่างระหว่าง แฮชแท็กได้อย่างชัดเจนด้วยกราฟ

บทที่ 2

ทฤษฎีและงานวิจัยที่ เกี่ยวข้อง



ทฤษฎีที่เกี่ยวข้อง ประกอบด้วย



1. ประวัติของ ทวิตเตอร์
2. ข้อมูลสำคัญและการใช้งาน API ทวิตเตอร์
3. ภาษา Python
4. Bootstrap framework
5. ภาษา HTML
6. ภาษา CSS
7. ภาษา Javascript
8. โปรแกรม Jupyter Notebook

ประวัติและความเป็นมาของ Twitter

เรื่องราวเริ่มขึ้นในปี 2006 บริษัท ODEO ซึ่งทำธุรกิจเกี่ยวกับพอดแคชซึ่งกิจการกำลังย่ำแย่ จึงต้องสร้างสิ่งใหม่ขึ้นมาเพื่อกอบกู้สถานะการณ หลังจากที่ผ่านมาการประชุมแล้ว Jack Dorsey ได้เสนอไอเดียใหม่เป็นการแชร์ข้อความสั้นๆคล้ายกับ SMS แต่จะเป็นการแชร์บนโลกออนไลน์ โดยที่เราจะทำการโพสต์สถานะของเราแล้วเพื่อนของเราก็จะสามารถเห็นได้ ซึ่งในแบบแรกจัดทำมาในรูปแบบเว็บไซต์ก่อนใช้ชื่อว่า Twtter.com ซึ่งย่อมาจากคำว่า Twitter จากนั้นวันที่ 22 มีนาคม 2549 หลังจากนั้น Jack Dorsey ได้ทวีตข้อความแรกว่า “Just setting up my twttr ” และหลังจากนั้น ทวิตเตอร์ก็แพลตฟอร์มที่ได้รับความนิยมเรื่อยมาและทวิตเตอร์ก็มีการพัฒนาอยู่ตลอดซึ่งเป็นที่มาของความนิยม





API ของทวิตเตอร์และวิธีใช้งาน

API หรือ Application Programming

API ทำหน้าที่ช่วยในการเข้าถึงข้อมูลได้ง่าย โดยการได้รับ API ซึ่งเป็นกุญแจจะสามารถทำให้เข้าถึงข้อมูลของแพลตฟอร์มหรือเว็บไซต์ที่เป็นเจ้าของข้อมูลนั้นๆ



ภาษา Python

เป็นภาษาเขียนโปรแกรมระดับสูงได้รับความนิยมค่อนข้างมากใน
หมู่นักเขียนโปรแกรม
เพราะรูปแบบโครงสร้างของภาษาที่ค่อนข้างง่าย และมีพื้นฐาน
ภาษาอังกฤษ
ก็สามารถเรียนรู้และเข้าใจได้อย่างรวดเร็ว และยังมีคุณสมบัติ
เป็นการเขียนโปรแกรมหลายรูปแบบ
เช่น การเขียนโปรแกรมเชิงวัตถุ การเขียนโปรแกรมแบบ
ฟังก์ชัน
การเขียนโปรแกรมแบบขั้นตอน และยังมีไลบรารีจำนวนมากที่
ครอบคลุมการทำงานที่หลากหลาย
ไม่ว่าจะเป็นการทำงานในรูปแบบ IoT หรือ
การสร้างเซิร์ฟเวอร์ที่ใช้ในการเชื่อมต่อส่วนของ Backend และ
Frontend



Bootstrap framework

Bootstrap คือชุดคำสั่งที่ประกอบด้วยภาษา CSS, HTML และ Javascript เป็นชุดคำสั่งที่ถูกพัฒนาขึ้นมาเพื่อกำหนดกรอบหรือรูปแบบการพัฒนาเว็บไซต์ในส่วนของการปฏิสัมพันธ์กับผู้ใช้งานเว็บไซต์
เราจึงสามารถเรียก Bootstrap ว่าเป็น Front-end framework คือใช้สำหรับ พัฒนาเว็บไซต์ส่วนการแสดงผล



ภาษา HTML

ภาษาที่ใช้สำหรับเขียน website ซึ่ง HTML5 นี้เป็นภาษาที่ถูกพัฒนาต่อมาจากภาษา HTML4 และพัฒนาขึ้นมาโดย WHATWG (The Web Hypertext Application Technology Working Group) HTML 5.0 เป็นมาตรฐานของภาษา HTML ที่อยู่ในระหว่างการพัฒนาภาษา มาร์กอัป สำหรับ รุ่นต่อไป ได้ออกเผยแพร่การใช้งานเมื่อวันที่ 22 มกราคม พ.ศ.2551 โดยภาษา HTML 5.0 มีรูปแบบสองแบบที่ร่วมกันคือ แบบดั้งเดิม และแบบที่สองคือ XHTML ที่ใช้ในการจัดโครงสร้างและการแสดงผลของเนื้อหาสำหรับเวิลด์ไวด์เว็บ



ภาษา CSS

การจัดทำเว็บไซต์ที่มีประสิทธิภาพต้องมีการวางแผนและการออกแบบระบบที่ดี โดยในยุคแรก ๆ จะใช้ภาษา HTML ในการจัดทำระบบการแสดงผลทางด้านโครงสร้างและข้อมูลของเว็บ แต่ปัจจุบันมีการพัฒนามาจนถึง HTML5 และยังมีการพัฒนาภาษาที่ใช้เพื่อกำหนดรูปแบบการแสดงผลทางหน้าเว็บไซต์ที่หลากหลายและมีความยืดหยุ่น เช่น สีอักษร สีพื้นหลัง ขนาดตัวอักษร จัดการเลย์เอ๊าท์ ให้สวยงามและอื่นๆ ซึ่งนั่นก็คือ CSS หรือ Style Sheets และในความหมายของทางโปรแกรมเมอ์นั้น คือ โครงสร้างการแสดงผลของหน้าตาเว็บไซต์



ภาษา Javascript

จาวาสคริปต์ (JavaScript) เป็นภาษาสคริปต์ ที่มีลักษณะการเขียนแบบโพรโทไทป์(Prototyped-based Programming) ส่วนมากใช้ในหน้าเว็บเพื่อประมวลผลข้อมูลที่ฝั่งของผู้ใช้งาน แต่ก็ยังมีใช้เพื่อเพิ่มเติมความสามารถในการเขียนสคริปต์โดยฝังอยู่ในโปรแกรมอื่น ๆ

The image shows the letters 'JS' in a bold, black, sans-serif font, centered on a bright yellow square background. This is a common logo for JavaScript.



โปรแกรม Jupyter Notebook

Jupyter Notebook คือหน้าเว็บไซต์ที่ทำงานด้วยเครื่องคอมพิวเตอร์ของตนเอง
โดยหน้าเว็บที่ใช้ในการเขียนโค้ดจะเปิดขึ้นเมื่อเปิดแฟ้มงานและ
สั่งการทำงานในแฟ้มงานภายในเครื่อง
หน้าเว็บรูปแบบจะเป็น กล่องข้อความที่ให้ใส่
โค้ดที่ต้องการจะเขียนลงไปและสามารถสั่งทำงานได้ทันทีเพื่อให้
แสดงผลลัพธ์ของโค้ดที่เขียนลงไป



งานวิจัยที่ เกี่ยวข้อง



การระบุตัวผู้เขียนข้อความออนไลน์ ภาษาไทยด้วยซอฟต์แวร์ แมชชีนและต้นไม้ตัดสินใจ

ระบุตัวผู้เขียนข้อความนิรนาม
โดยวิธีการที่เลือกใช้คือการจำแนกด้วยซอฟต์แวร์
แมชชีนและต้นไม้ตัดสินใจเมื่อทดสอบกับข้อความ
ขนาดสั้น



DETECTION OF SPAMMERS IN TWITTER APPLICATIONS

ตรวจจับสแปมในทวิตเตอร์ แอปพลิเคชันหรือทวิตที่
มีการสแปมเกิดขึ้น



การพัฒนาโมเดลการจำแนกความรู้สึก ของข้อความภาษาไทยโดยใช้เทคนิค การเรียนรู้ด้วยเครื่อง

ๆ การวิเคราะห์ความรู้สึกจากข้อความ
สามารถแบ่งแยกข้อความที่แสดงความรู้สึกทางด้าน
บวก ทางด้านลบ
และไม่แสดงความรู้สึกหรือเป็นกลางออกจากกันได้จะ
เป็นประโยชน์ในการควบคุมคุณภาพสินค้าหรือรักษา
คุณภาพการให้บริการ



ระบบเฝ้าติดตามความเห็นบนสื่อสังคม ออนไลน์สำหรับมหาวิทยาลัย

รวบรวมข้อมูลความคิดเห็นบนทวิตเตอร์ที่กล่าวถึง
มหาวิทยาลัยไทย
โดยเลือกมหาวิทยาลัยตัวอย่างจาก ภาครัฐและเอกชน
อย่างละ 20 แห่ง



การพัฒนาโมเดลตรวจจับคำหยาบภาษา ไทยบนสื่อออนไลน์ด้วยเทคนิคดาต้าไมน์ นิ่ง

เปรียบเทียบกระบวนการวิเคราะห์คำหยาบและคำไม่
สุภาพภาษาไทยบนสื่อออนไลน์ด้วยเทคนิคดาต้าไมน์นิ่ง

บทที่ 3

วิธีการดำเนินงาน



การศึกษาความเป็นไปได้ของโครงการ

1

บุคลากร

สมาชิกภายในทีม

- 1) นายเจษฎา ทิพย์แสง
- 2) นายรัชนาทลักษณ์
- 3) นายธีระพงษ์ วีระพงษ์สุชาติ

3

เครื่องมือที่ใช้ในการดำเนินการประกอบด้วย

- 1) Jupyter notebook
- 2) Visual studio code 1.50
- 3) PyCharm Community Edition 2020.2.3
- 4) Python flask 1.1.2
- 5) Bootstrap framework 4

2

ระยะเวลาในการดำเนินงานใช้ระยะเวลาดำเนินการทั้งสิ้นจำนวน 10 เดือน

- ม.ค. - มี.ค. ศึกษาความเป็นไปได้โครงการ
มี.ค. - พ.ค. รวบรวมข้อมูลความต้องการ
มี.ค. - มิ.ย. การวิเคราะห์และออกแบบระบบ
ส.ค. - ก.ย. ออกแบบส่วนปฏิสัมพันธ์กับผู้ใช้งาน
มิ.ย. - ก.ค. ออกแบบโมเดล
มี.ค. - ต.ค. เขียนและแก้ไขโปรแกรม
ก.ย. - ต.ค. ทดสอบระบบ
ก.ย. - ต.ค. จัดทำเอกสารเข้าเล่ม
ต.ค. ทดสอบการใช้งานโดยอาจารย์ที่ปรึกษา

การเก็บรวบรวมข้อมูล

1

ทวิตเตอร์แพลตฟอร์ม

ผู้พัฒนาได้ทำการเก็บรวบรวมข้อมูลจากสิ่งที่ผู้คนกล่าวถึงมากที่สุด 5 อันดับบนทวิตเตอร์แพลตฟอร์ม

2

งานวิจัยสำหรับการรวบรวมข้อมูลและวิเคราะห์ข้อมูลจากทวิตเตอร์แพลตฟอร์ม

ผู้พัฒนาได้ศึกษาการบวณการวิเคราะห์ข้อมูลและขั้นตอนในบางส่วนเพื่อนำมาพัฒนาในตัวโปรแกรมให้มีคุณภาพมากขึ้น

การวิเคราะห์และออกแบบระบบ

- 1 ภาพรวมการทำงานของระบบ
- 2 การออกแบบกระบวนการทำงานของโปรแกรม
- 3 ลำดับขั้นตอนในการทำงานของเว็บไซต์

1 ภาพรวมการทำงานของระบบ

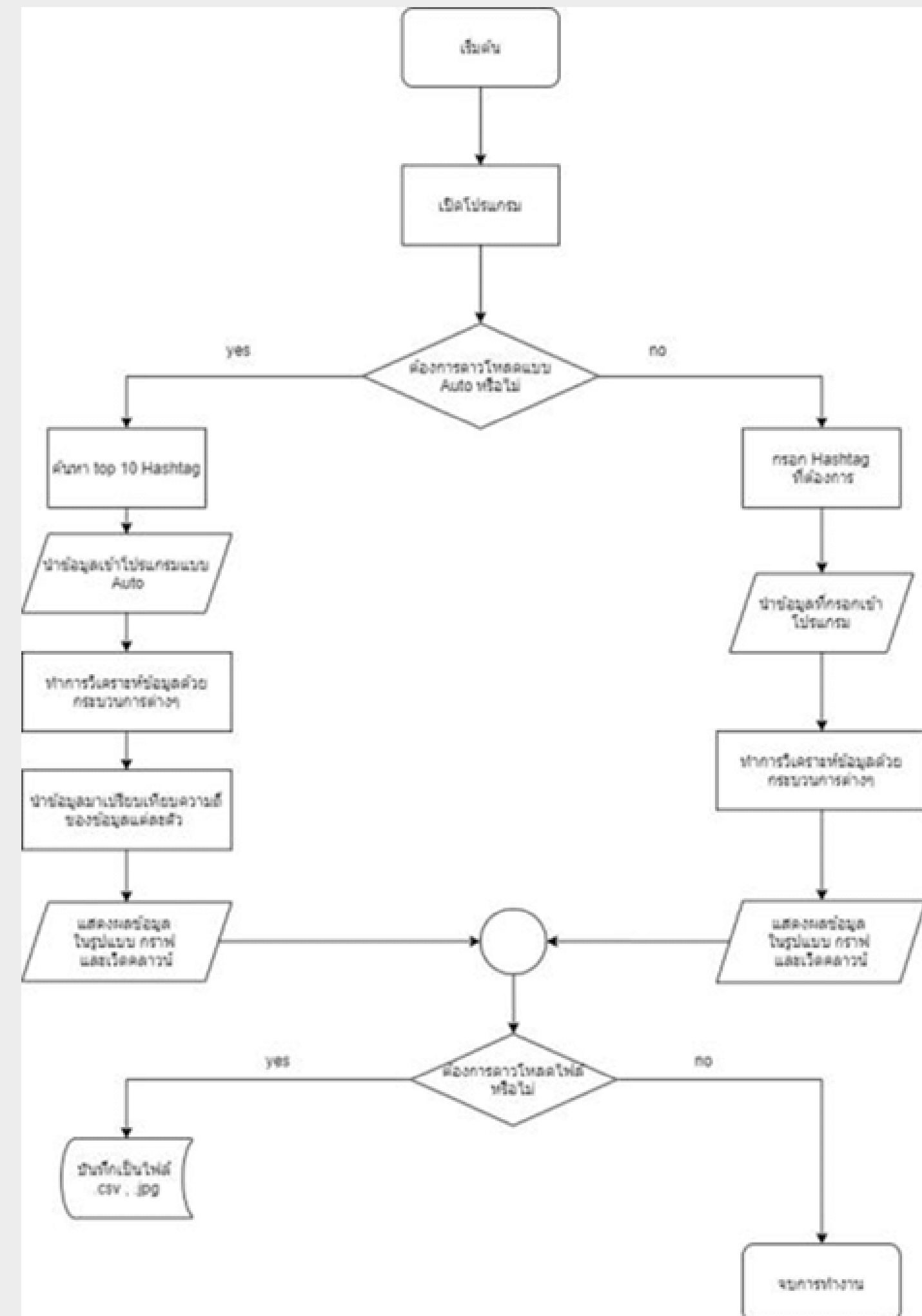
ค้นหา Twitter Hashtag ที่คุณสนใจ

ค้นหา

กดปุ่มด้านล่างค้นหา TOP5 แบบอัตโนมัติ

AUTO SEARCH

name	tweet_volume
#MinimalWarmByLOEY	277481
#ฉันคู่2GetherLiveOnStage	38659
#EnchantedMomentWithYin	214324
#DestinyClinicxOhmFluke	38852
#มีชัย1พฤษภาคม	133517



2

การออกแบบกระบวนการทำงานของโปรแกรม

API

Data Frame

Data set

save local

open Data set

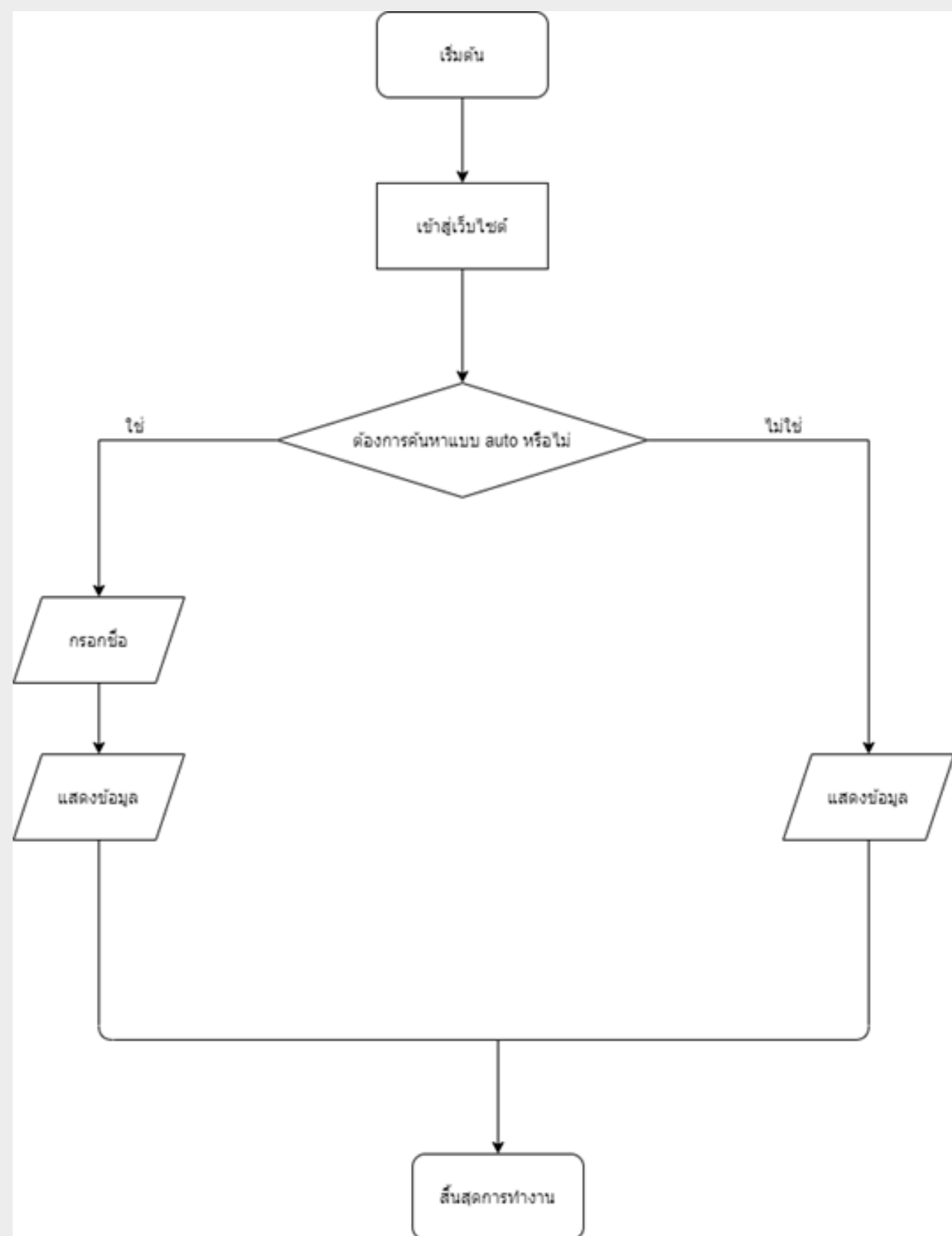
Wordcloud

กระบวนการทำงานมีอยู่ 6 ขั้นตอน ได้แก่

- 1) ดึงข้อมูลจากทวิตเตอร์โดยใช้ API
- 2) จัดเรียงข้อมูลในรูปแบบ DATA FRAME
- 3) บันทึกข้อมูลให้อยู่ในรูปแบบชุดข้อมูล FILE.CSV
- 4) จัดเก็บไฟล์ไว้ในเครื่อง
- 5) ทำการอ่านไฟล์.CSV
- 6) แสดงผลลัพธ์เป็นเว็ดคลาวด์และเก็บไว้ในโฟลเดอร์

3

ลำดับขั้นตอนในการทำงานของเว็บไซต์



ค้นหา Twitter Hashtag ที่คุณสนใจ

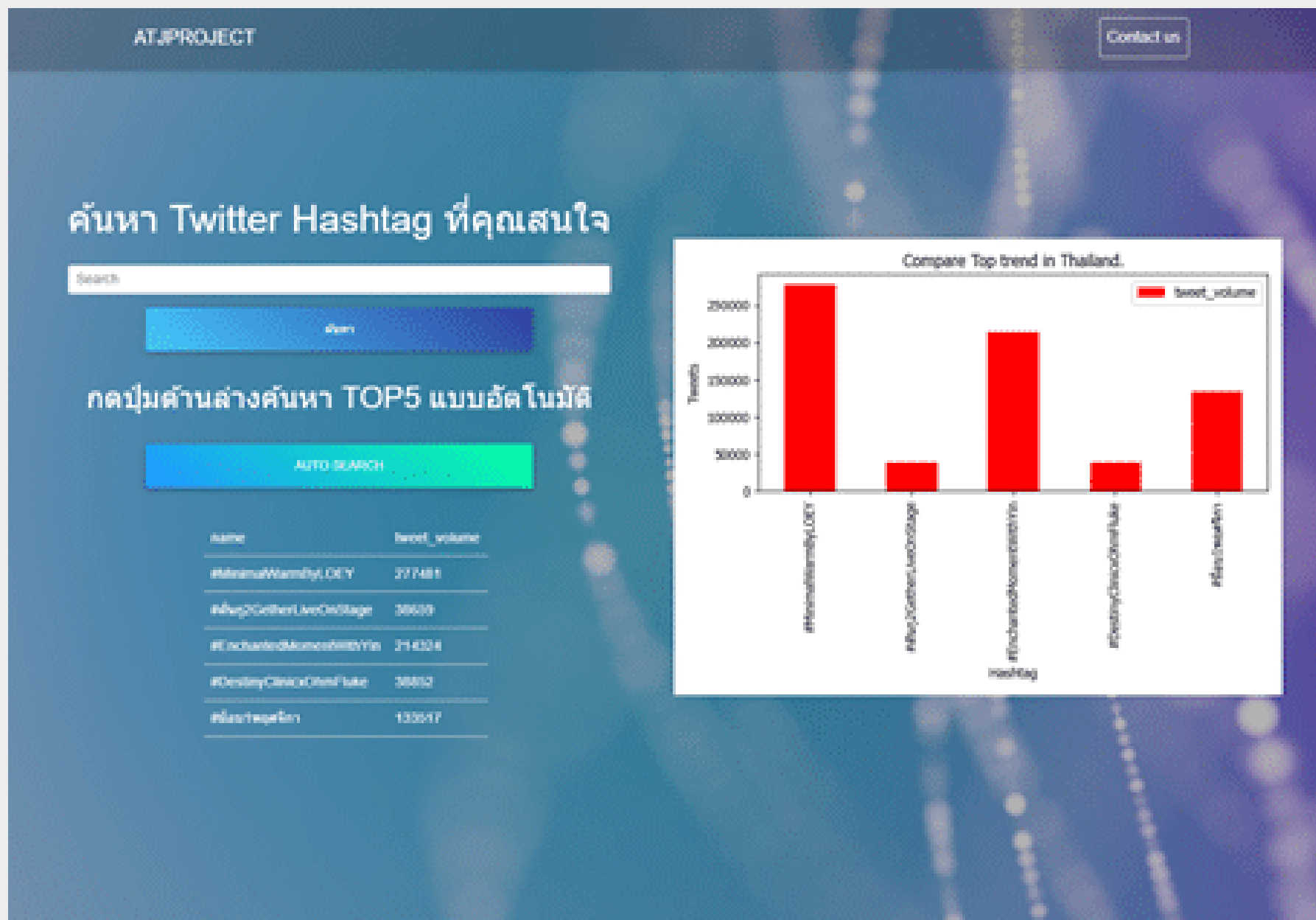
Search

ค้นหา

กดปุ่มด้านล่างค้นหา TOP5 แบบอัตโนมัติ

AUTO SEARCH

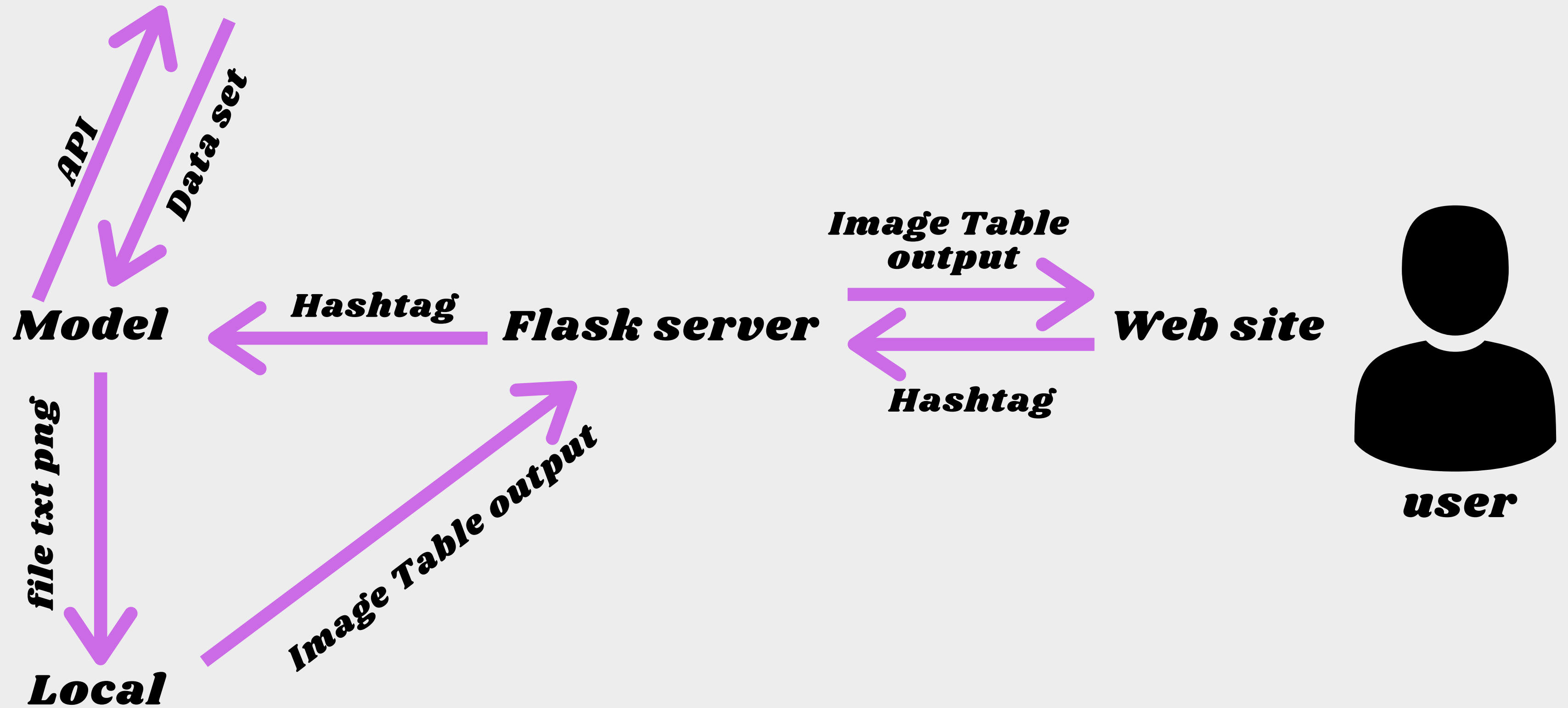
การออกแบบส่วนปฏิสัมพันธ์กับผู้ใช้งาน



ส่วนแสดงผลลัพท์



การออกแบบโมเดล



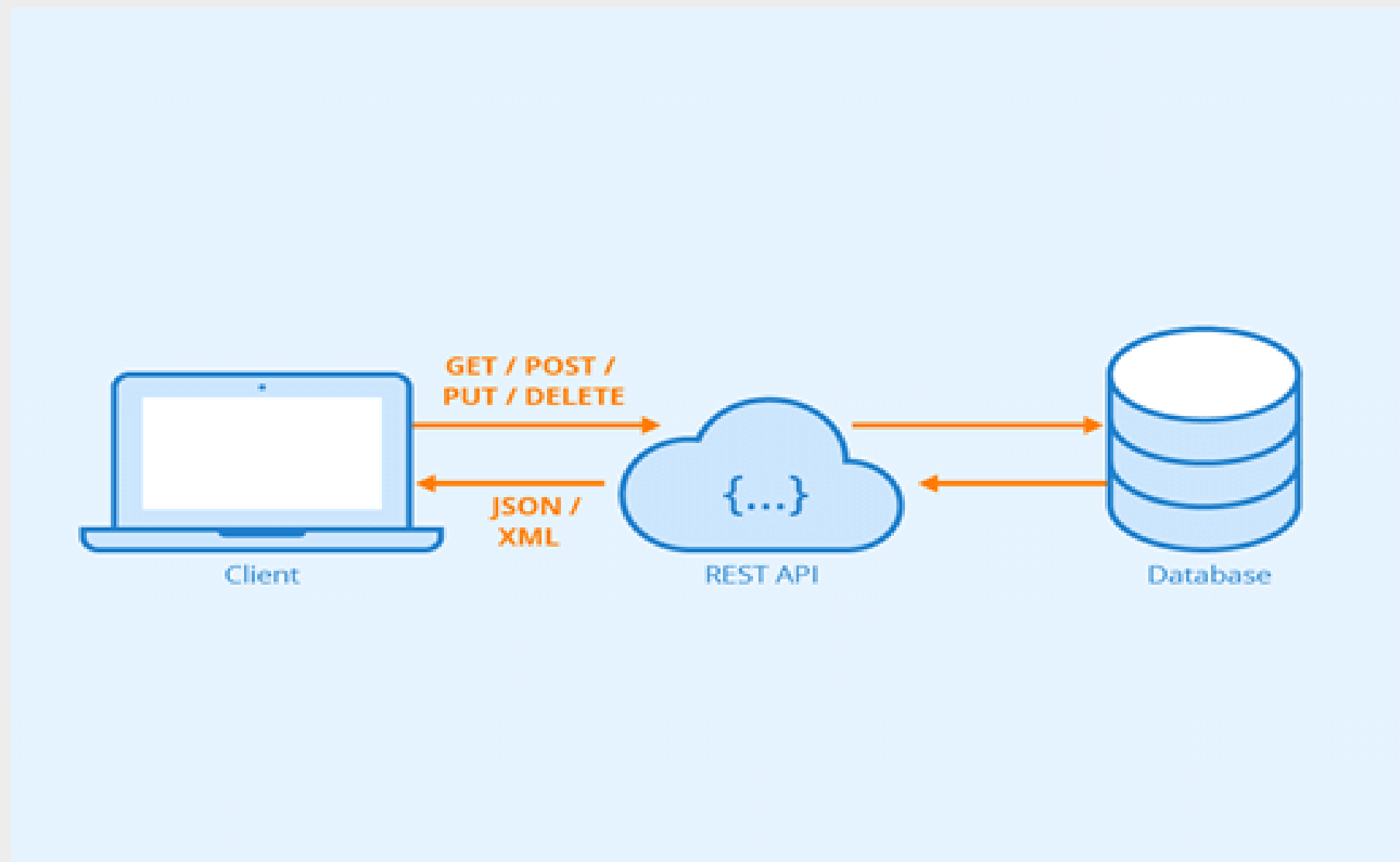
บทที่ 4

ผลการดำเนินงาน



การร้องขอ API จากทาง Twitter

ซึ่งโปรเจกต์นี้มีการใช้ข้อมูลจากTwitter ดังนั้นวิธีที่จะได้ข้อมูลที่ดีที่สุดอาจจะเป็นการใช้ API จากทางเจ้าของแพลตฟอร์มโดยการสมัครAccountทางTwitter แล้วส่งเรื่องขอ API ไปยัง Twitter Developer



หลักการทำงานของโปรแกรมมี 2 รูปแบบ

1 การทำงานแบบ Manual

โดยการทำงานแบบ Manual นั้นให้ User ป้อน Hashtag ที่ต้องการค้นหาลงในช่อง Search แล้วกดปุ่มค้นหาแล้วทางโปรแกรมจะทำการค้นหาข้อมูล Hashtag ตามที่userค้นหาโดยมี Limit ข้อมูลที่ได้รับได้ คือ 1000 ตัว

ค้นหา Twitter Hashtag ที่คุณสนใจ

ค้นหา

2 การทำงานแบบ Auto

โดยการทำงานแบบ Auto นั้นให้ User กดปุ่มค้นหาแบบอัตโนมัติทางโปรแกรมจะทำการค้นหาข้อมูลที่ดี Top Hashtag 5 อันดับและทำการค้นหาโดยการค้นหาแต่ละตัวนั้นมี Limit ข้อมูลที่รับต่อ 1 Hashtag คือ 1000 ตัว

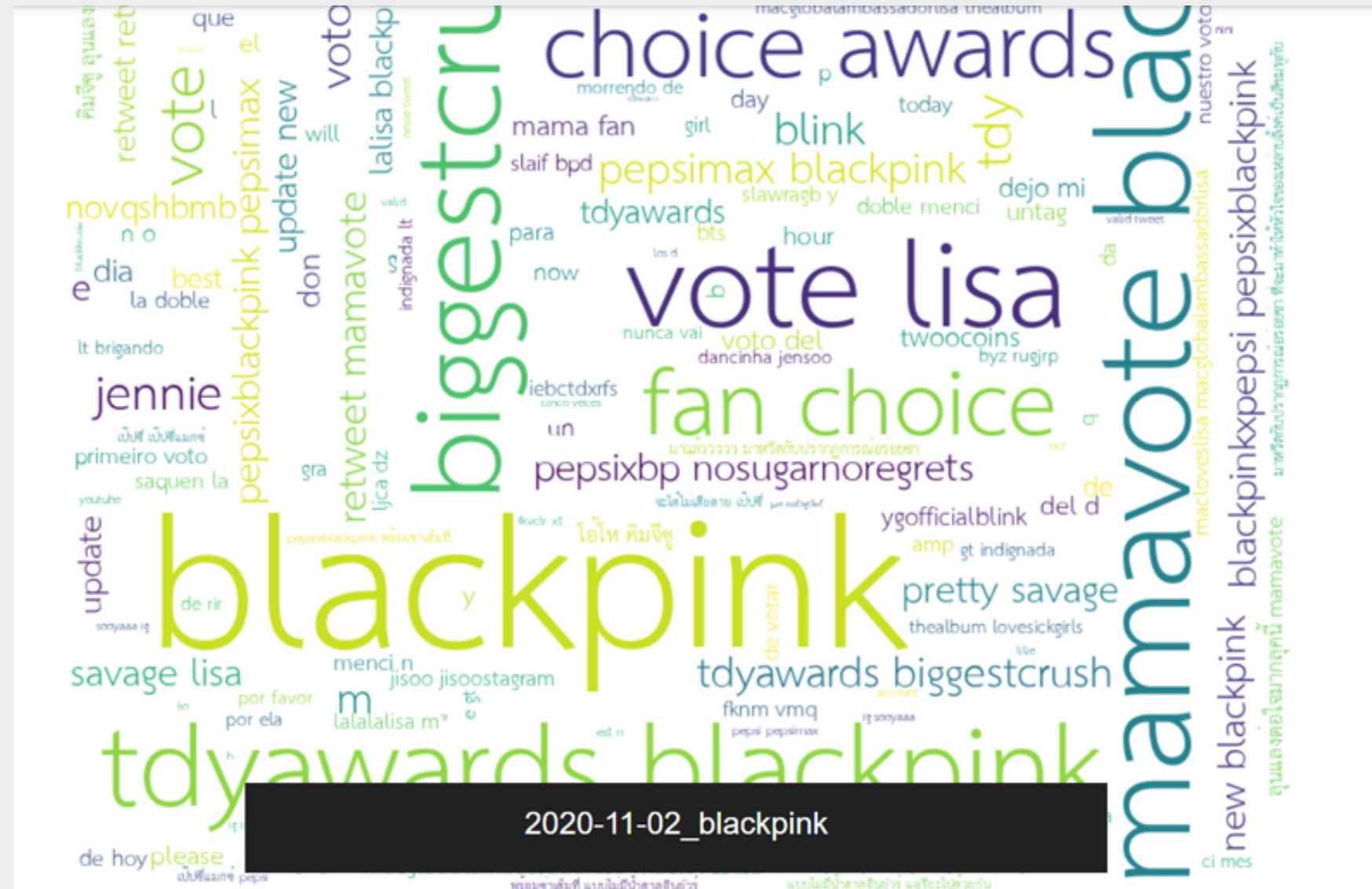
กดปุ่มด้านล่างค้นหา TOP5 แบบอัตโนมัติ

ค้นหาอัตโนมัติ

การแสดงผลการข้อมูลจะมีอยู่ 2 รูปแบบ

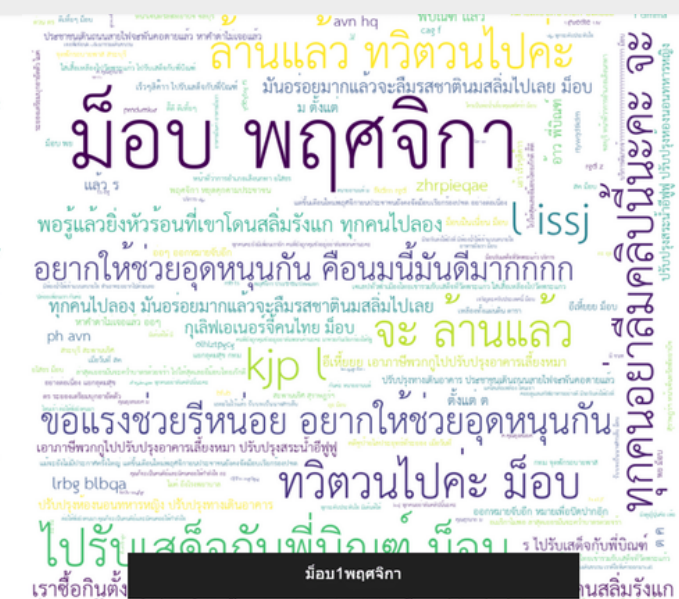
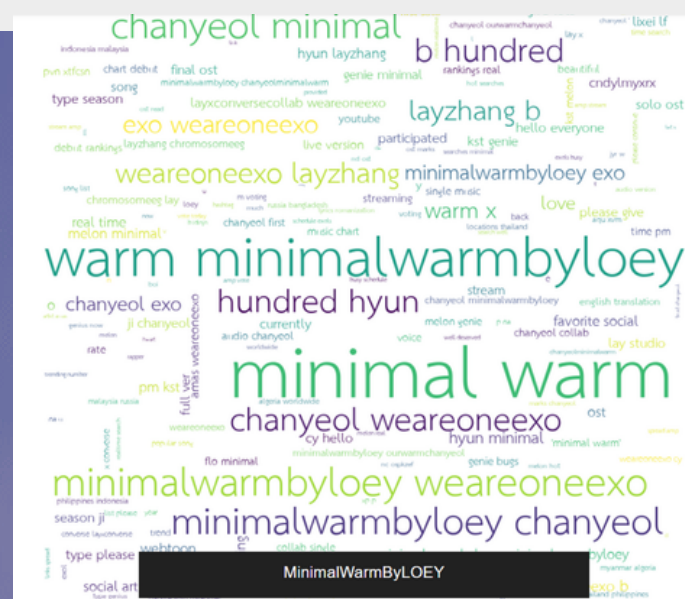
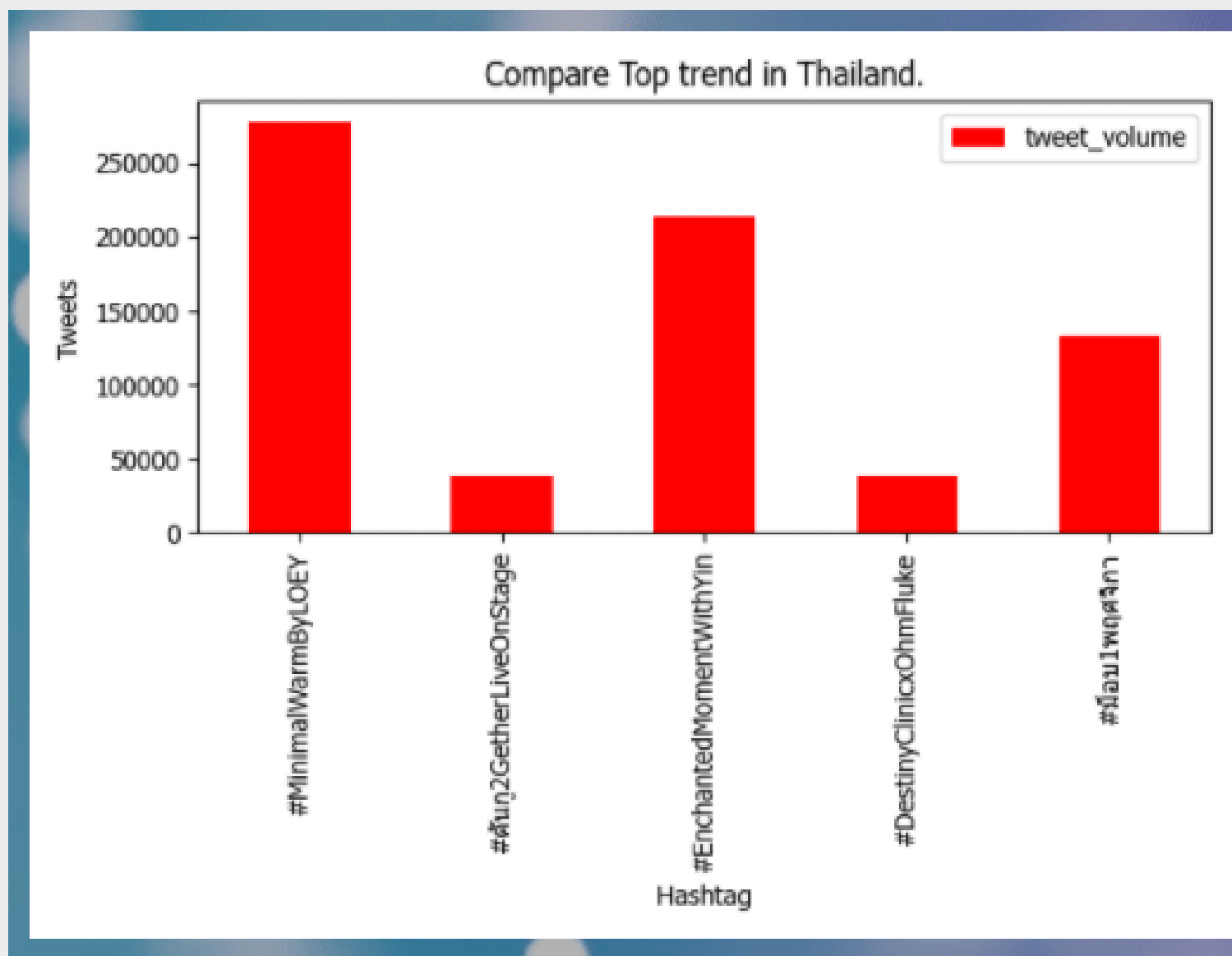
1

การแสดงผลในรูปแบบ Manual จะเป็นการที่ทางโปรแกรมรับค่าและค้นหาตามที่ User ต้องการโดยมี Limit ข้อมูล 1000 ตัวมาแสดงในรูปแบบของ "Word Cloud"



2

การแสดงผลในรูปแบบ Auto จะเป็นการที่ทางโปรแกรม ค้นหาข้อมูลที่ติด Top Hashtag 5อันดับ โดยแสดงในรูปแบบกราฟข้อมูล และ "Word Cloud" ในแต่ละ Hashtag



บทที่ 5

สรุปผลและข้อเสนอแนะ



สรุป

1

การเข้าแอปพลิเคชัน ผู้ทำโครงการได้สร้างหน้าเว็บที่เข้าถึงคือ <https://atjproject.herokuapp.com/> และมีการทำงานเพียง 3 หน้าเท่านั้น

2

เป็นเว็บแอปพลิเคชันที่เชื่อมต่อโดย Python ในรูปแบบ Framework ของ Flask และดึงข้อมูลส่งผ่านไปยังหน้าเว็บแอปพลิเคชัน

3

เป็นการใช้ API จาก Twitter เพื่อดึงข้อมูลที่ต้องการไม่ว่าจะเป็น Hashtag ที่เราต้องการค้นหาโดยเฉพาะ หรือ แบบ Top Hashtag และบันทึกข้อมูลในรูปแบบ Data แล้วนำมาจัดแสดงในรูปแบบ กราฟ หรือ Word Cloud

อุปสรรค และ ปัญหา

- 1.) มีความคลาดเคลื่อนของเวลาของข้อมูลที่ไหลดมา
- 2.) การยื่นเรื่องขอ API Twitter ติดปัญหาเนื่องจากต้องผ่านการกรองที่เข้มงวดจากทางบริษัท Twitter จึงใช้เวลานานจึงจะได้รับการอนุมัติ
- 3.) ปัญหาของ VersionPython ในบางกรณี version ใหม่ยังพบปัญหาอยู่จึงต้องกลับไปใช้ version เก่า
- 4.) มีขอบเขตของการดึงข้อมูล
- 5.) การดึงข้อมูลจากแพลตฟอร์มต้องใช้เวลาค่อนข้างเยอะเพื่อดึงโพสต์มาจากผู้ใช้งานตามเวลา

ข้อเสนอแนะ

- 1.)ถ้ามีงบประมาณจำนวนหนึ่งจะสามารถซื้อแพคเกจรับข้อมูลได้มากขึ้นและลดเวลาการDelayของข้อมูลได้
- 2.)สามารถนำชุดข้อมูลไปใช้งานด้านการวิเคราะห์อื่นๆได้
- 3.)สามารถนำไปต่อยอดในด้านธุรกิจได้