ระบบวิเคราะห์ระดับภาษาด้วยเหมืองข้อความ

THE THAI LANGUAGE LEVEL ANALYSIS SYSTEM WITH TEXT MINING

ผู้จัดทำโครงงาน

นายสาริน สงครินทร์ 613021008-4

นายอภิสิทธิ์ ไกรยะโส 613021008-4

อาจารย์ที่ปรึกษา : อ.ดร.วรัญญา วรรณศรี

วัตถุประสงค์

1. เพื่อสร้างชุดข้อมูล สำหรับนำไปวิเคราะห์กับโมเดล

2. เพื่อสร้างโมเดล สำหรับวิเคราะห์ประโยคว่าเป็นทางการหรือไม่เป็นทางการ

3. เพื่อเปรียบเทียบประสิทธิภาพโมเดล สำหรับโมเดลที่เหมาะสมกับระบบที่สุด

4. เพื่อสร้างระบบวิเคราะห์ระดับของภาษาไทย

ขอบเขตงานวิจัย

1. ระบบจะให้ผู้ใช้ทำการใส่ข้อความหรือประโยคที่ต้องการจะตรวจสอบ และแสดงผลลัพธ์ในรูปแบบข้อความ
2. ระบบมีการแนะนำส่วนที่ต้องแก้ไขในประโยคหรือข้อความ

ข้อจำกัด

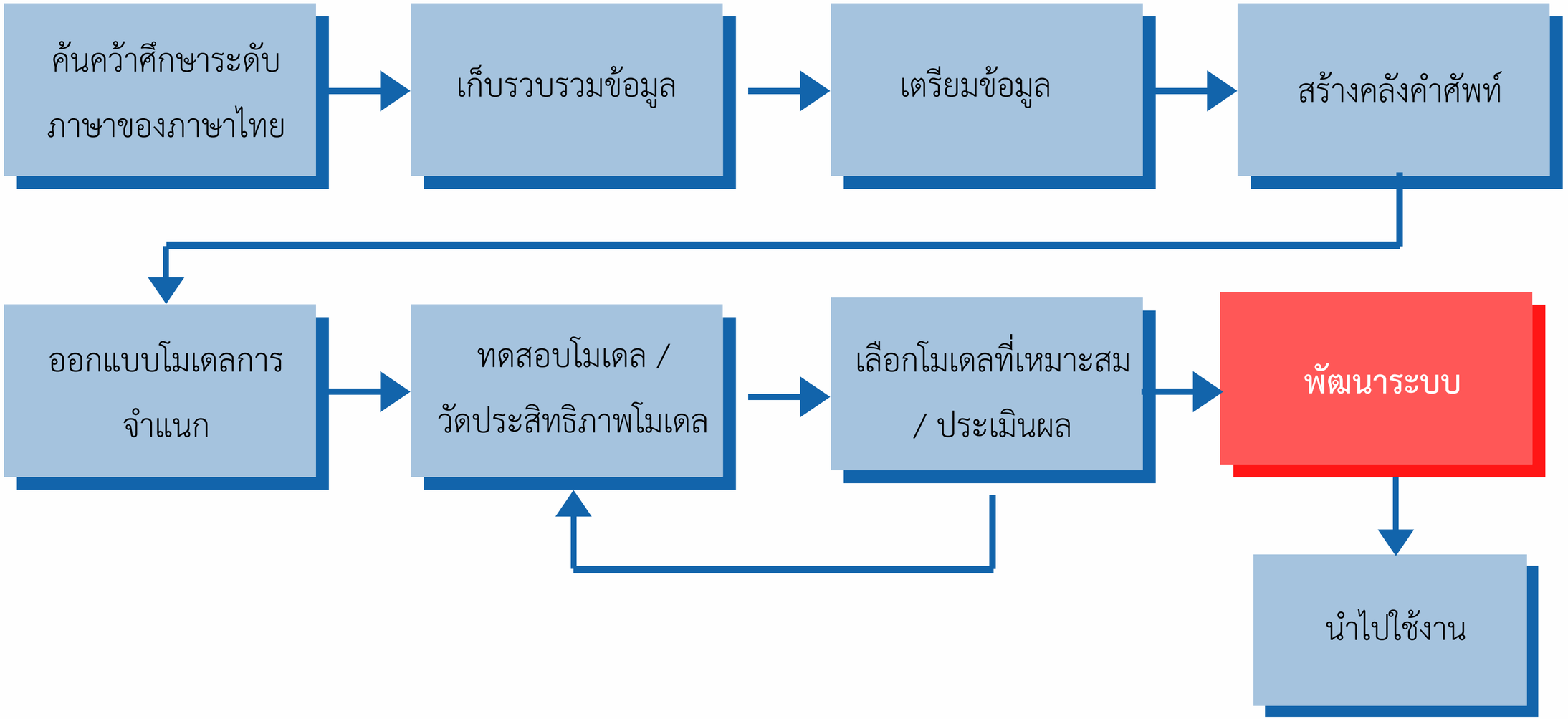
1. จำกัดเฉพาะภาษาไทย
2. ตรวจสอบเฉพาะภาษาที่ใช้ในรายงาน
3. ระบบจะรับข้อมูลเข้าในรูปแบบข้อความเท่านั้น

ความก้าวหน้าจาก Final

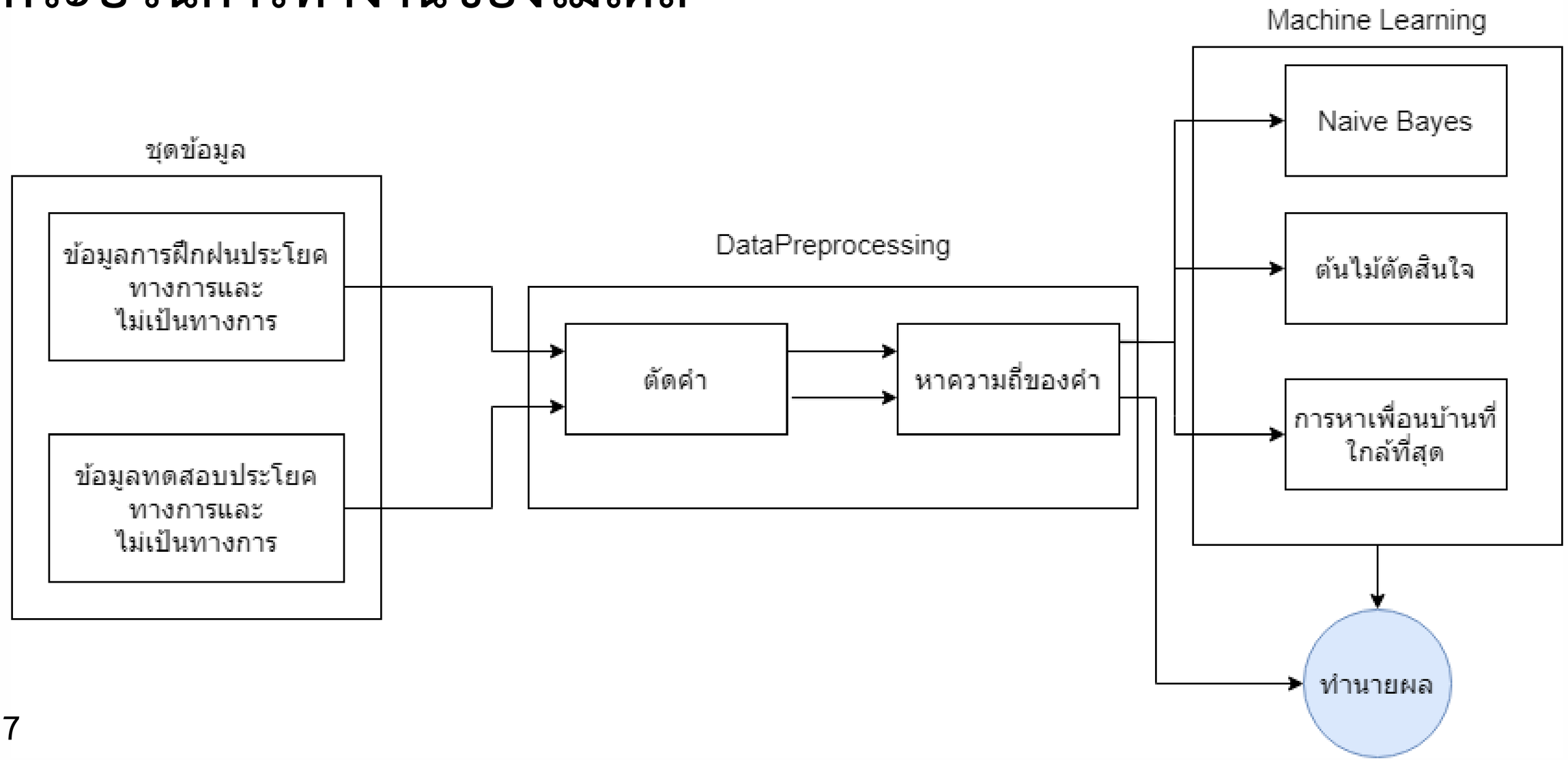
1.พัฒนาโมเดล

2. พัฒนาเว็บแอปพลิเคชัน

ขั้นตอนการดำเนินงาน



กระบวนการทำงานของโมเดล



ชุดข้อมูล

ชุดข้อมูลทั้งหมด 2,000 ประโยคที่จะนำไปใช้ในโมเดลจะประกอบไปด้วย

	ประโยค
ประโยคทางการ	1,000
ประโยคไม่เป็นทางการ	1,000

ตัวอย่าง - ชาวนาในขอนแก่นมีการทำการเกษตรเป็นจำนวนมาก

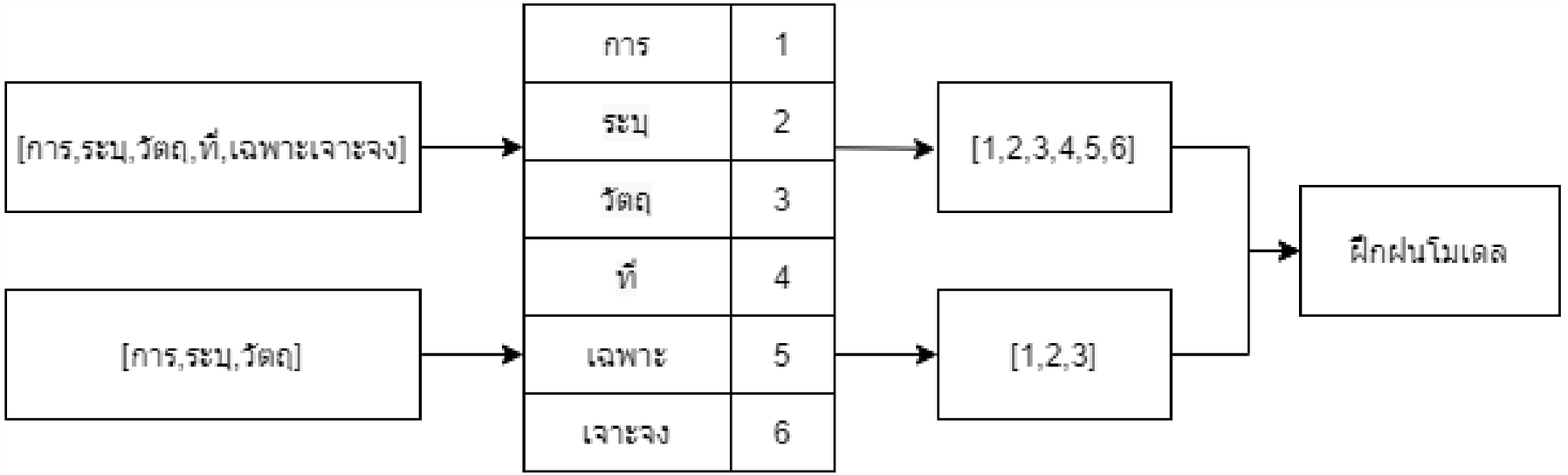
- เกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่สถานที่รับซื้อต่าง ๆ

ตัวอย่างชุดข้อมูล

Sent	Result
ชาวสวนในจังหวัดขอนแก่นมีการทำการเกษตรเป็นจำนวนมาก	**ไม่เป็นทางการ**
เกษตรกรในจังหวัดขอนแก่นมีการทำการเกษตรเป็นจำนวนมาก	**ทางการ**
เกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่โรงรับซื้อต่าง ๆ	**ไม่เป็นทางการ**
เกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่สถานที่รับซื้อต่าง ๆ	**ทางการ**

ตัวอย่างการเตรียมข้อมูล

การระบุวัตถุที่เฉพาะเจาะจง → การ | ระบุ | วัตถุ | ที่ | เฉพาะเจาะจง



โมเดลที่นำมาทดสอบ

- Naive Bayes
- การหาเพื่อนบ้านที่ใกล้ที่สุด
- ต้นไม้ตัดสินใจ

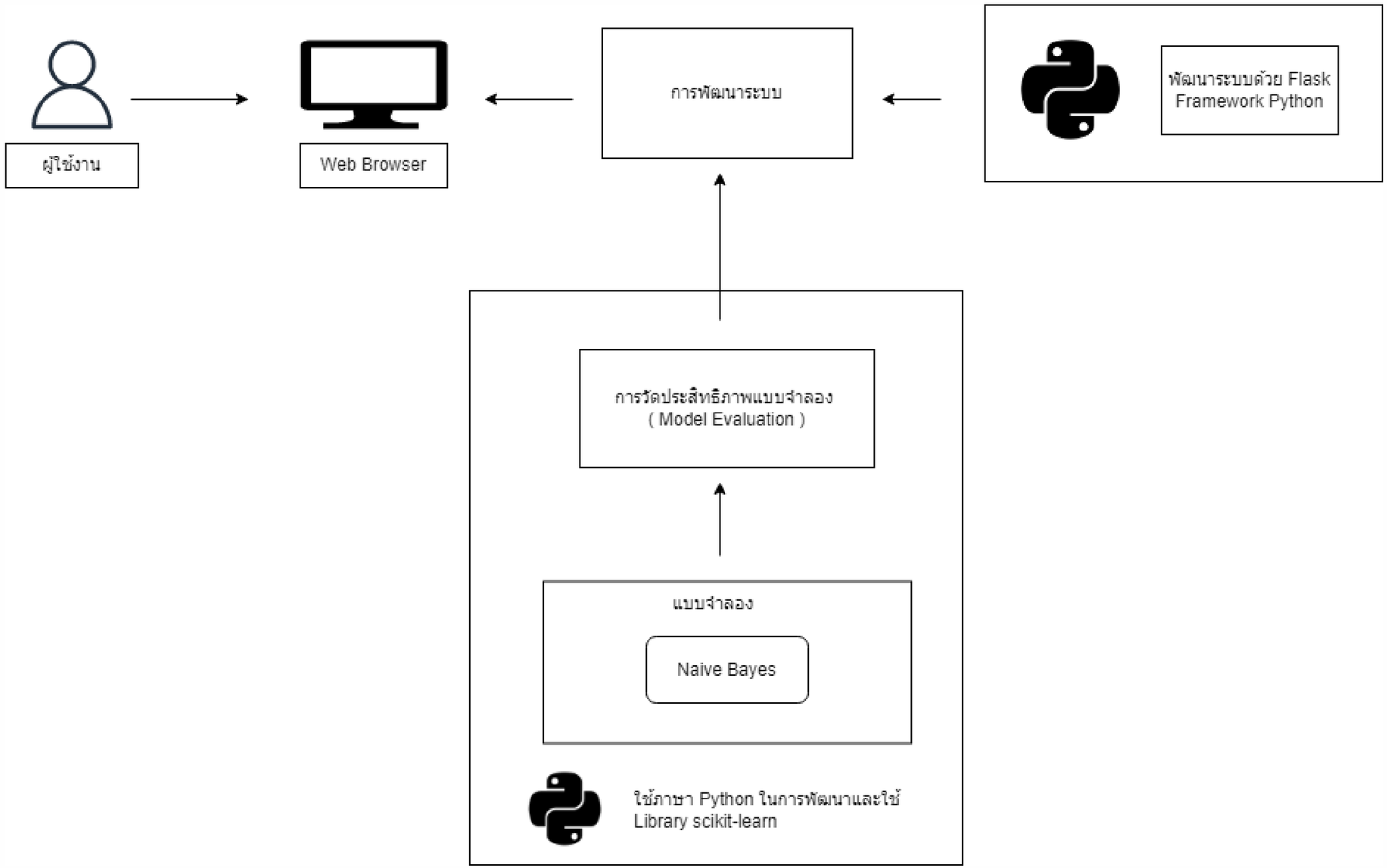
ผลการเปรียบเทียบวัดประสิทธิภาพของระบบด้วย

K-fold Cross Validation

จากการทดสอบ 3 โมเดลสรุปได้ว่า โมเดล Naive Bayes มีประสิทธิภาพสูงที่สุดอยู่ที่ 10-fold ซึ่งมีค่าความแม่นยำอยู่ที่ 75.63%

Model	Accuracy
Naive Bayes	75.63%
ต้นไม้ตัดสินใจ	63.37%
เพื่อนบ้านที่ใกล้ที่สุด (k = 5)	58.37%

สถาปัตยกรรมของระบบ

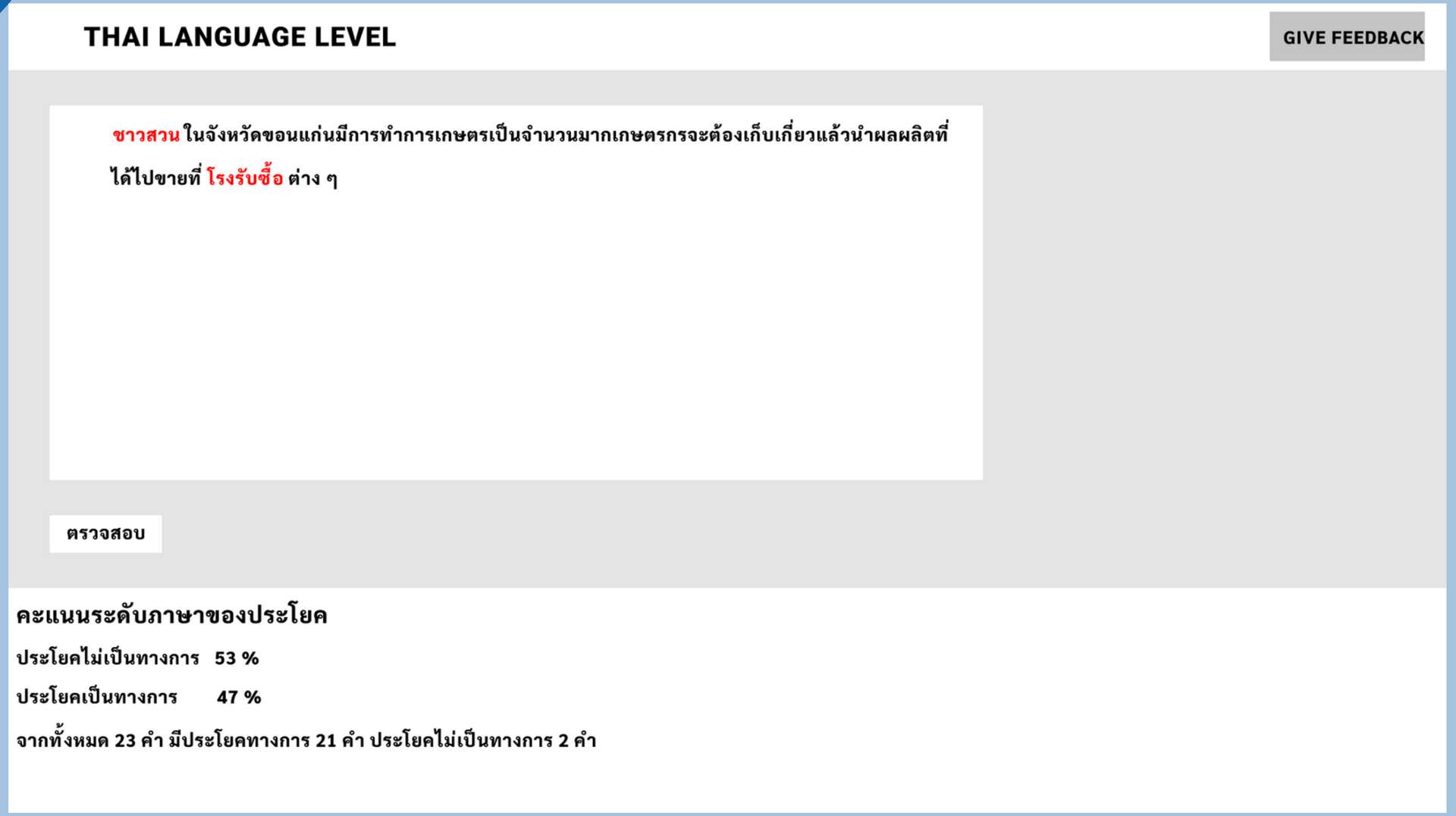


ความสามารถของแอปพลิเคชัน

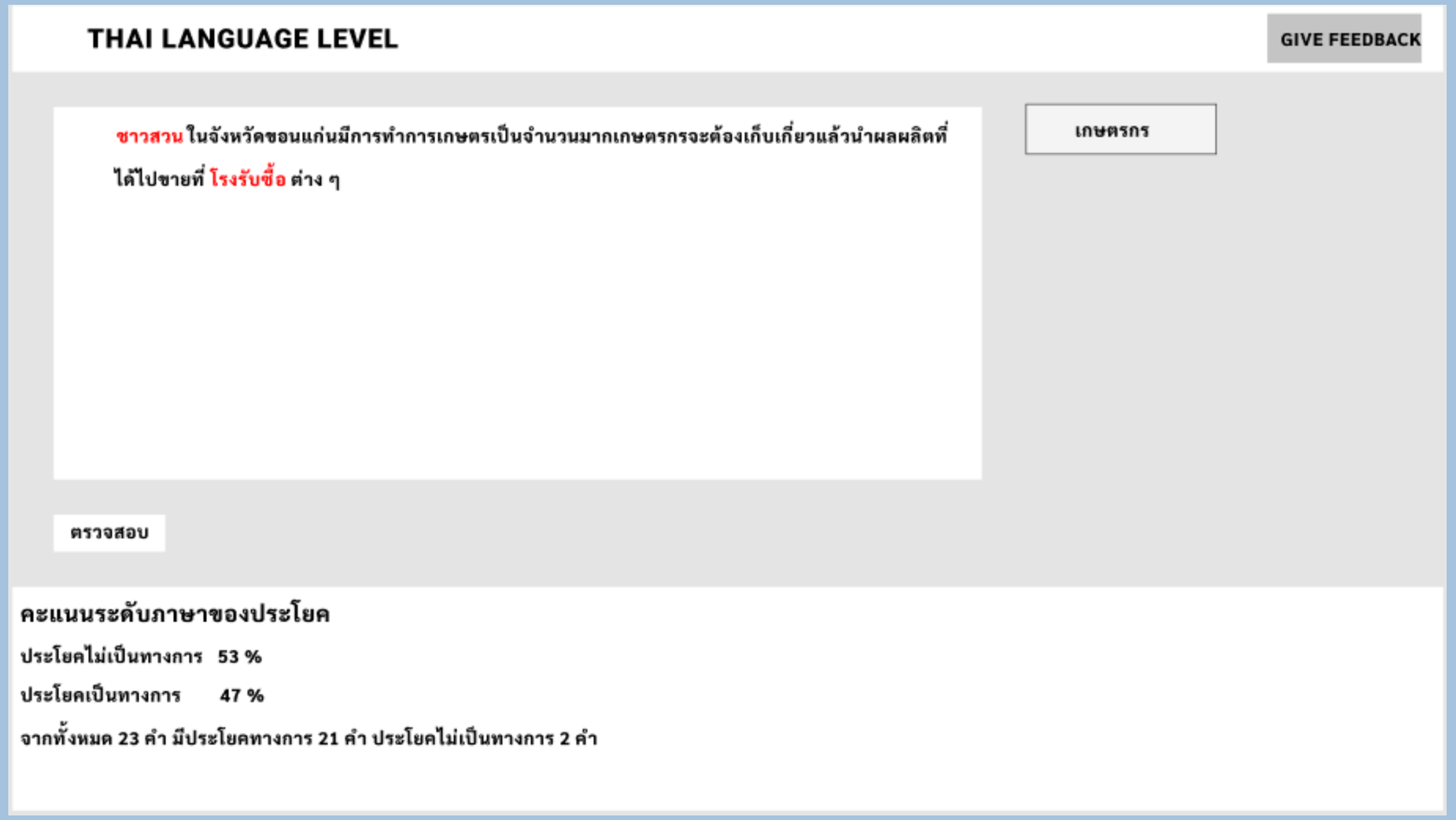
- ฟังก์ชันการตรวจสอบประโยคที่ให้ผลลัพธ์เป็นทางการหรือไม่เป็นทางการ
- ฟังก์ชันการแนะนำคำที่เป็นทางการ
- ฟังก์ชันให้ผู้ใช้เสนอแนะหากเว็บไซต์ตรวจสอบประโยคผิดพลาดหรือแนะนำคำที่ไม่เป็นทางการให้ผู้วิจัยได้ทราบ

ตัวอย่างเว็บไซต์

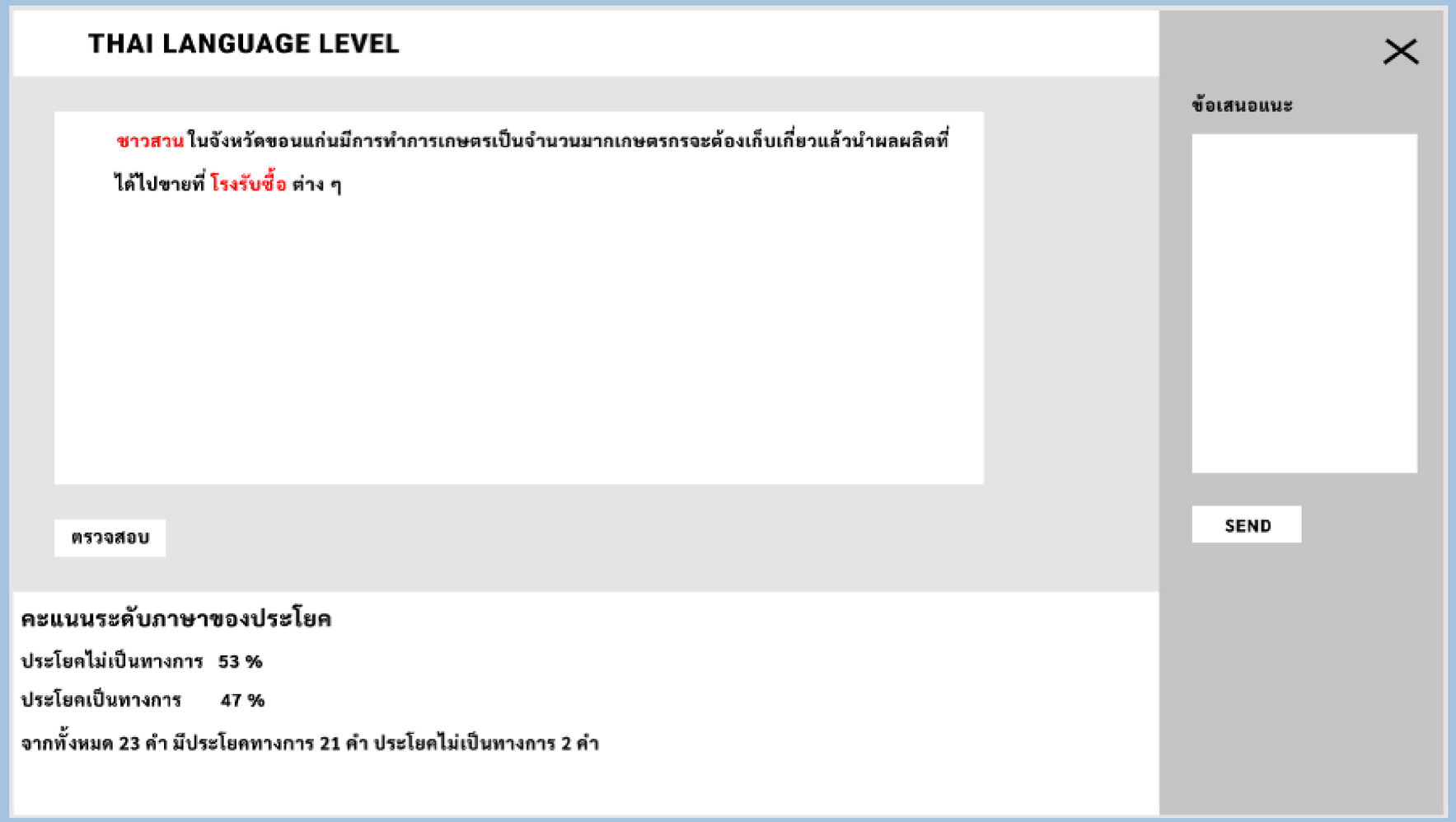
ตัวอย่างเว็บไซต์



ตัวอย่างเว็บไซต์



ตัวอย่างเว็บไซต์



ส่วนที่พัฒนาเสร็จเรียบร้อยแล้ว

1.โมเดลทั้ง 3 แบบ คือ Naive Bayes, ต้นไม้ตัดสินใจ, การหาเพื่อนบ้านที่ใกล้ที่สุด

2.วัดประสิทธิภาพของระบบด้วย K-fold Cross Validation

3.เว็บแอปพลิเคชันบางส่วน

การพัฒนาต่อไป

1.พัฒนาหน้าเว็บแอปพลิเคชันให้สมบูรณ์

2.ทดสอบเว็บแอปพลิเคชัน

Q/A

[1] ราชวิทย์ทิพย์เสนา, ฉัตรเกล้า เจริญผล, แกมกาญจน์สมประเสริฐศรี. (2556). การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความ. คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

[2] Avinash Navlani. (2564). KNN Classification using Scikit-learn, ค้นวันที่ 30 กันยายน 2564 จาก https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

[3] รักษ์พงศ์ ธรรมผุสนา. ระดับภาษาและการใช้ภาษาที่ถูกต้อง. ค้นวันที่ 12 กุมภาพันธ์ 2563 จากhttp://gened.siam.edu/wp-content/uploads/2018/07/thaic-handout-03.pdf

[4] เอกรัฐ บุญเชียง. การแบ่งกลุ่มข้อมูลและการจำแนกประเภทข้อมูล. เชียงใหม่ : ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่, 2561.

[5] อุมารินทร์ นอกตะแบก. ระดับภาษาและคำราชาศัพท์. ค้นวันที่ 20 กันยายน 2563. จาก https://sites.google.com/site/018schoolnet/bth-thi1-radab-phasa-laea-kha-rachasaphth/1

[6] Moshe Koppel, Jonathan Schler, Kfir Zigdon. Determining an Author's Native Language by Mining a Text for Errors. Computer Science Department Bar-lian University. 624-628.

[7] Hanumanthappa, Narayana Swamy. (2015).Indian Language Text Mining. Department of computer Applications Bangalore University Bangalore.

[8] Ekaterina Shutova, Patricia Lichtenstein. (2016). Psychologically Motivated Text Mining. Computer Laboratory University of Cambridge, Dept. of Cognitive and Information Sciences University of California, Merced.

[9] Fadi ABU SHEIKHA, Diana INKPEN. Automatic Classification of Documents by Formality. University of Ottawa, SITE University of Ottawa, SITE 800 King Edward, Ottawa, ON, Canada

[10] ราชวิทย์ทิพย์เสนา, ฉัตรเกล้า เจริญผล, แกมกาญจน์สมประเสริฐศรี. (2556). การจำแนกกลุ่มคำถามอัตโนมัติบนเครือข่ายสังคมออนไลน์. คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม