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

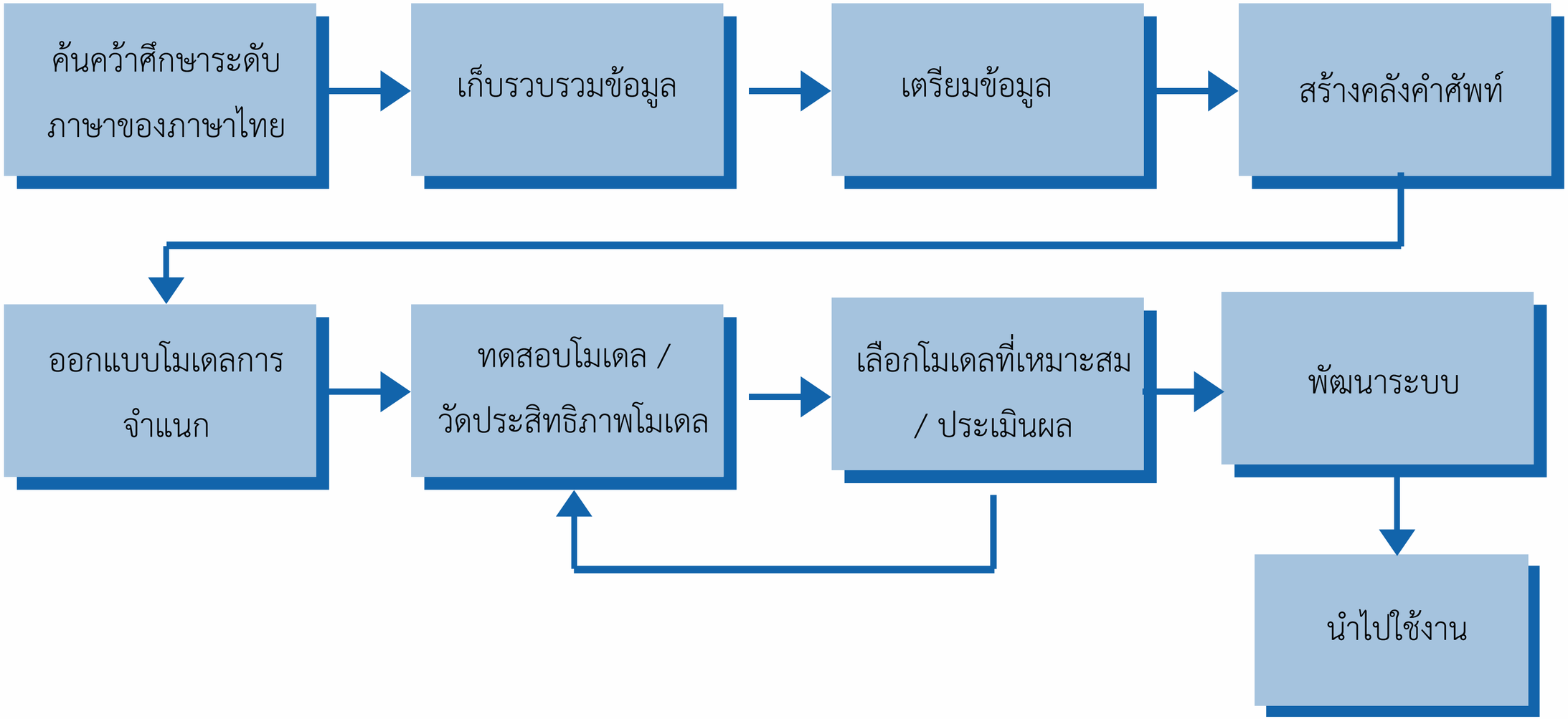
1. จำกัดเฉพาะภาษาไทย
2. ตรวจสอบเฉพาะภาษาที่ใช้ในรายงาน
3. ระบบจะรับข้อมูลเข้าในรูปแบบข้อความเท่านั้น

ความก้าวหน้าจาก Midterm

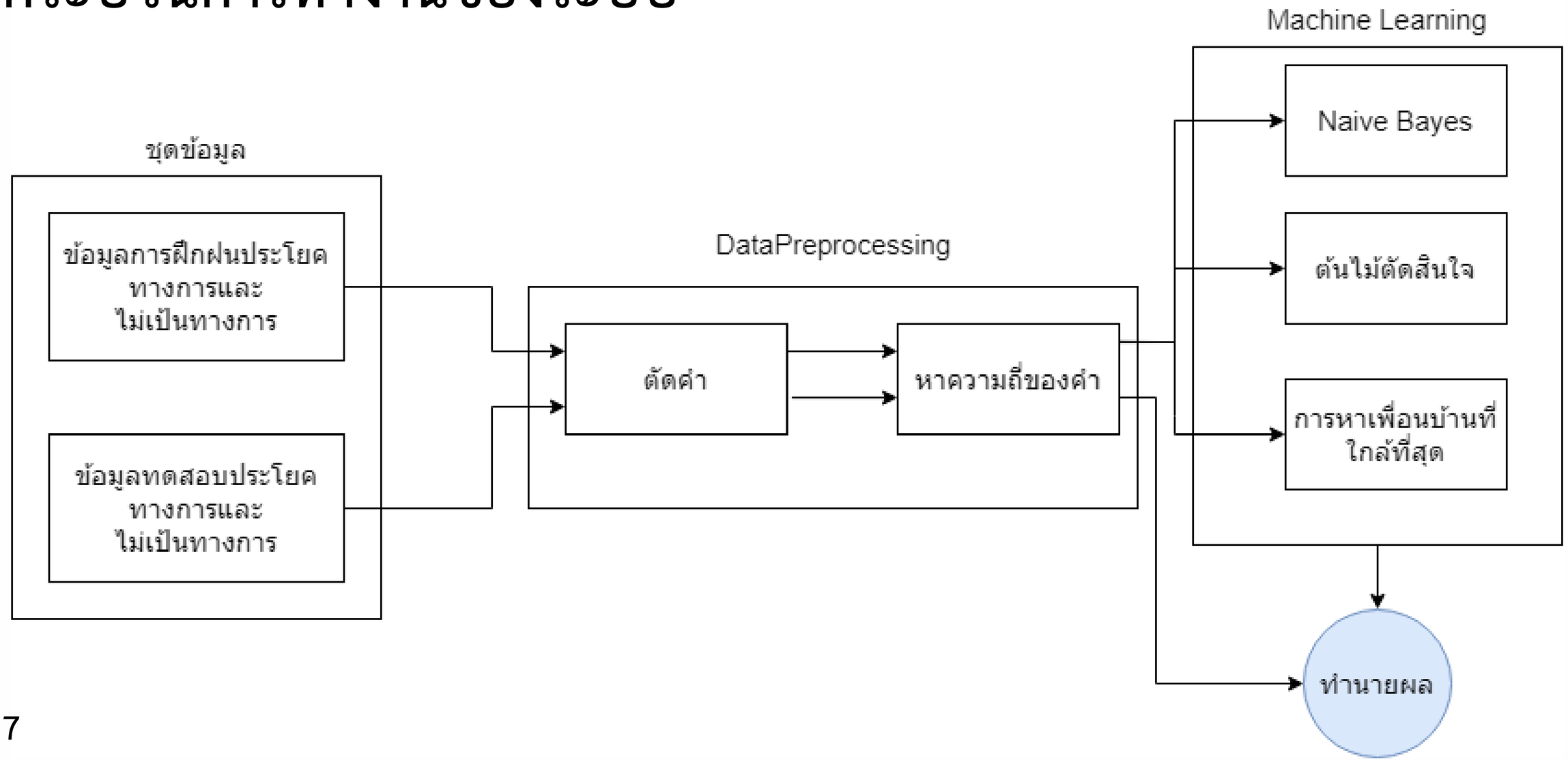
1. เพิ่มประโยคจาก 1,500 ประโยค เป็น 2,000 ประโยค

2. วัดประสิทธิภาพของระบบด้วย K-fold Cross Validation เพื่อเลือกโมเดลที่จะนำมาใช้ในระบบ

ขั้นตอนการดำเนินงาน



กระบวนการทำงานของระบบ



ชุดข้อมูล

ชุดข้อมูลทั้งหมด 2,000 ประโยคที่จะนำไปใช้ในโมเดลจะประกอบไปด้วย

	ประโยค
ประโยคทางการ	1,000
ประโยคไม่เป็นทางการ	1,000

ตัวอย่าง - ชาวนาในขอนแก่นมีการทำการเกษตรเป็นจำนวนมาก

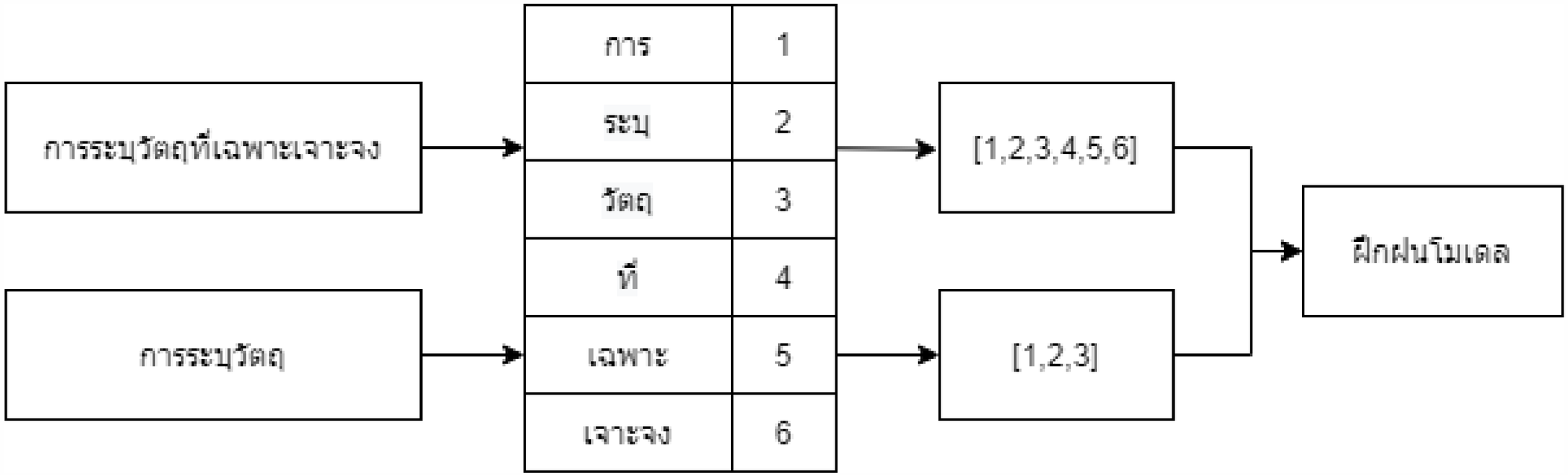
- เกษตรกรจะต้องเก็บเกี่ยวแล้วนำผลผลิตที่ได้ไปขายที่สถานที่รับซื้อต่าง ๆ

ตัวอย่างชุดข้อมูล

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

ตัวอย่างการเตรียมข้อมูล

การระบุวัตถุที่เฉพาะเจาะจง → การ | ระบุ | วัตถุ | ที่ | เฉพาะเจาะจง



1

0

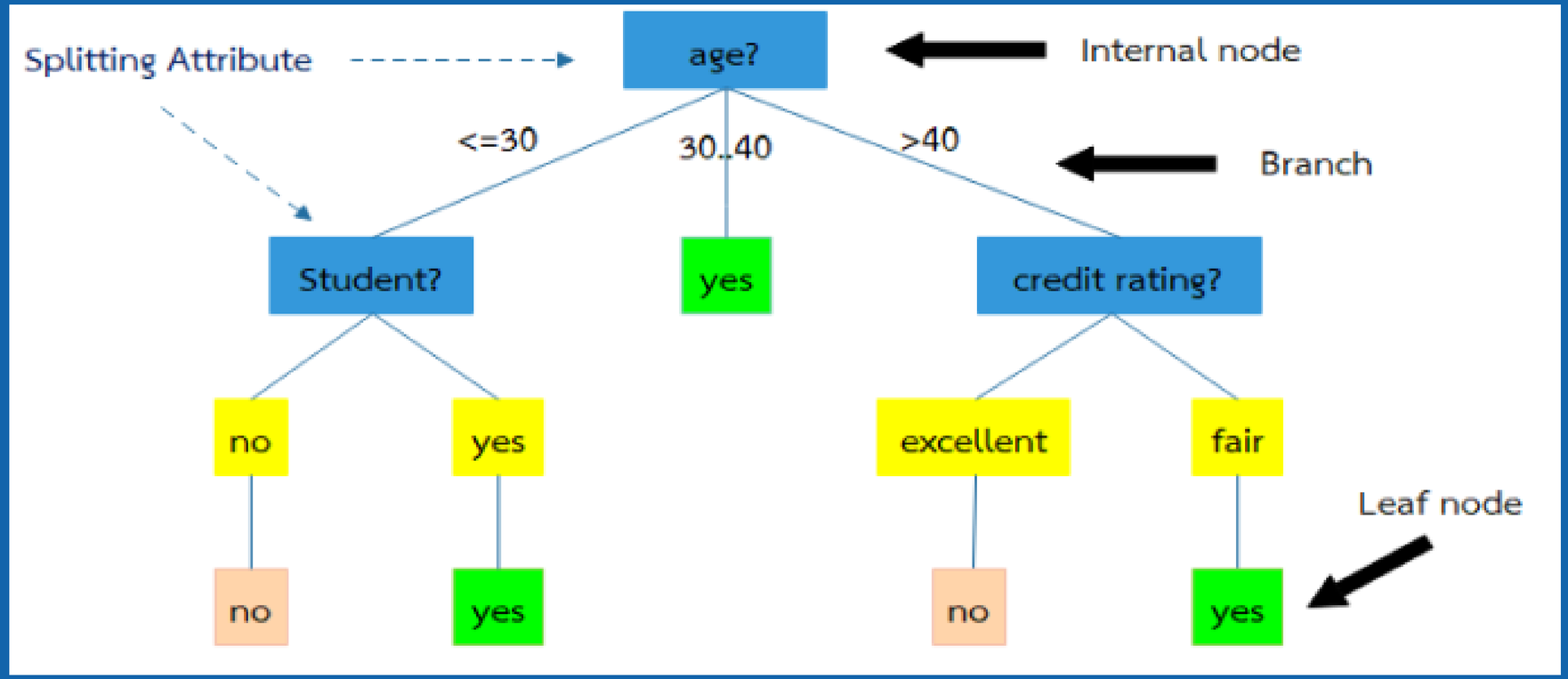
โมเดลที่นำมาทดสอบ

- Naive Bayes
- การหาเพื่อนบ้านที่ใกล้ที่สุด

Midterm

- ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ


Splitting Attribute
age?
Internal node
<=30
30..40
>40
Branch
Student?
yes
credit rating?
no
yes
excellent
fair
Leaf node
no
yes
no
yes

ต้นไม้ตัดสินใจ (ต่อ)

ส่วนประกอบของ Decision Tree ดังนี้

- Root node (โหนดราก) คือ โหนดแรกของต้นไม้ตัดสินใจ
- Internal node (โหนดภายใน) คือ คุณลักษณะต่างๆ ของข้อมูล ซึ่งเมื่อ ข้อมูลมาถึงโหนด จะใช้คุณลักษณะของข้อมูลเพื่อตัดสินใจว่าข้อมูลจะไปในทิศทางใดต่อไป โดยโหนดภายในจะมีจุดเริ่มต้นคือ Root node (โหนดราก)
- Branch (กิ่ง) คือ ค่าคุณลักษณะของแต่ละโหนดที่แตกกิ่งออกไป โดยจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าของคุณลักษณะในโหนดภายในนั้นๆ
- Leaf node (โหนดใบ) คือ กลุ่มของผลลัพธ์ในการจำแนกประเภทข้อมูล

การทำการของ ต้นไม้ตัดสินใจ

1.รากจะเริ่มจากจุดเริ่มต้นของเหตุการณ์

2.เลือก Attribute โดยมีสูตรดังนี้

- ถ้าข้อมูลนำเข้า T มีการแบ่ง class ทั้งหมด n classes Gini index, Gini(T) คือ

$$Gini(t) = 1 - \sum_{i=0}^{C-1} [p(i|t)]^2$$

โดยที่ $p(i|t)$ เป็นควำมถี่ของคลำส i ในข้อมูลนำเข้า t

การทำการของ ต้นไม้ตัดสินใจ (ต่อ)

- ถ้าข้อมูลนำเข้า T แบ่งออกเป็น 2 กลุ่ม คือ T1 และ T2 และมีความถี่ โดยรวม N1 และ N2 ตามลำดับ

$Ginisplit(T)$ คือ

$$Gini_{split}(T) = \frac{N_1}{N} gini(t_1) + \frac{N_2}{N} gini(t_2)$$

ข้อสังเกต Attribute ที่ทำให้ $Ginisplit(T)$ น้อยที่สุด จะเป็น Attribute ที่ดีที่สุด

ผลการทดสอบ ต้นไม้ตัดสินใจ

โมเดล ต้นไม้ตัดสินใจ ทำการหาค่า Accuracy โดยใช้การแบ่งชุดข้อมูลเทรน 1600 ชุดข้อมูล ทดสอบ 400

	Accuracy
ต้นไม้ตัดสินใจ	68.33%

ตารางเปรียบเทียบผลการทดสอบทั้ง 3 โมเดล

ทำการหาค่า Accuracy โดยใช้การแบ่งชุดข้อมูลเทรน 1600 ชุดข้อมูลทดสอบ 400

	k	Prediction
เพื่อนบ้านที่ใกล้ที่สุด	3	55.86%
	5	54.36%
	7	56.62%

ตารางเปรียบเทียบผลการทดสอบทั้ง 3 โมเดล

ทำการหาค่า Accuracy โดยใช้การแบ่งชุดข้อมูลเทรน 1600 ชุดข้อมูลทดสอบ 400

	Prediction
Naive Bayes	73.57%
ต้นไม้ตัดสินใจ	68.33%

ผลการทดสอบ K - fold ของ การหาเพื่อนบ้านที่ใกล้ที่สุด

K-Fold	KNN	ACC
5	3	38.6%
	5	40.6%
	7	37.55%
10	3	47.3%
	5	48.35%
	7	45.74%

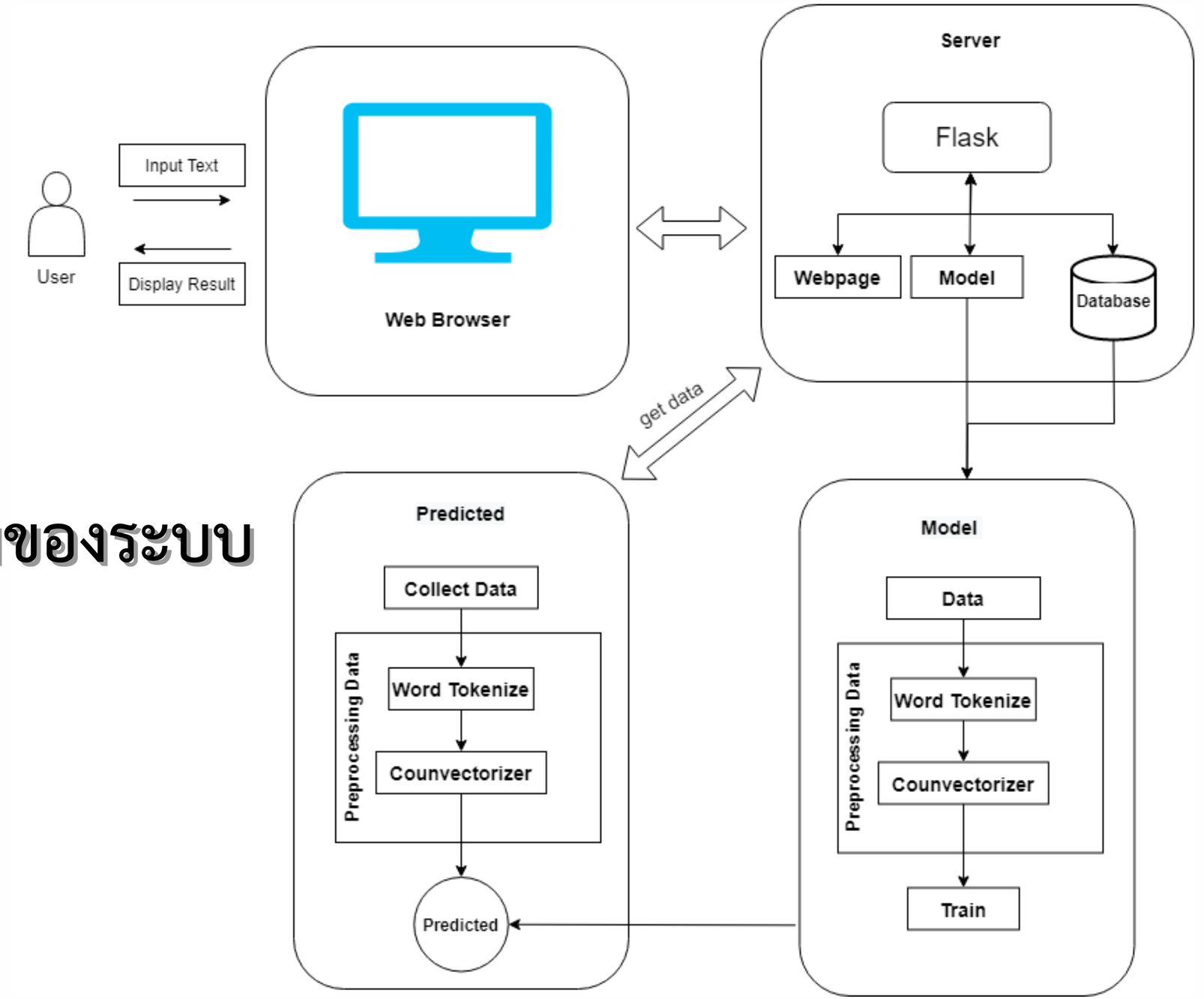
ผลการทดสอบ K - fold ของ Naive Bayes และ ต้นไม้ตัดสินใจ

K-fold	Model	ACC
5	Naive Bayes	74.56%
	ต้นไม้ตัดสินใจ	62.18%
10	Naive Bayes	75.63%
	ต้นไม้ตัดสินใจ	63.37%

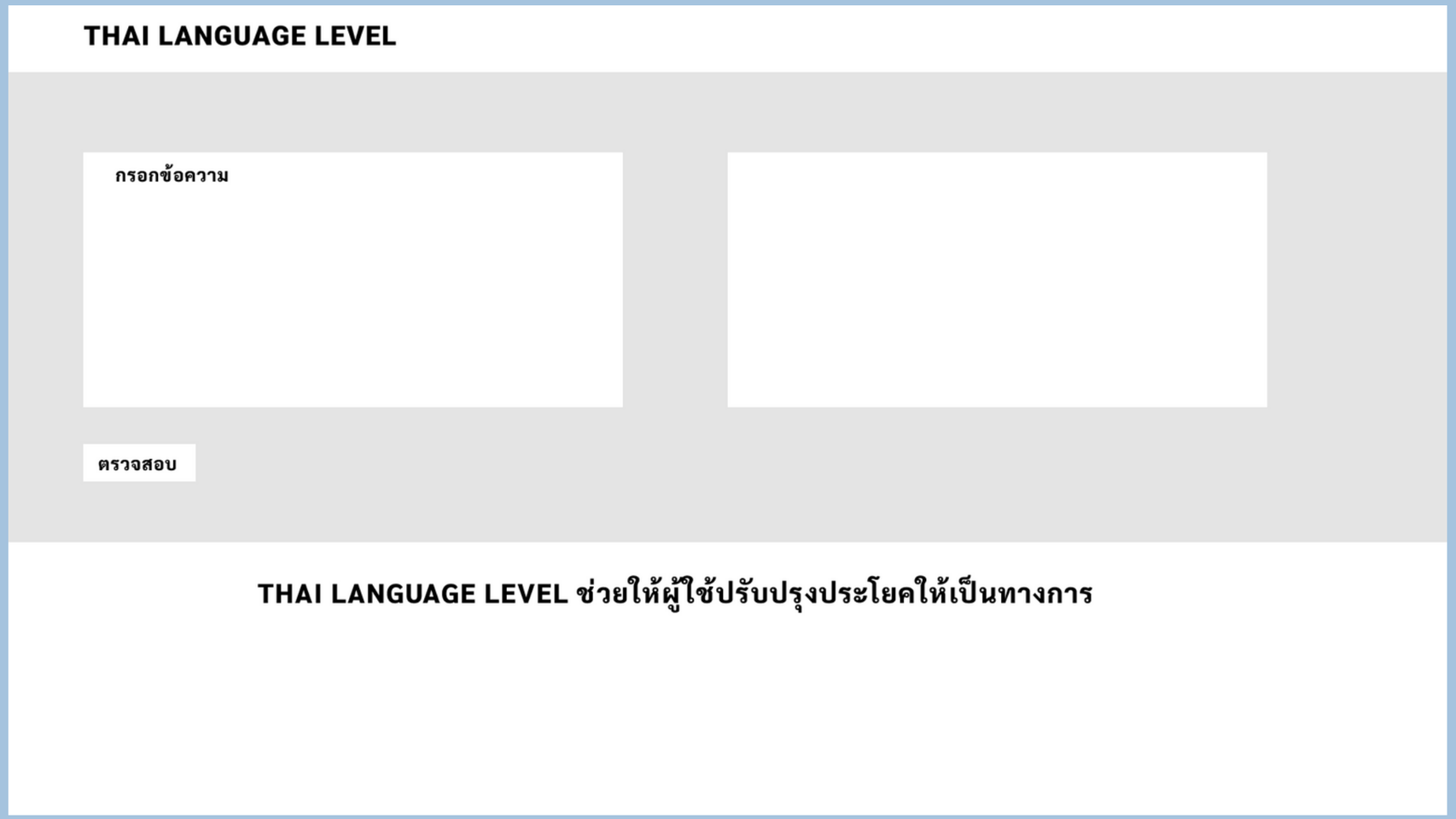
สรุปผลการทดสอบ K - fold

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

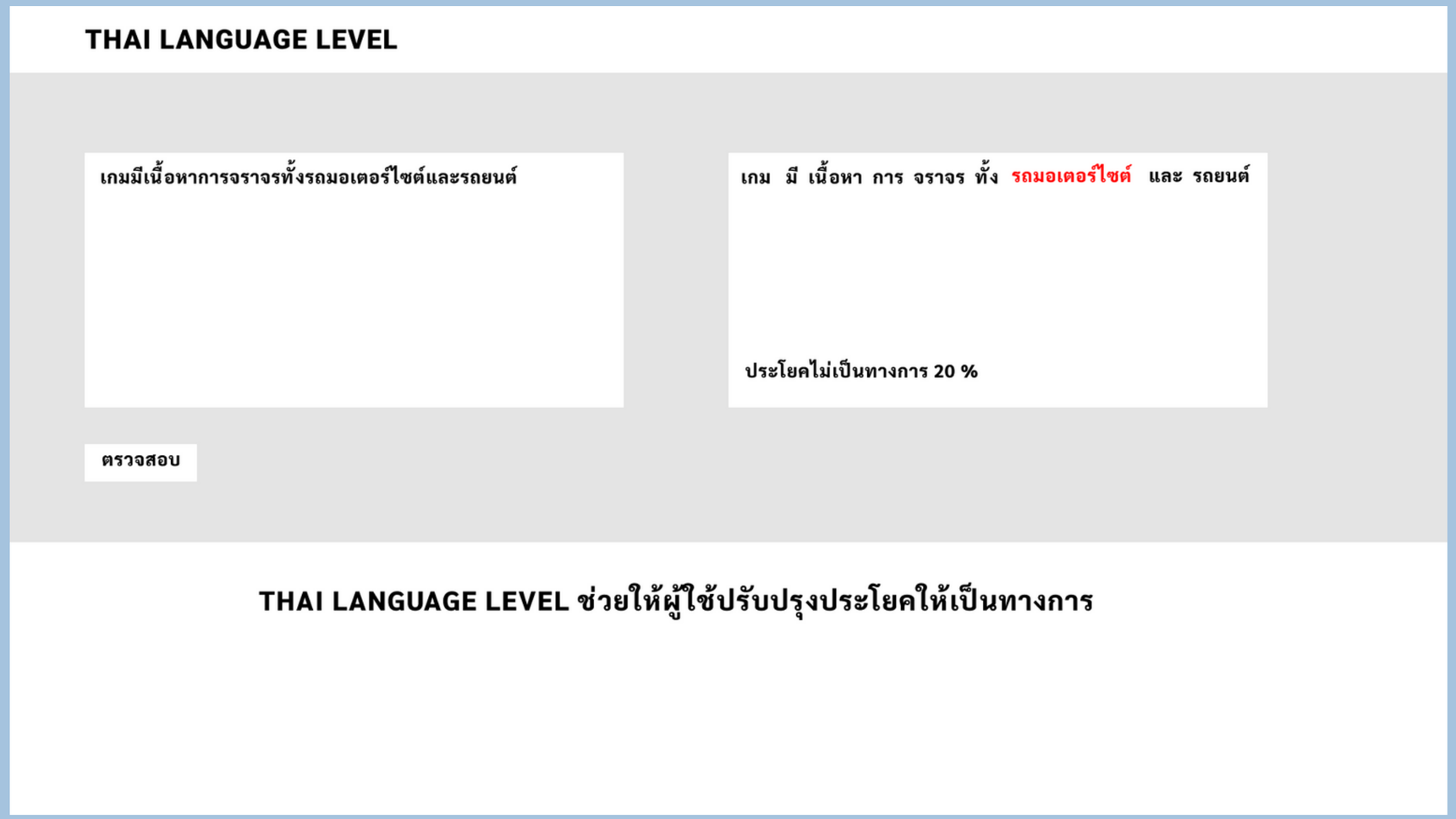
สถาปัตยกรรมของระบบ



ตัวอย่างเว็บไซต์



ตัวอย่างเว็บไซต์



ส่วนที่พัฒนาเสร็จเรียบร้อยแล้ว

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

2.การวัดประสิทธิภาพโมเดล

คิดเป็น 60% ของงานทั้งหมด

การพัฒนาต่อไป

1.พัฒนาโมเดลเพื่อเพิ่มประสิทธิภาพของโมเดล โดยศึกษาพารามิเตอร์ของโมเดลจากงานวิจัย

2.พัฒนาหน้าเว็บแอปพลิเคชั่นของระบบ Thai Language level

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