โครงการ
การเรียนการสอนเพื่อเสริมประสบการณ์

ชื่อโครงการ โปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล

Soccer News Summarization

ชื่อนิสิต นาย กรวิชญ์ กำปั่นทอง

นาย อภิชัย สมนาม

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์

สาขาวิชา วิทยาการคอมพิวเตอร์

ปีการศึกษา 2561

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

โปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล

นาย กรวิชญ์ กำปั่นทอง
นาย อภิชัย สมนาม

โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
สาขาวิชา วิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

Soccer News Summarization

Korawitch Kampanthong

Apichai Somnam

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of Bachelor of Science Program in Computer Science

Department of Mathematics and Computer Science

Faculty of Science

Chulalongkorn University

Academic Year 2018

หัวข้อโครงงาน	โปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล
โดย	นาย กรวิชญ์ กำปั่นทอง
	นาย อภิชัย สมนาม
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาโครงงานหลัก	อ.ดร.นฤมล ประทานวณิช

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับโครงงานฉบับนี้เป็นส่วนหนึ่ง ของการศึกษาตามหลักสูตรปริญญาบัณฑิต ในรายวิชา 2301499 โครงงานวิทยาศาสตร์ (Senior Project)

.....
(ศ.ดร. กฤษณะ เนียมมณี) หัวหน้าภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

คณะกรรมการสอบโครงงาน

.....
(อ.ดร.นฤมล ประทานวณิช) อาจารย์ที่ปรึกษาโครงงานหลัก

.....
(ศ.ดร.ชิดชนก เหลือสินทรัพย์) กรรมการ

.....
(รศ.ดร.วิมลรัตน์ งามอร่ามวรางกูร) กรรมการ

นาย กรวิชญ์ กำปั่นทอง, นาย อภิชัย สมนาม:โปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล (Soccer News Summarization) อ.ที่ปรึกษาโครงงานหลัก : อ.ดร.นฤมล ประทานวณิช, 52 หน้า.

โครงการวิจัยในชั้นเรียน เรื่อง “โปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล” มีวัตถุประสงค์ คือ พัฒนาโปรแกรมสรุปเนื้อหาข่าวสารเกี่ยวกับกีฬาฟุตบอล โดยนำเทคนิคการเรียนรู้ของเครื่อง มาประยุกต์ใช้ เริ่มจากการใช้ word embedding ซึ่งเป็นวิธีที่ใช้สำหรับเปลี่ยนคำเป็นเวกเตอร์จำนวนจริง หลังจากนั้นใช้ตัวแบบทางคณิตศาสตร์ sequence-to-sequence มาใช้ในการประมวลข้อมูลข่าวเพื่อสร้างสรุปข่าว ในส่วนของการวัดผลนั้น เพื่อเปรียบเทียบคะแนนที่ได้จากสรุปที่ได้จากตัวแบบทางคณิตศาสตร์และสรุปที่ได้จากการสุ่ม คณะผู้จัดทำใช้ BLEU scores ซึ่งวัดว่ามีจำนวนคำที่เหมือนกับสรุปข่าวจริงอยู่กี่คำ ทั้งแบบ 1-gram ที่พิจารณาคำแต่ละคำแยกกัน และแบบ 2-gram ที่พิจารณาสองคำที่อยู่ติดกัน ผลการวิจัยที่ได้แสดงให้เห็นว่า สรุปที่ได้จากตัวแบบทางคณิตศาสตร์นั้นดีกว่าสรุปที่ได้จากการสุ่มโดยเฉพาะเมื่อพิจารณาแบบสองคำที่อยู่ติดกัน นอกจากนี้สรุปที่ได้จากโมเดลสามารถอ่านแล้วพอเข้าใจได้ว่าทีมใดชนะ

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต กรวิชญ์ กำปั่นทอง

ลายมือชื่อนิสิต อภิชัย สมนาม

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาโครงงานหลัก นฤมล

ปีการศึกษา 2561 ลายมือชื่อ อ.ที่ปรึกษาโครงงานร่วม

5833604723, 5833667223: MAJOR COMPUTER SCIENCE

KEYWORDS: NEURAL NETWORK / SEQUENCE-TO-SEQUENCE / TEXT SUMMARIZATION

Korawitch Kampanthong, Apichai Somnam: Soccer News Summarization. ADVISOR: ASSOC. PROF. Ph.D. Naruemon Pratanwanich, 52 pp.

The objective of this project is to apply machine learning techniques for soccer's news summarization. First, we used a word embedding technique which converts words into numerical vectors. Then, we applied a sequence-to-sequence model to learn the conversion of news scripts to their corresponding summaries. To evaluate the model performance, we used 1-gram and 2-gram BLEU scores to compute the number of words that the model and the random procedure recalled from the true summaries. Our results reveal that the summaries from the sequence-to-sequence model had higher BLEU scores than the output from random summarization, especially on the 2-gram BLEU scores, indicating more readability. Additionally, the model's summaries were moderately understandable which team was the winner.

Department: Mathematics and Computer Science Student's Signature กรวิชญ์ กำปั่นทอง

Student's Signature อภิชัย สมนาม

Field of Study : Computer Science Advisor's Signature

Academic Year : 2018 Co-advisor's Signature

กิตติกรรมประกาศ

โครงงานโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอลสามารถสำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์อย่างดียิ่งของ อาจารย์ ดร.นฤมล ประทานวณิช อาจารย์ที่ปรึกษาโครงการ และ ผศ.ดร.ฑิตยา หวานวารี ซึ่งเสียสละเวลาให้ความรู้ คำปรึกษา และสนับสนุนด้วยความเอาใจใส่อย่างยิ่งจนทำให้โครงงานสำเร็จลุล่วงได้ด้วยดี

ขอขอบพระคุณ กรรมการคุมสอบ ศ.ดร.ชิดชนก เหลือสินทรัพย์ และ รศ.ดร.วิมลรัตน์ งามอร่ามวรางกูร ผู้เป็นกรรมการคุมสอบที่ช่วยแนะแนวทางต่าง ๆ ที่เป็นประโยชน์ต่อโครงงานนี้

สุดท้ายขอขอบคุณทุกท่านที่ไม่ได้กล่าวนามไว้ข้างต้น ที่ให้การสนับสนุนในด้านต่าง ๆ ที่คอยผลักดันให้โครงงานสำเร็จลุล่วงไปได้ด้วยดี

คณะผู้จัดทำ

สารบัญ

หน้า

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและเหตุผลการวิจัย

ปัจจุบันมีข่าวกีฬามากมายที่คนบนโลกโซเชียลส่วนใหญ่นิยมเข้าไปอ่าน หนึ่งในข่าวกีฬาที่ได้รับความนิยมเป็นอย่างมากคือ ข่าวกีฬาฟุตบอลซึ่งมีหลากหลายรายการแข่งขัน ไม่ว่าจะเป็น การแข่งขันฟุตบอลโลก การแข่งขันของลีคแต่ละประเทศ รวมถึงนัดอุ่นเครื่องอีกมากมาย ซึ่งจะเห็นได้ว่ามีนัดการแข่งขันเป็นจำนวนมาก ในขณะที่ผู้คนก็ต้องการรู้ เหตุการณ์ต่าง ๆ ในการแข่งขันเป็นจำนวนมากเช่นกัน ในปัจจุบันการสรุปเนื้อหาโดยใช้คอมพิวเตอร์ยังไม่แพร่หลายนัก เนื่องจากขาดความแม่นยำและข่าวที่สรุปนั้นอ่านแล้วเข้าใจได้ยาก จึงนิยมสรุปโดยใช้แรงงานมนุษย์ อย่างไรก็ตามจำนวนการแข่งขันนั้นมีเป็นจำนวนมาก การใช้แรงงานมนุษย์จึงไม่เพียงพอต่อจำนวนนัดการแข่งขัน

ปัจจุบันมีคนวิจัยการสรุปเนื้อหาข่าวกีฬาฟุตบอลอยู่บ้าง แต่เห็นผลงานที่เป็นรูปธรรมเช่น ซอฟต์แวร์หรือโปรแกรมประยุกต์ ออกมาเป็นจำนวนน้อย ในการศึกษาพบว่ามีงานวิจัยจำนวนน้อยที่นำหลักของตัวแบบทางคณิตศาสตร์ (model) เข้ามาประยุกต์ใช้ ดังนั้นทางคณะผู้จัดทำจึงได้สนใจนำหลักของตัวแบบทางคณิตศาสตร์เข้ามาประยุกต์ใช้เพื่อให้ข้อความข่าวที่สรุปมีความถูกต้องแม่นยำและรวดเร็วมากขึ้น

ดังนั้นคณะผู้จัดทำจึงมีความสนใจที่จะพัฒนาโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอลที่เป็นภาษาอังกฤษโดยนำหลักการเรียนรู้ของเครื่อง (machine learning) มาประยุกต์ใช้เพื่อสร้างตัวแบบคณิตศาสตร์ในการสรุปข่าว เพื่อให้ผู้ใช้งานได้เสพข่าวที่มีความรวดเร็วขึ้นและทำให้ช่วยลดภาระในการสรุปข่าว

1.2 วัตถุประสงค์ของการวิจัย

พัฒนาโปรแกรมสรุปเนื้อหาข่าวสารเกี่ยวกับกีฬาฟุตบอล โดยนำเทคนิคการเรียนรู้ของเครื่อง มาประยุกต์ใช้

1.3 ขอบเขตการวิจัย

1. การศึกษานี้จะมุ่งศึกษาเกี่ยวกับการสรุปสคริปท์จากรายงานสดกีฬาฟุตบอล
2. การศึกษานี้จะใช้แหล่งที่มาของคลิปข่าวจากเว็บไซต์ https://www.sportsmole.co.uk เท่านั้น
3. การศึกษานี้จะใช้ข่าวที่เป็นภาษาอังกฤษเท่านั้น
4. ผลลัพธ์ที่ได้จะอยู่ในรูปแบบข้อความภาษาอังกฤษและไม่สนใจไวยากรณ์
5. ใช้ตัวแบบทางคณิตศาสตร์ที่มีชื่อว่า sequence-to-sequence เท่านั้น

1.4 ขั้นตอนการวิจัย

1. ศึกษาเนื้อหาและบทความเกี่ยวกับการสรุปข้อความโดยใช้การเรียนรู้ของเครื่อง ซึ่งใช้ RNN ประเภท sequence-to-sequence จากคลังโปรแกรมของ Google ที่มีชื่อว่า TensorFlow

มาเป็นตัวต้นแบบในการพัฒนา โดยจะทำเป็น supervised learning โดยสร้างสรุปข่าวจากสคริปท์ข้อมูลที่เป็นข้อมูลขาเข้า

2. รวบรวมข้อมูลสคริปท์ข่าวและสรุปของข่าวเดียวกันโดยใช้คลังโปรแกรมที่มีชื่อว่า beautiful soup 3
3. ศึกษาข้อมูล โปรแกรมและเทคนิคที่สามารถนำมาใช้ในการดำเนินงาน โดยสรุปจากเอกสารที่ใช้อ้างอิงได้ดังนี้ ตัวแบบคณิตศาสตร์ที่ใช้คือ sequence-to-sequence model เป็น DNN (Deep Neural Network) ที่แบ่งเป็น 2 ส่วนคือ Encoder และ Decoder โดยส่วน encoder จะเข้ารหัส (encode) สคริปท์และส่วน decoder จะถอดรหัส (decode) ออกมาเป็นสรุป นอกจากนั้นในเอกสารอ้างอิงก็จะใช้ LSTM แทน standard RNN ในการสร้างส่วนเข้ารหัสและส่วนถอดรหัสอีกด้วยด้วย [2]
4. กำหนดขอบเขตและวิธีการดำเนินงาน
5. ออกแบบโมเดลและวิธีการสรุปสคริปท์ข่าวที่รวบรวมมา
6. พัฒนาโปรแกรมสรุปสคริปท์ข่าว
7. ทดสอบการใช้โปรแกรมที่พัฒนาแล้ว โดยใช้ BLEU คิดคะแนนจากสรุปที่ได้กับโปรแกรมเทียบกับสรุปจากการสุ่มคำ (random) และทำการสัมภาษณ์ให้คนอ่านอ่านสรุปที่ได้จากโปรแกรมแล้วสามารถบอกได้ว่าทีมไหนชนะหรือไม่ [3]
8. ตรวจสอบความถูกต้องและแก้ไขความผิดพลาดของโปรแกรมที่พัฒนา
9. ทำการวัดผลเพื่อเปรียบเทียบการใช้งานระหว่างใช้งานโปรแกรมและไม่ได้ใช้งาน ในแง่ของเวลาและอ่านแล้วจับใจความได้
10. สรุปผลการดำเนินงาน ข้อเสนอแนะและการจัดทำเอกสาร

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่ได้รับจากการวิจัยในครั้งนี้มีดังนี้

1. ประโยชน์ต่อนิสิตที่ทำโครงการ
   1.1. ได้ศึกษาเรียนรู้เกี่ยวกับ Machine Learning
   1.2. ได้ศึกษาเรียนรู้เกี่ยวกับ การสรุปข่าวกีฬาฟุตบอล
2. ประโยชน์ที่ได้จากโครงงานที่พัฒนาขึ้น
   2.1. สามารถสรุปข่าวฟุตบอลได้อย่างรวดเร็วแม่นยำ

1.6 โครงสร้างของรายงาน

บทที่ 2 จะกล่าวถึงเอกสาร ความรู้และงานวิจัยที่เกี่ยวข้องกับโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล

บทที่ 3 จะกล่าวถึงวิธีการวิจัยและพัฒนา ซึ่งจะประกอบไปด้วย วิธีการเก็บข้อมูลที่นำมาใช้ในการพัฒนาโปรแกรม ตัวอย่างข้อมูลที่ใช้ในการพัฒนา วิธีการปรับแต่งข้อมูลที่ใช้ในการพัฒนาก่อนที่จะเข้าสู่ตัว

แบบทางคณิตศาสตร์ วิธีการพัฒนาตัวแบบทางคณิตศาสตร์ และวิธีการวัดผลโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล

บทที่ 4 จะกล่าวถึงผลการวิจัยและสรุปผลการวิจัย

บทที่ 5 จะกล่าวถึงสรุปผลการดำเนินงาน เป้าหมายในอนาคต ปัญหาของงานวิจัยและวิธีการแก้ไข

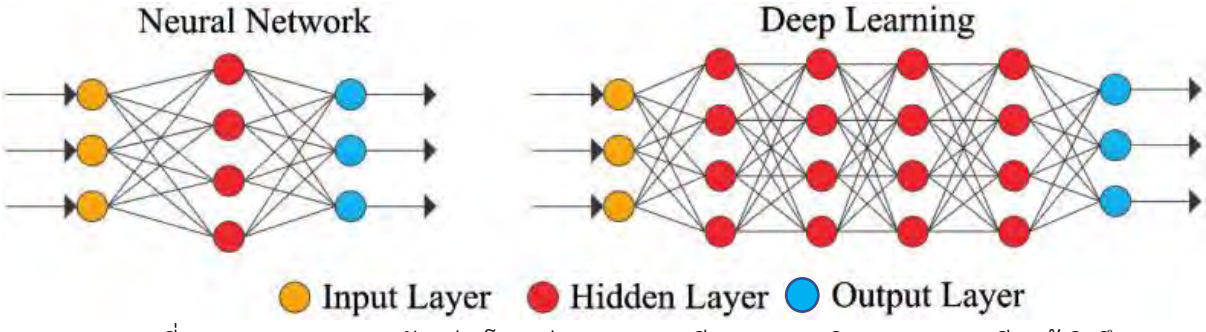
บทที่ 2
เอกสาร ความรู้และงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึง ความรู้และงานวิจัยที่เกี่ยวข้องกับโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล

2.1 การเรียนรู้ของเครื่อง (Machine learning) และ การเรียนรู้เชิงลึก (Deep learning)

การเรียนรู้ของเครื่อง คือ ส่วนการเรียนรู้ของเครื่องคอมพิวเตอร์ เป็นสาขาย่อยของปัญญาประดิษฐ์ (Artificial Intelligence : AI) ซึ่งการเรียนรู้ของเครื่องถูกใช้งานเสมือนเป็นสมองของปัญญาประดิษฐ์ อาจพูดได้ว่าปัญญาประดิษฐ์ใช้การเรียนรู้ของเครื่องในการสร้างความฉลาด โดยการเรียนรู้ของเครื่องมักจะใช้เรียกตัวแบบทางคณิตศาสตร์ที่เกิดจากการเรียนรู้ของปัญญาประดิษฐ์ กล่าวคือมนุษย์มีหน้าที่เขียนโปรแกรมให้ปัญญาประดิษฐ์เรียนรู้จากข้อมูลเท่านั้น ที่เหลือเครื่องจัดการทำงานเอง

การเรียนรู้เชิงลึก เป็นสาขาของการเรียนรู้ของเครื่อง พื้นฐานของการเรียนรู้เชิงลึกคือ อัลกอริทึมที่พยายามจะสร้างแบบจำลองเพื่อแทนความหมายของข้อมูลในระดับสูงโดยการสร้างสถาปัตยกรรมข้อมูลขึ้นมาที่ประกอบไปด้วยโครงสร้างย่อย ๆ หลายอัน และแต่ละอันนั้นได้มาจากการแปลงที่ไม่เป็นเชิงเส้น (non-linear)



รูปภาพที่ 2.1 แผนภาพแสดงตัวอย่างโครงข่ายประสาทเทียมแบบปกติและแบบการเรียนรู้เชิงลึก

จากรูปภาพที่ 2.1 จะเห็นได้ว่าโครงข่ายประสาทเทียมนั้นประกอบไปด้วยเส้นหรือน้ำหนัก (weight) ระหว่างเซลล์ และ เซลล์ที่มี 3 แบบคือ Input layer หรือชั้นนำเข้า Hidden layer หรือชั้นซ่อน และ Output layer หรือชั้นผลลัพธ์

ชั้นนำเข้าเป็นส่วนที่รับข้อมูลนำเข้ามาส่งต่อให้กับเซลล์ต่อไป

ชั้นซ่อนเป็นส่วนที่นำข้อมูลที่ได้รับมาจากเซลล์ที่แล้วมาคำนวณหาค่าใหม่โดยใช้สูตรด้านล่าง

$$h = f(\sum wx + b)$$

โดยที่ h แทนชั้นซ่อน f(·) เป็น activation function ที่ไม่เป็นเชิงเส้น w แทนน้ำหนัก x แทนข้อมูลเข้าจากเซลล์ก่อนหน้า และ b เป็นค่าคงตัวที่กำหนดไว้ของแต่ละโครงข่าย

ชั้นผลลัพธ์เป็นชั้นที่รับค่ามาจากเซลล์ก่อนหน้าเพื่อส่งค่าออกไปเป็นผลลัพธ์ไปยังภายนอกโครงข่าย

จะเห็นได้ว่าโครงข่ายประสาทเทียมแบบการเรียนรู้เชิงลึกนั้นแตกต่างจากแบบปกติตรงที่มีจำนวนชั้นของชั้นซ่อนมากกว่าและมีหลายสถาปัตยกรรมเช่น โครงข่ายประสาทแบบป้อนไปหน้า (Feed-forward

Neural Networks : FNNs) ที่เป็นการส่งค่าจากเซลล์ไปสู่อีกเซลล์หนึ่งทางเดียวต่อไปเรื่อย ๆ ส่วนมากใช้กับข้อมูลทั่วไปที่ไม่มีความซับซ้อน โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Networks : CNNs) ที่เป็นการจำลองการมองเห็นของมนุษย์ที่มองพื้นที่เป็นที่ย่อย ๆ และนำกลุ่มของพื้นที่ย่อย ๆ มาผสานกัน เพื่อหาว่าสิ่งนั้นคืออะไร ส่วนมากใช้กับข้อมูลที่เป็นภาพ และ โครงข่ายแบบวนซ้ำ (Recurrent neural networks : RNNs) เป็นโครงข่ายหลายชั้นที่สามารถเก็บข้อมูลไว้ที่เซลล์จึงทำให้มันสามารถรับข้อมูลเป็นแบบลำดับและให้ผลลัพธ์ออกเป็นลำดับของข้อมูลได้ ส่วนมากใช้กับข้อมูลที่เป็นลำดับ ซึ่งในโครงงานนี้คณะผู้วิจัยได้นำโครงข่ายแบบวนซ้ำมาใช้ในตัวแบบทางคณิตศาสตร์

2.2 การประมวลภาษาธรรมชาติ (NLP: Natural language processing)

การประมวลภาษาธรรมชาติ เป็นสาขาย่อยของปัญญาประดิษฐ์และภาษาศาสตร์ที่ศึกษาปัญหาในการประมวลผลและใช้งานภาษาธรรมชาติ รวมทั้งการทำความเข้าใจภาษาธรรมชาติ ทั้งนี้เพื่อให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์ได้

2.3 รูปแบบการแทนข้อความ (Text representation)

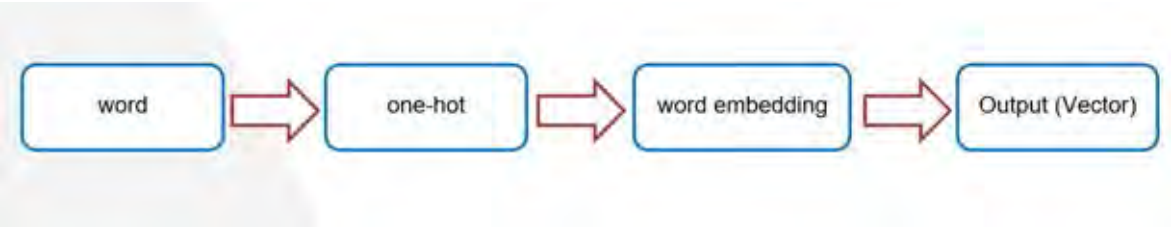
เนื่องจากการนำข้อความมาใช้ในการเรียนรู้ของเครื่องโดยตรงนั้นทำได้ยาก คณะผู้จัดทำจึงได้เลือกใช้รูปแบบการแทนข้อความต่าง ๆ ดังต่อไปนี้

2.3.1 One-hot

One-hot คือชุดของบิท (bit) ของข้อมูลประเภทต่าง ๆ โดยบิทที่แทนคำนั้นจะเป็น 1 แค่ตัวเดียวและที่เหลือจะเป็น 0 การใช้การเข้ารหัสแบบ one-hot จะไม่สนลำดับคำ หรือคำอื่น ๆ ที่อยู่ลำดับติดกัน จะสนใจแค่ว่ามีคำนั้น ๆ อยู่หรือไม่เท่านั้น

2.3.2 Word embedding

Word embedding คือการแปลงคำเป็นเวกเตอร์ ถือเป็นหนึ่งในวิธีในการสร้างตัววัด (features) จากคำวิธีหนึ่ง โดยจะทำการลดขนาดของปริภูมิเวกเตอร์ (vector space) ลงด้วย ในการศึกษานี้คณะผู้จัดทำได้ทำการแบ่งข้อมูลขาเข้า (input) ซึ่งก็คือเนื้อหาสคริปต์ข่าวเป็นประโยค และทำการแบ่งประโยคออกเป็นคำเดี่ยว ๆ เพื่อนำไปป้อนให้กับ word embedding ผลลัพธ์ที่ได้จะเป็นเวกเตอร์ของคำ เพื่อนำไปใช้กับตัวแบบทางคณิตศาสตร์ที่มีชื่อว่า sequence-to-sequence ต่อไป



รูปภาพที่ 2.2 แผนภาพแสดงหลักการทำงานโดยง่ายของ word embedding

2.3.3 CN (Conceptnet Numberbatch)

CN คือ อัลกอริทึมแบบ unsupervised learning สำหรับการทำ word embedding ในการศึกษานี้ได้นำ CN มาใช้เป็นคลังคำศัพท์ (corpus) เพื่อใช้ในการอ้างอิงเวกเตอร์ในการทำ word embedding ซึ่งใน CN ประกอบด้วยคำศัพท์จาก GloVe word2vec และ fastText รวมกัน โดยดาวน์โหลดจาก https://github.com/commonsense/conceptnet-numberbatch

2.4 ตัวแบบคณิตศาสตร์

ในการศึกษาครั้งนี้คณะผู้จัดทำได้ตัดสินใจใช้ตัวแบบคณิตศาสตร์ที่มีชื่อว่า sequence-to-sequence เท่านั้น

2.4.1 sequence-to-sequence

sequence-to-sequence เป็นตัวแบบคณิตศาสตร์ชนิดหนึ่งของโครงข่ายประสาทเทียมแบบวนซ้ำ (RNN: Recurrent Neural Networks) โดยหลักการโดยย่อของ sequence-to-sequence คือการป้อนข้อมูลขาเข้าเป็นลำดับ (sequence) และผลลัพธ์ที่ได้ก็เป็นลำดับ ตัวแบบคณิตศาสตร์นี้ถูกนำไปใช้ในงานหลาย ๆ งานเช่น เครื่องแปลภาษา (Machine translation) การจำแนกเสียง (Speech recognition) หรือ คำบรรยายวิดีโอ (Video captioning)

sequence-to-sequence มีชื่อเรียกอีกชื่อหนึ่งว่า RNN Encoder-Decoder ซึ่งเป็นตัวแบบคณิตศาสตร์ที่จะถูกนำมาใช้ในการประมวลข้อมูลข่าว สามารถแบ่งได้เป็น 3 ส่วนดังนี้ (ดังรูปภาพที่ 2.2)

1. ตัวเข้ารหัส (Encoder) เป็นส่วนที่มีไว้เข้ารหัสข้อมูลขาเข้าที่เป็นลำดับของคำให้เป็นเวกเตอร์ที่กำหนดขนาดไว้ ลำดับการป้อนข้อมูลคือชุดของคำทั้งหมด แต่ละคำจะถูกแทนด้วย x_t โดยที่ t คือลำดับของคำนั้น สถานะที่ซ่อนอยู่ (hidden state) h_t ถูกคำนวณโดยใช้สูตรด้านล่าง

$$h_t = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t)$$

สมการง่าย นี้แสดงให้เห็นถึงผลลัพธ์ของ RNN ทั่วไป โดย f(·) เป็น activation function และในช่วงของการฝึกสอน ตัวแบบทางคณิตศาสตร์จะหาน้ำหนักที่เหมาะสม ($W^{(hh)}$, $W^{(hx)}$) สำหรับสถานะที่ซ่อนไว้ก่อนหน้านี้ $h_{(t-1)}$ และเวกเตอร์ขาเข้า x_t

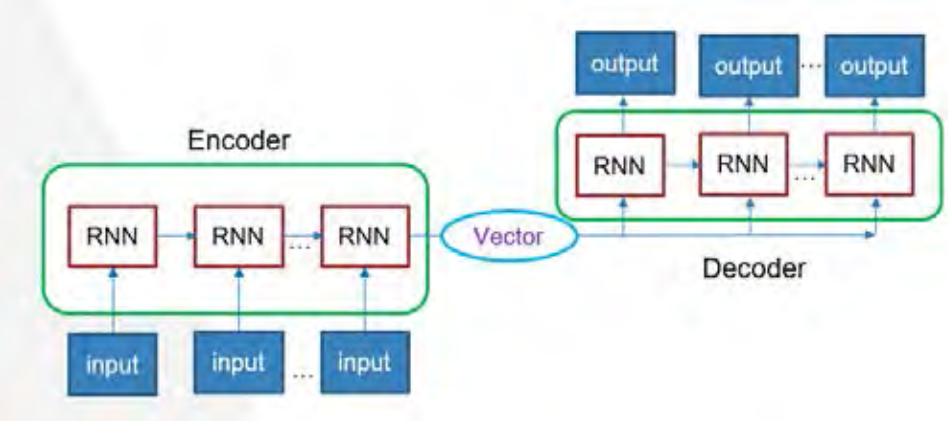
2. เวกเตอร์สื่อกลาง (Intermediate vector หรือ Encoded vector) เป็น สถานะที่ซ่อนอยู่ตัวสุดท้ายที่ผลิตจากส่วนตัวเข้ารหัส ซึ่งคำนวณโดยใช้สูตรด้านบน เวกเตอร์นี้มีจุดมุ่งหมายเพื่อรวบรวมข้อมูลไว้เป็นกลุ่มเดียวกันสำหรับองค์ประกอบขาเข้าทั้งหมดเพื่อช่วยให้ตัวถอดรหัสทำการทำนาย และยังทำหน้าที่เป็นสถานะที่ซ่อนตัวแรกของตัวถอดรหัส
3. ตัวถอดรหัส (Decoder) เป็นกลุ่มของหน่วย RNN ที่แต่หน่วยละทำนายผลลัพธ์ของคำในลำดับ i (y_i) จากค่าของสถานะซ่อนและผลลัพธ์ที่ทำนาย จากหน่วยก่อนหน้า h'_{t-1} และ y_i คำตอบสุดท้ายคือชุดของคำทั้งหมดที่ทำนายจากหน่วย RNN ทุกอัน แต่ละคำจะถูกแทนด้วย y_i โดยที่ i คือลำดับของคำนั้น แต่ละสถานะที่ซ่อนอยู่ (h'_i) ของตัวถอดรหัสคำนวนได้จากสูตรด้านล่าง

$$h'_i = f(W^{(h'h')}h'_{i-1})$$

ซึ่งใช้สถานะที่ซ่อนอยู่ก่อนหน้านี้เพื่อคำนวณสถานะถัดไป ผลลัพธ์ y_i ณ เวลาที่ขั้นตอน i คำนวณโดยใช้สูตรด้านล่าง

$$y_i = softmax(W^S h'_i)$$

ซึ่งจะคำนวณผลลัพธ์โดยใช้สถานะที่ซ่อนอยู่ในขั้นตอนเวลาปัจจุบันพร้อมกับน้ำหนักที่เกี่ยวข้อง W^S ฟังก์ชั่น Softmax ใช้เพื่อสร้างเวกเตอร์ความน่าจะเป็นซึ่งจะช่วยให้เรากำหนดผลลัพธ์สุดท้ายได้ว่าเป็นคำใด



รูปภาพที่ 2.3 แผนภาพแสดงหลักการทำงานโดยง่ายของตัวแบบคณิตศาสตร์ sequence-to-sequence

2.5 วิธีการวัดผล

ทางคณะผู้จัดทำได้ศึกษาวิธีการวัดผล 2 วิธีดังต่อไปนี้

2.5.1 BLEU score (BiLingual Evaluation Understudy score)

BLEU score เป็นตัวชี้วัดในการประเมินคุณภาพของข้อความที่ได้รับการแปลด้วยเครื่อง (machine translation) จากภาษาธรรมชาติ (natural language) หนึ่งไปยังอีกภาษาหนึ่ง ซึ่งคะแนนที่ได้จะมีค่าได้ตั้งแต่ 0 ถึง 1 คณะผู้จัดทำได้เลือกวิธีนี้ในการวัดผลเพราะเป็นที่นิยม โดยจะใช้แบบ 1 gram และ 2 grams ซึ่งผล 1 gram หมายถึง เจอคำในสรุปจากตัวแบบทางคณิตศาสตร์ตรงกับสรุปจากเว็บ 1 คำได้1 คะแนน และผล 2 gram หมายถึงเจอคำในสรุปจากตัวแบบทางคณิตศาสตร์ตรงกับสรุปจากเว็บ 2 คำติดกันได้ 1 คะแนน

2.5.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score

ROUGE score เป็นอีกหนึ่งตัวชี้วัดในการวัดผลการสรุปอัตโนมัติ (automatic summarization) และซอฟต์แวร์การแปลด้วยเครื่องในการประมวลภาษาธรรมชาติ

2.6 งานวิจัยที่เกี่ยวข้อง : Summarizing Text with Amazon Reviews

Summarizing Text with Amazon Reviews [4] เป็น tutorial หลักที่เราศึกษาและนำมาใช้อ้างอิงในโครงงานครั้งนี้ tutorial นี้สร้างขึ้นเพื่อสรุปข้อความรีวิวของอเมซอนให้สั้นกะทัดรัดได้ใจความ ซึ่งแบ่งออกเป็นส่วนต่าง ๆ คือเริ่มโดยการดูข้อมูลที่เราใช้ในการวิจัย จากนั้นจึงเตรียมข้อมูลก่อนที่จะนำมาทำการสร้างตัวแบบทางคณิตศาสตร์ โดยเริ่มจากการเปลี่ยนคำย่อในข้อมูลให้เป็นคำเต็ม จากนั้นทำการลบตัวอักษรและพวกคำหยุด (stopwords) ที่ไม่ต้องการออก แล้วจึงนับคำแล้วสร้าง dict ของชนิดคำในสรุปและข้อความ ต่อจากนั้นก็ทำ word embedding แล้วทำกระบวนการเปลี่ยนคำเป็นตัวเลข และตัวเลขเป็นคำ เพื่อใช้ในการสอน

(train) ตัวแบบคณิตศาสตร์ และเรียงสรุปและข้อความตามความยาวที่กำหนดจากน้อยไปมาก หลังจากนั้นจะเป็นการสร้างตัวแบบทางคณิตศาสตร์ โดยการเริ่มที่สร้าง placeholder ของข้อมูลนำเข้า จากนั้นก็สร้างตัวเข้ารหัสและตัวถอดรหัส แล้วจึงเป็นการสอนตัวแบบทางคณิตศาสตร์ โดยเลือกข้อมูลมาชุดหนึ่งจากข้อมูลทั้งหมดจากนั้นก็ทำการสอนตัวแบบทางคณิตศาสตร์จากข้อมูลนั้น จะได้ตัวแบบทางคณิตศาสตร์ที่ได้รับการสอนแล้วซึ่งจะนำมาใช้หาสรุปจากข้อความที่ต้องการ ซึ่งคณะผู้วิจัยได้นำตัวแบบคณิตศาสตร์นี้มาปรับใช้กับโครงงานนี้ ซึ่งจะอธิบายวิธีการนำมาปรับใช้ในบทต่อไป

บทที่ 3
วิธีการวิจัยและพัฒนา

ในบทนี้ ผู้พัฒนาจะวิธีการวิจัยและพัฒนาโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล โดยจะกล่าวถึงข้อมูลที่นำมาใช้พัฒนา วิธีการพัฒนาและวัดผล ซึ่งแบ่งออกเป็นหัวข้อหลัก ดังนี้

1. วิธีการเก็บข้อมูลที่นำมาใช้ในการพัฒนาโปรแกรม
2. วิธีการปรับแต่งข้อมูลที่ใช้ในการพัฒนาก่อนที่จะเข้าสู่ตัวแบบทางคณิตศาสตร์
3. ตัวแบบทางคณิตศาสตร์ที่ใช้
4. วิธีการวัดผลโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล

3.1 วิธีการเก็บข้อมูลที่นำมาใช้ในการพัฒนาโปรแกรม

ในการศึกษาครั้งนี้คณะผู้วิจัยได้ใช้ข้อมูลจาก https://www.sportsmole.co.uk คณะผู้วิจัยได้ใช้ Beautiful Soup 3 ซึ่งเป็นคลังโปรแกรม (library) หนึ่งในภาษา python3 ที่มีไว้สำหรับดึงข้อมูลจากหน้าเว็บหรือ HTML มาใช้ในการดึงข้อมูลข่าวจากเว็บมาทั้งหมดก่อน จากนั้นจึงเลือกเก็บข่าวที่ตรงกับคำหลัก (keyword) ที่คณะผู้วิจัยกำหนดไว้กล่าวคือคำว่า "RED CARD" "UPDATE" "CHANCE" "GOAL" "HALF-TIME" และ "FULL-TIME" ซึ่งเป็นเนื้อหาข่าวหลักสำหรับกีฬาฟุตบอล โดยเลือกเก็บทั้งย่อหน้าที่มีคำหลักนั้น ๆ แล้วนำมารวมกันเป็นข้อมูลขาเข้า เพราะตัวแบบทางคณิตศาสตร์ที่ใช้นั้นไม่สามารถรับข้อมูลขาเข้าได้ทีละมาก ๆ เนื่องจากทรัพยากรในการทำงานไม่เพียงพอ จึงต้องเลือกข้อมูลขาเข้าที่สำคัญมาสรุปข่าว

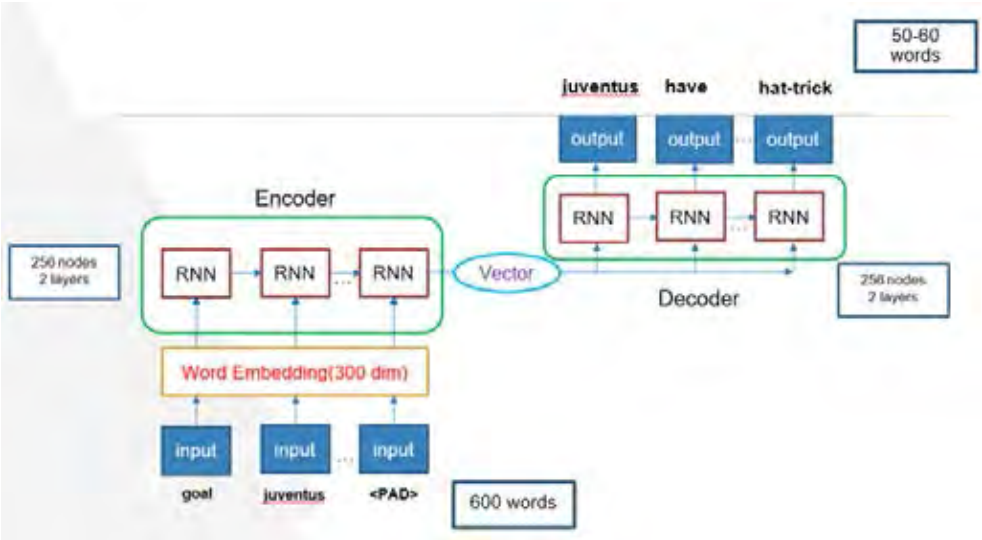
3.2 วิธีการปรับแต่งข้อมูลที่ใช้ในการพัฒนาก่อนที่จะเข้าสู่ตัวแบบทางคณิตศาสตร์

หลังจากที่ได้ข่าวมาเป็นที่เรียบร้อยจะทำความสะอาดข้อมูลด้วย regular expression หรือ regex ตัดข้อมูลจำพวก emoji หรือ สัญลักษณ์ต่าง ๆ ที่ไม่เกี่ยวกับเนื้อหาข่าวและในเนื้อหาตัวสรุปออก จากนั้นทำการตัดคำจากข้อความข่าวให้ออกมาเป็นคำเดี่ยว ๆ (Tokenize) แล้วจึงทำ dict ที่เปลี่ยนคำให้เป็นตัวเลขและเปลี่ยนตัวเลขเป็นคำ เพื่อที่จะสามารถทำการเข้าและถอดรหัสในตัวแบบทางคณิตศาสตร์ได้ จากนั้นก็ทำการเทียบกับคลังคำศัพท์ซึ่งในที่นี้คือ CN ว่ามีคำที่สามารถใช้ในการสอนตัวแบบทางคณิตศาสตร์ได้กี่คำ จากนั้นก็ทำ word embedding เปลี่ยนคำแต่ละคำเป็นเวกเตอร์ 300 มิติ หลังจากนั้นคำที่เหลือที่ไม่ได้อยู่ใน CN จะถูกนับเป็น UNK และเพิ่มลงใน dict ที่ทำไว้ก่อนหน้าเป็นตัวเลขสุดท้าย แล้วนำข้อมูลมาแบ่งข้อมูลออกเป็น 2 ชุดแบบสุ่ม (split test) ด้วยคลังโปรแกรม ของ scikit-learn ใช้ข้อมูลเพียง 80% ในการฝึกฝนตัวแบบทางคณิตศาสตร์ (train set) และทดสอบผลของตัวแบบทางคณิตศาสตร์ (test set) ด้วยข้อมูล 20% จากนั้นเรานำข่าวบางข่าวในข้อมูลชุดฝึกฝน ออกบางส่วนซึ่งเป็นเนื้อหาในแต่ละข่าวและสรุปในแต่ละข่าวที่ยาวเกินไป โดยกำหนดจำนวนคำในเนื้อหาข่าว และ จำนวนคำในเนื้อหาสรุปสูงสุดที่ 600 คำ เพื่อให้ทรัพยากรเพียงพอและจำนวน UNK ในแต่ละข่าวอยู่ที่ 50 คำ

3.3 ตัวแบบทางคณิตศาสตร์ที่ใช้

ทางคณะผู้วิจัยได้ใช้ตัวแบบทางคณิตศาสตร์ของงานวิจัยที่กล่าวไปในหัวข้อ 2.6 ข้างต้น เนื่องจากเป็นตัวแบบทางคณิตศาสตร์ที่ทำงานได้ผลลัพธ์ที่ดีอยู่แล้ว จึงปรับแค่ค่าของไฮเปอร์พารามิเตอร์เท่านั้น ผลที่ได้จะเป็นตัวเลข จะต้องทำการเปลี่ยนกลับเป็นคำ เพื่อให้ได้สรุปที่ต้องการ

ตัวแบบทางคณิตศาสตร์ที่นำมาใช้ในงานวิจัยนี้นั้นคือ sequence-to-sequence ที่ใช้ two-layered bidirectional RNN with LSTMs กับข้อมูลเข้า และใช้ two layers, each with an LSTM using bahdanau attention กับผลลัพธ์และใช้ word embedding 300 มิติ จะฝึกสอนตัวแบบทางคณิตศาสตร์ 10 ครั้ง ทำการทดสอบผลลัพธ์ 2 ครั้ง



รูปภาพที่ 3.1 แผนภาพแสดงหลักการทำงานของตัวแบบคณิตศาสตร์ sequence-to-sequence ที่ใช้ในโครงงาน

จากรูปภาพด้านบน การทำงานของตัวแบบทางคณิตศาสตร์สามารถอธิบายได้เป็นขั้นตอนดังนี้

1. เตรียมข้อมูลขาเข้าเป็นเวกเตอร์ของคำโดยกำหนดความยาวไม่เกิน 600 คำ ถ้าเกินจะคัดข่าวนั้นออก ถ้าไม่ถึง จะเติมคำว่า <PAD> แล้วนำเวกเตอร์นั้นเข้าสู่ชั้น word embedding
2. ในชั้น word embedding คำในแถวที่เป็นข้อมูลขาเข้า จะถูกแปลงเป็นเวกเตอร์ 300 มิติที่ประกอบไปด้วยตัวเลขโดยอ้างอิงจากคลังคำศัพท์ CN จากนั้นนำเวกเตอร์ที่แปลงแล้วนั้นเข้าสู่ส่วนเข้ารหัสต่อไป
3. เมื่อเข้าสู่ส่วนเข้ารหัสจะนำเวกเตอร์นั้นเข้าสู่เซลล์ RNN ที่มี 256 เซลล์ ส่วนเข้ารหัสมี 2 ชั้น ทีละคำเพื่อทำการเข้ารหัสและส่งเวกเตอร์ที่เข้ารหัสแล้วไปสู่ส่วนถอดรหัส
4. เมื่อเข้าสู่ส่วนถอดรหัสจะนำเวกเตอร์มาเข้าสู่เซลล์ RNN ที่มี 256 เซลล์ ส่วนถอดรหัสมี 2 ชั้น เพื่อทำการถอดรหัสมาเป็นคำที่รวมกันแล้วได้เป็นสรุปข่าว

5. ในส่วนของผลลัพธ์ ในตอนเรียนรู้ เราจะให้คู่สรุปข่าวจริงไว้ให้ตัวแบบทางคณิตศาสตร์เรียนรู้ แต่ในตอนทดสอบ เราจะให้แต่ตัวสคริปต์ข่าวเท่านั้น ตัวสรุปข่าวจริงจะใช้ในการเปรียบเทียบเพื่อวัดผล

3.4 วิธีการวัดผลโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล

3.5.1 BLEU score

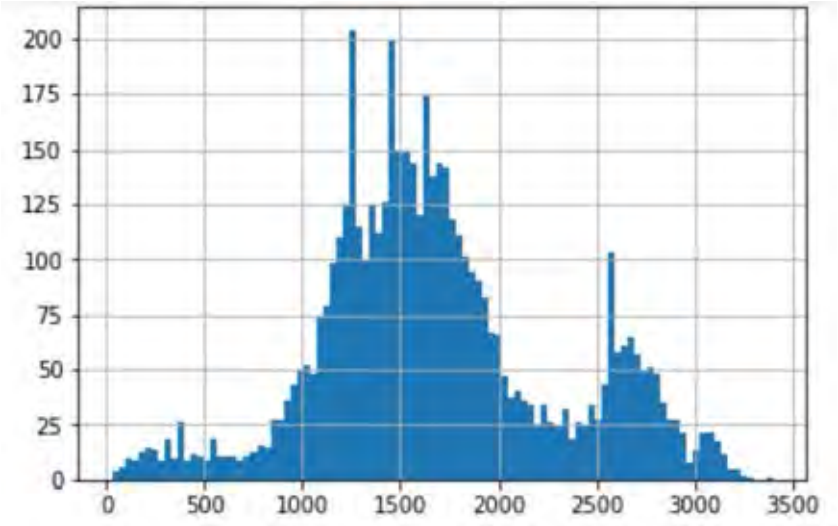
ทางคณะผู้วิจัยได้ทำการวัดผลสรุปจากตัวแบบทางคณิตศาสตร์เปรียบเทียบกับสรุปที่ได้จากการสุ่ม โดยการใช้ BLEU score โดยใช้แค่ผล 1-gram และ 2-gram เทียบกันเท่านั้น

บทที่ 4
ผลการวิจัย

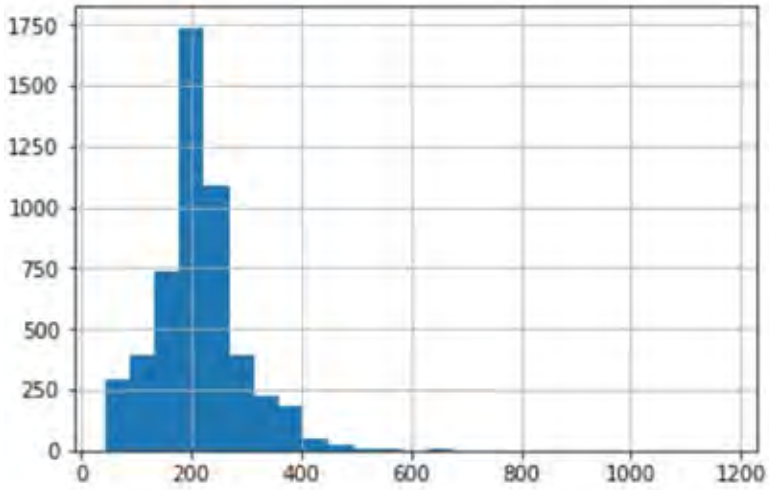
ในบทนี้จะกล่าวถึงผลของการดำเนินการวิจัยสำหรับการสรุปข่าวอัตโนมัติโดยใช้ตัวแบบทางคณิตศาสตร์ sequence-to-sequence

4.1 ลักษณะของข้อมูลที่ใช้

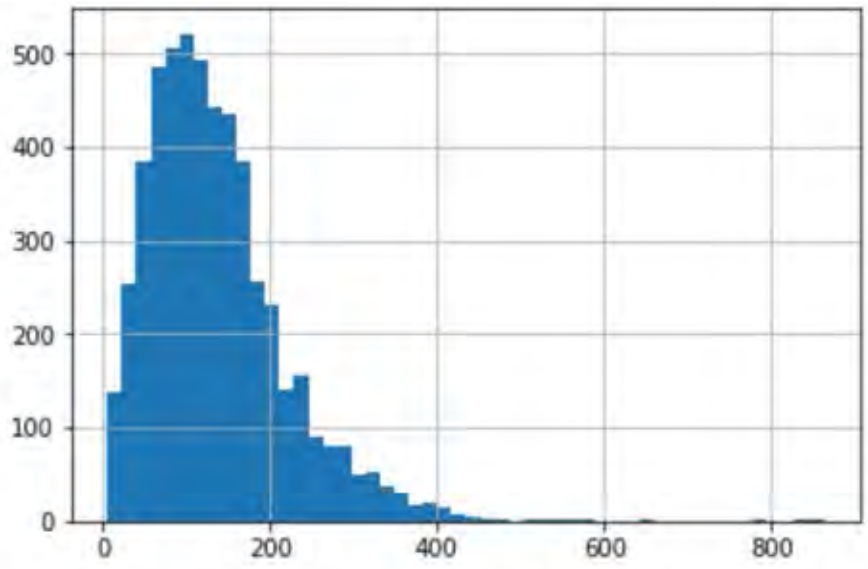
ข้อมูลข่าวที่ได้ดึงมาจาก https://www.sportsmole.co.uk จากหัวข้อ 3.1 ก่อนการกรองคำด้วยคำหลักจะมีความยาวโดยเฉลี่ยตั้งแต่ 1,000 ถึง 2,000 คำและมีการกระจายตัวดังรูปที่ 4.1 เนื่องจากจำนวนคำในเนื้อหาข่าวมีลักษณะที่ไม่เพียงพอที่ทรัพยากรจะใช้ทำงานได้ จึงต้องทำการดึงเนื้อหาข่าวที่มีคำหลักกล่าวคือ "RED CARD" "UPDATE" "CHANCE" "GOAL" "HALF-TIME" และ "FULL-TIME" เพื่อลดจำนวนคำในเนื้อหาข่าวให้สามารถทำงานได้สำหรับทรัพยากรที่มีทำให้มีความยาวโดยรวมลดลงดังรูปที่ 4.3



รูปภาพที่ 4.1 กราฟแสดง จำนวนคำของสคริปต์ข่าวแต่ละอัน



รูปภาพที่ 4.2 กราฟแสดง จำนวนคำของสรุปข่าวแต่ละอัน



รูปภาพที่ 4.3 กราฟแสดง จำนวนคำของข่าวที่ได้รับการกรองคำแล้วแต่ละอัน

ส่วนหนึ่งของตัวอย่างสคริปต์ข่าวหลังจากกรองตามคำหลักที่กำหนดไว้แล้ว

GOAL! Juventus 1-0 Atletico Madrid (Cristiano Ronaldo) Reuters GOAL! Game on! Bernardeschi delivers a superb cross from the left-hand side of the penalty area and it finds Ronaldo, who powers a header inside the near post after getting the better of Juanfran. CHANCE! Spinazzola delivers another peach of a cross from the left flank but on this occasion, Ronaldo heads wide of the post. The Portuguese should have done better. CHANCE! An outswinging corner from the right finds Chiellini by the penalty spot, and the defender sees his powerful header tipped over the crossbar by Oblak. Juventus are finishing this first half well on top. CHANCE! Big chance for Atletico! A cross from the right appears perfect for Morata to find the far corner with his header, but the Spaniard heads over! How important could that be? GOAL! Juventus 2-0 Atletico Madrid (Cristiano Ronaldo) Reuters GOAL! Juventus are back on level terms! Initially, it looks like Oblak has pulled off an oustanding save, but goalline technology rules that Ronaldo's towering header from 12 yards has gone a matter of inches over the line. CHANCE! What a chance for Kean! A long ball is deflected into the path of Kean and the youngster is through on goal. However, the forward scuffs his volley from 10 yards wide of the post. GOAL! Juventus 3-0 Atletico Madrid (Cristiano Ronaldo) GOAL! Ronaldo steps up and dispatches the penalty into the bottom corner. Oblak dived the opposite way. Atletico need a goal!

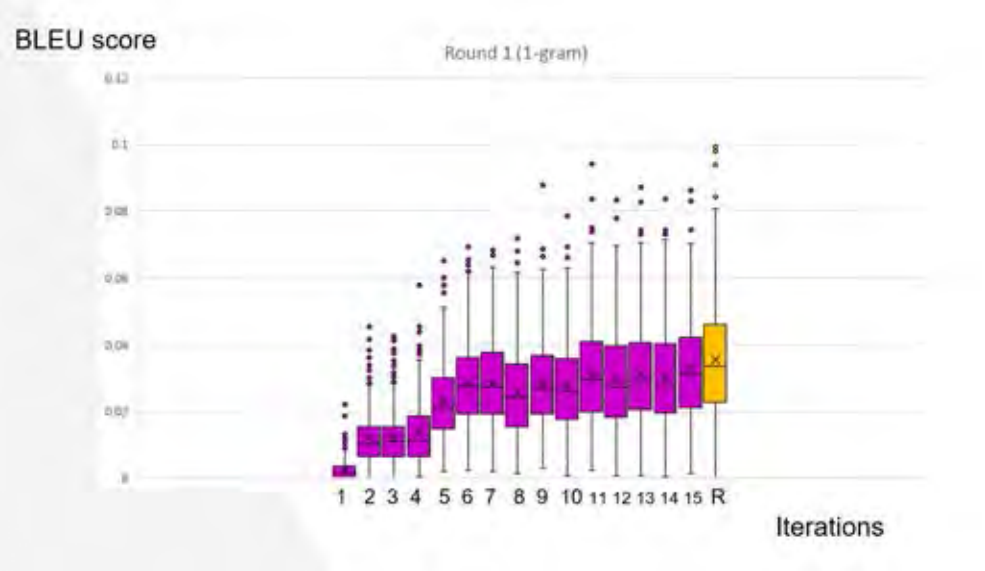
ส่วนหนึ่งของตัวอย่างข้อมูลสรุปของสคริปต์ข่าวด้านบน

Juventus have progressed through to the Champions League quarter-finals with a 3-0 victory over Atletico Madrid on Tuesday night.Massimiliano Allegri's side trailed 2-0 on

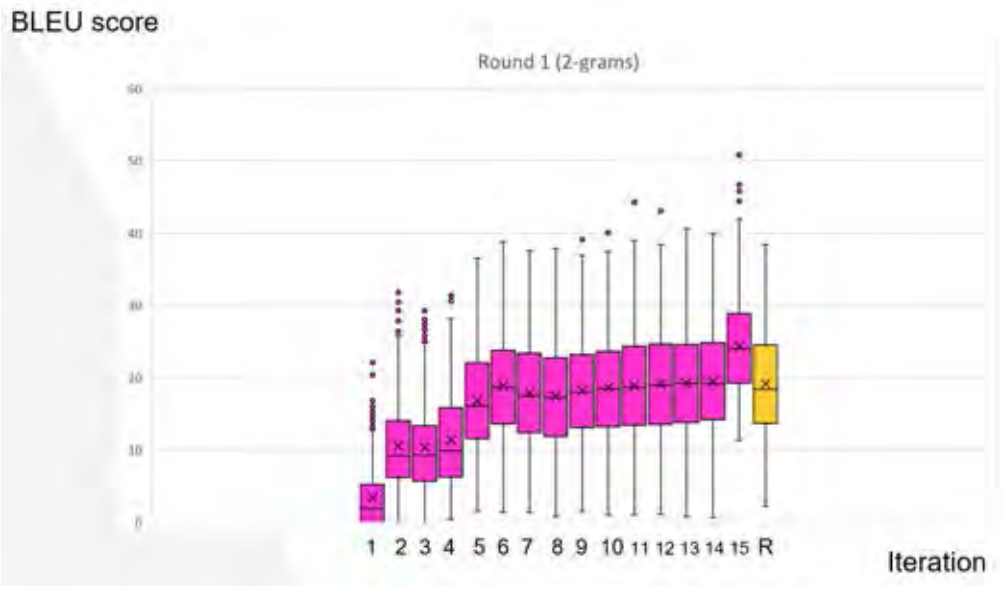
aggregate after the first leg in Spain, but Cristiano Ronaldo inspired the Italian champions to a deserved win.Ronaldo got Juventus back on level terms with two towering headers, with his hat-trick being completed from the penalty spot with six minutes remaining.Find out how all of the action unfolded in Turin courtesy of Sports Mole's minute-by-minute updates below.Massimiliano Allegri's side trailed 2-0 on aggregate after the first leg in Spain, but Cristiano Ronaldo inspired the Italian champions to a deserved win.Ronaldo got Juventus back on level terms with two towering headers, with his hat-trick being completed from the penalty spot with six minutes remaining.Find out how all of the action unfolded in Turin courtesy of Sports Mole's minute-by-minute updates below.Ronaldo got Juventus back on level terms with two towering headers, with his hat-trick being completed from the penalty spot with six minutes remaining.Find out how all of the action unfolded in Turin courtesy of Sports Mole's minute-by-minute updates below.

4.2 ความถูกต้องของตัวแบบคณิตศาสตร์ sequence-to-sequence ในการสรุปข่าว

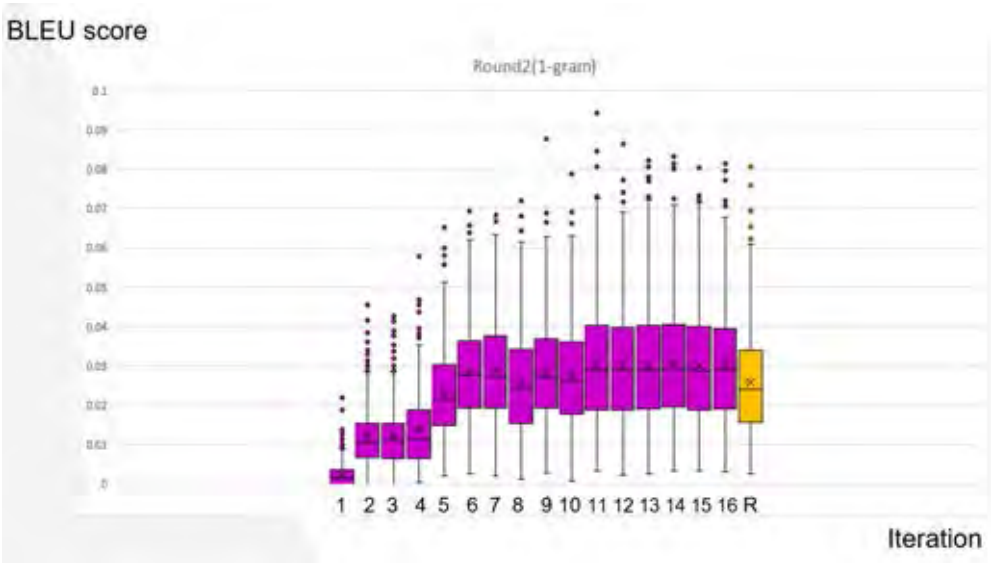
หลังจากที่ตัวแบบทางคณิตศาสตร์ได้ทำนายสรุปออกมาแล้ว นอกจากจะใช้การวัดผลแบบ 1-gram BLEU scores แล้ว คณะผู้วิจัยได้ใช้การวัดผลแบบ 2-gram BLEU score ซึ่งสนใจสองคำที่อยู่ติดกันด้วย วิธีที่ได้คะแนนนี้สูงกว่ามีแนวโน้มที่จะอ่านรู้เรื่องได้มากกว่า



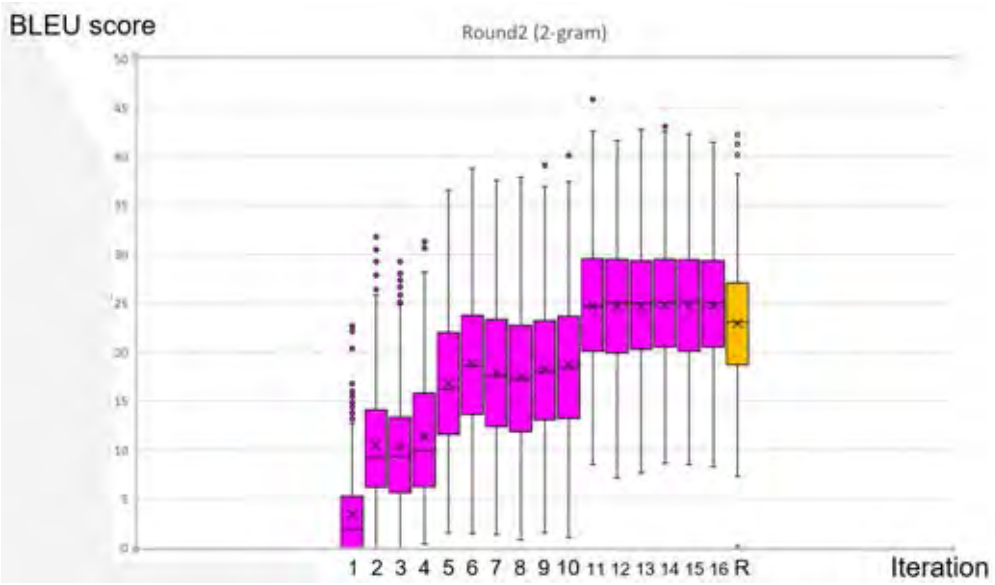
รูปภาพที่ 4.4 แสดง Box plot ของผล BLEU score (1-gram) รอบที่ 1



รูปภาพที่ 4.5 Box plot แสดงผล BLEU score (2-gram) รอบที่ 1



รูปภาพที่ 4.6 Box plot แสดงผล BLEU score (1-gram) รอบที่ 2



รูปภาพที่ 4.7 Box plot แสดงผล BLEU score (2-gram) รอบที่ 2

ตัวอย่างข่าว

สคริปต์ข่าว

goal juventus 1-0 atletico madrid cristiano ronaldo reuters goal game bernardeschi delivers superb cross left-hand side penalty area finds ronaldo powers header inside near post getting better juanfran chance spinazzola delivers another peach cross left flank occasion ronaldo heads wide post portuguese done better chance outswinging corner right finds chiellini penalty spot defender sees powerful header tipped crossbar oblak juventus finishing first half well top chance big chance atletico cross right appears perfect morata find far corner header spaniard heads important could half time juventus 1-0 atletico madrid goal juventus 2-0 atletico madrid cristiano ronaldo reuters goal juventus back level terms initially looks like oblak pulled oustanding save goalline technology rules ronaldo towering header 12 yards gone matter inches line chance chance kean long ball deflected path kean youngster goal however forward scuffs volley 10 yards wide post goal juventus 3-0 atletico madrid cristiano ronaldo goal ronaldo steps dispatches penalty bottom corner oblak dived opposite way atletico need goal full time juventus 3-0 atletico madrid

สรุปข่าวจริง

juventus have progressed through to the champions league quarter-finals with a 3-0 victory over atletico madrid on tuesday night massimiliano allegri s side trailed 2-0 on aggregate after the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below massimiliano allegri s side trailed 2-0 on aggregate after the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

juventus have progressed through to the champions league quarter-finals with a 3-0 victory over atletico madrid on tuesday night massimiliano allegri s side trailed 2-0 on aggregate after the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with two towering headers with his hat-trick

(ดูตัวอย่างผลลัพธ์เพิ่มเติมได้ในภาคผนวก ก)

จากผลการวิจัยพบว่า ผลของ BLEU score (1-gram) ของสรุปที่ได้จากตัวแบบทางคณิตศาสตร์นั้นมากกว่าหรือเท่ากับสรุปที่ได้จากการสุ่ม และผล BLEU score (2-gram) ของสรุปที่ได้จากตัวแบบทางคณิตศาสตร์นั้นมากกว่าสรุปที่ได้จากการสุ่ม และเราได้ทำการวัดผลทั้งหมด 2 รอบเพื่อเป็นตัวยืนยันผลการวิจัย ส่วนผลการสัมภาษณ์จากกลุ่มตัวอย่างพบว่าสรุปจากตัวแบบทางคณิตศาสตร์นั้นสามารถบอกได้ว่าทีมใดชนะในนัดการแข่งขันนั้นๆและอ่านได้เข้าใจมากกว่าการสรุปข่าวที่ได้จากการสุ่ม

บทที่ 5
ข้อสรุปและข้อเสนอแนะ

5.1 สรุปผลการดำเนินงาน

ในงานนี้คณะผู้วิจัยได้ใช้ตัวแบบคณิตศาสตร์ sequence-to-sequence ในการพัฒนาโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอลโดยทำการฝึกฝน 2 ครั้งก่อนที่จะสร้างเป็นตัวแบบคณิตศาสตร์ที่ใช้

ผลการทดลองของ BLEU score บอกได้ว่าสรุปจากตัวแบบทางคณิตศาสตร์ดีกว่าสรุปจากการสุ่ม รวมถึงผลจากการสัมภาษณ์บอกได้ว่าสรุปจากตัวแบบทางคณิตศาสตร์นั้นสามารถบอกได้ว่าอ่านรู้เรื่องมากกว่าสรุปที่ได้จากการสุ่ม

5.2 เป้าหมายในอนาคต

คณะผู้วิจัยจะพัฒนาตัวแบบทางคณิตศาสตร์และปรับปรุงแก้ไขข้อมูลให้ได้ผลลัพธ์ดียิ่งขึ้น และจะนำตัวแบบทางคณิตศาสตร์ไปพัฒนาให้เป็นระบบต่อไป

5.3 ปัญหาของงานวิจัยและวิธีการแก้ไข

ปัญหาที่ 1 ระหว่างทำการวิจัยได้ลองทดสอบการทำงานของตัวแบบทางคณิตศาสตร์พบว่าไม่สามารถทำงานได้

วิธีแก้ไขปัญหา ต้องแก้ไขรุ่นของภาษาและคลังโปรแกรมที่ใช้ ตัวแบบทางคณิตศาสตร์จึงสามารถทำงานได้

ปัญหาที่ 2 ระหว่างทำการวิจัยได้ลองทดสอบการทำงานของตัวแบบทางคณิตศาสตร์ โดยป้อนข้อมูลขาเข้าโดยที่ยังไม่ได้เลือกจากคำหลัก พบว่าไม่สามารถทำงานได้เพราะทรัพยากรในการทำงานไม่เพียงพอ

วิธีแก้ไขปัญหา ตัดข้อมูลขาเข้าบางส่วนออกโดยเลือกใช้ย่อหน้าที่มีคำหลักเท่านั้น

ปัญหาที่ 3 Router รีสตาร์ทอินเทอร์เน็ตเอง ทำให้การทำงานของตัวแบบทางคณิตศาสตร์หยุดทำงาน ทำให้ตัวแบบทางคณิตศาสตร์เรียนรู้ได้ไม่มากพอ

วิธีแก้ไขปัญหา บันทึกผลตัวแบบทางคณิตศาสตร์ที่ดีสุดก่อนที่ Router จะรีสตาร์ท

ปัญหาที่ 4 เนื้อหาข่าวและสรุปบางส่วนในเว็บมีภาษาที่กำกวม เช่น เนื้อหาของข่าวที่ดึงออกมาไม่มีการแจ้งรายละเอียดที่เพียงพอสำหรับการสรุป หรือ เนื้อหาข่าวที่เป็นสรุป สรุปออกมาได้ไม่ดี

วิธีแก้ไขปัญหา พยายามนำข่าวประเภทนี้ออกให้มากที่สุด แต่เนื่องจากมีข่าวอยู่เป็นจำนวนมากจึงไม่สามารถเอาออกได้หมด

รายการอ้างอิง

[1] Jeffrey Pennington. GloVe: Global Vectors for Word Representation. Available from https://nlp.stanford.edu/projects/glove/ [2014, August]

[2] Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." *arXiv preprint arXiv:1602.06023* (2016).

[3] Graham, Yvette. "Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE." *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015.

[4] Amazon Fine Food Reviews. Available from: https://www.kaggle.com/snap/amazon-fine-food-reviews [2018, October 31]

ภาคผนวก ก

ตัวอย่างข่าวที่ 1

สคริปต์ข่าว

chance bright start spurs everton carve first chance seamus coleman lofts ball path leon osman drops shoulder beat bentaleb opens opportunity midfielder get enough curl shot edge area ball drifts wide target chance first real sight goal spurs although relatively speculative effort christian eriksen player hit ball 30 yards dip enough lands roof howard net chance big opportunity spurs defender dawson mistimed jump could well given hosts lead rose well meet eriksen corner jump little early met ball way result ball flew harmlessly bar chance mirallas greedy everton would probably 1-0 first belgian shows great pace beat vertonghen dawson done naismith found acres space inside area mirallas went goal instead fired way top understandably naismith frustrated teammate half time spurs 0-0 everton chance neat link play everton duo baines pienaar left frees mirallas sidesteps walker leave shooting chance 18 yards goal goes far post always bending away target lloris watches safety chance meanwhile end adebayor wriggles free barry meet eriksen corner header powerful enough much height skims crossbar goal spurs 1-0 everton adebayor chance dembele collects possession midway inside everton half spins spot rolls ball adebayor feet turns quickly shoots little backlift 20 yards close enough worry howard ball flies foot crossbar full time spurs 1-0 everton

สรุปข่าวจริง

tottenham hotspur played host to everton at white hart lane with both sides looking to close the gap on liverpool in fourth as it was following a goalless first half emmanuel adebayor scored the only goal of the game after the restart to seal all three points for the home side you can find out how the match unfolded with our minute-by-minute updates below as it was following a goalless first half emmanuel adebayor scored the only goal of the game after the restart to seal all three points for the home side you can find out how the match unfolded with our minute-by-minute updates below you can find out how the match unfolded with our minute-by-minute updates below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

tottenham hotspur have moved into the top of the premier league table courtesy of a 1-0 win over galatasaray at the hands of the match the visitors dominated the lead in the first half but they were unable to breach the deadlock as the hosts were held out to secure their place in the table

ตัวอย่างข่าวที่ 2

สคริปต์ข่าว

chance first opportunity evening comes home side jese dances osasuna box finding benzema frenchman denied fine arribas challenge vital moment jese collects moments later denied inside box bright opening six minutes copa del rey encounter chance plenty osasuna box javi garcia pleased opening nine minutes side showing nice touches madrid half riera possession right finding cejudo inside box low effort pushed wide post iker casillas goal real madrid 1-0 osasuna benzema chance opportunity osasuna free kick torres felled arbeloa outside box resulting set piece met oier header drops wide post javi garcia seen enough side suggest cause problems madrid chance wonderful chance madrid make 2-0 jese drives box appearing put plate ronaldo arribas comes nowhere make fine diving challenge chances ends first period enters latter stages chance incredible chance osasuna level torres finds oier inside box low cross defender flash effort wide post opportunities come around often bernabeu half-time real madrid 1-0 osasuna chance incredible chance ronaldo bale finds portuguese inside box fires effort straight riesgo close range goal real madrid 2-0 osasuna jese chance another wonderful chance ronaldo modric finds run forward inside area somehow turns wide post osasuna sorts bother 63 minutes clock chance incredible chance madrid benzema finds bale inside box welshman effort appears heading bottom corner strikes teammate ronaldo dropping wide post osasuna strike lucky 15 minutes remaining full-time real madrid 2-0 osasuna

สรุปข่าวจริง

real madrid welcomed osasuna to the bernabeu on thursday for the first leg of their last-16 copa del rey clash a lacklustre first period brought just the one goal with karim benzema heading home after 19 minutes madrid made it 2-0 on the hour mark through jese rodriguez as carlo ancelotti s side secured an advantage ahead of the second leg read how it all unfolded in sports mole s minute-by-minute live commentary of the action below a lacklustre first period brought just the one goal with karim benzema heading home after 19 minutes madrid made it 2-0 on the hour mark through jese rodriguez as carlo ancelotti s side secured an advantage ahead of the second leg read how it all unfolded in sports mole s minute-by-minute live commentary of the action below madrid made it 2-0 on the hour mark through jese rodriguez as carlo ancelotti s side secured an advantage ahead of the second leg read how it all unfolded in sports mole s minute-by-minute live commentary of the action below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

real madrid welcomed osasuna to osasuna on sunday night in the second leg of their copa del rey clash at the bernabeu courtesy of the first leg the home side led 2-0 ahead after the interval courtesy of a smart finish from karim benzema and karim benzema made it 2-0 in the second

ตัวอย่างข่าวที่ 3

สคริปต์ข่าว

goal juventus 1-0 atletico madrid cristiano ronaldo reuters goal game bernardeschi delivers superb cross left-hand side penalty area finds ronaldo powers header inside near post getting better juanfran chance spinazzola delivers another peach cross left flank occasion ronaldo heads wide post portuguese done better chance outswinging corner right finds chiellini penalty spot defender sees powerful header tipped crossbar oblak juventus finishing first half well top chance big chance atletico cross right appears perfect morata find far corner header spaniard

heads important could half time juventus 1-0 atletico madrid goal juventus 2-0 atletico madrid cristiano ronaldo reuters goal juventus back level terms initially looks like oblak pulled oustanding save goalline technology rules ronaldo towering header 12 yards gone matter inches line chance chance kean long ball deflected path kean youngster goal however forward scuffs volley 10 yards wide post goal juventus 3-0 atletico madrid cristiano ronaldo goal ronaldo steps dispatches penalty bottom corner oblak dived opposite way atletico need goal full time juventus 3-0 atletico madrid

สรุปข่าวจริง

juventus have progressed through to the champions league quarter-finals with a 3-0 victory over atletico madrid on tuesday night massimiliano allegri s side trailed 2-0 on aggregate after the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below massimiliano allegri s side trailed 2-0 on aggregate after the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

juventus have progressed through to the champions league quarter-finals with a 3-0 victory over atletico madrid on tuesday night massimiliano allegri s side trailed 2-0 on aggregate after

the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with

ตัวอย่างข่าวที่ 4

สคริปต์ข่าว

chance wonderful chance away side take lead moreno breaks atletico box striker cannot find route past brilliant oblak spread half-time atletico 0-0 espanyol chance another wonderful chance espanyol moreno sets baptistao inside atletico box oblak make another important save team chance another chance home side carrasco delivers dangerous low cross griezmann reyes hand make fine clearance pressure continues chance stunning chance griezmann gaitan finds frenchman inside espanyol box fires straight arms diego lopez score full-time atletico 0-0 espanyol

สรุปข่าวจริง

atletico madrid welcomed in-form espanyol to the vicente calderon in la liga on saturday night the home side enjoyed close to 70 possession in the first 45 minutes but a strong rear-guard action from espanyol saw the two teams enter the interval on level terms atletico continued to press in the second period but diego simeone s side could not breach the espanyol defence as the points were shared in the spanish capital read how it all unfolded in sports mole s minute-by-minute live commentary of the action below the home side enjoyed close to 70 possession in the first 45 minutes but a strong rear-guard action from espanyol saw the two teams enter the interval on level terms atletico continued to press in the second period but diego simeone s side could not breach the espanyol defence as the points were shared in the spanish capital read how it all unfolded in sports mole s minute-by-minute live commentary of the action below atletico continued to press in the second period but diego simeone s side could not breach the espanyol defence as the points were shared in the

spanish capital read how it all unfolded in sports mole s minute-by-minute live commentary of the action below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

atletico madrid welcomed espanyol to the vicente calderon on saturday night knowing that three points would move them in the second leg of the season but neither could make the breakthrough as atletico held on a 0-0 draw in the first leg read how it all unfolded in sports mole s minute-by-minute live commentary of

ตัวอย่างข่าวที่ 5

สคริปต์ข่าว

goal juventus 1-0 atletico madrid cristiano ronaldo reuters goal game bernardeschi delivers superb cross left-hand side penalty area finds ronaldo powers header inside near post getting better juanfran chance spinazzola delivers another peach cross left flank occasion ronaldo heads wide post portuguese done better chance outswinging corner right finds chiellini penalty spot defender sees powerful header tipped crossbar oblak juventus finishing first half well top chance big chance atletico cross right appears perfect morata find far corner header spaniard heads important could half time juventus 1-0 atletico madrid goal juventus 2-0 atletico madrid cristiano ronaldo reuters goal juventus back level terms initially looks like oblak pulled oustanding save goalline technology rules ronaldo towering header 12 yards gone matter inches line chance chance kean long ball deflected path kean youngster goal however forward scuffs volley 10 yards wide post goal juventus 3-0 atletico madrid cristiano ronaldo goal ronaldo steps dispatches penalty bottom corner oblak dived opposite way atletico need goal full time juventus 3-0 atletico madrid

สรุปข่าวจริง

juventus have progressed through to the champions league quarter-finals with a 3-0 victory over atletico madrid on tuesday night massimiliano allegri s side trailed 2-0 on aggregate after

the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below massimiliano allegri s side trailed 2-0 on aggregate after the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below ronaldo got juventus back on level terms with two towering headers with his hat-trick being completed from the penalty spot with six minutes remaining find out how all of the action unfolded in turin courtesy of sports mole s minute-by-minute updates below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

juventus have progressed through to the champions league quarter-finals with a 3-0 victory over atletico madrid on tuesday night massimiliano allegri s side trailed 2-0 on aggregate after the first leg in spain but cristiano ronaldo inspired the italian champions to a deserved win ronaldo got juventus back on level terms with two towering headers with

ตัวอย่างข่าวที่ 6

สคริปต์ข่าว

goal tottenham hotspur 1-0 borussia dortmund son heung-min goal tottenham hotspur 1-1 borussia dortmund andriy yarmolenko goal tottenham hotspur 2-1 borussia dortmund harry kane chance lloris keeps sweeper keeper tag intact racing getting ball ahead aubameyang end spurs created first chance since going ahead second time kane wasted blasting bar inside box chance two big chances tottenham space two minutes simple like kane latches ball top squares son took one touch set curling bar great position goal disallowed second time tonight

dortmund goal ruled shocker decision truth gabonese striker aubameyang well onside volleying ball past lloris back post goal tottenham hotspur 3-1 borussia dortmund harry kane chance another chance comes goes tottenham christian eriksen bends ball ball - via deflection went unspotted officials - perhaps could played son alongside red card absolute shocker decision referee poor evening jan vertonghen caught gotze stray arm worthy yellow card surely late impact game

สรุปข่าวจริง

harry kane scored twice as tottenham hotspur picked up a rare victory at wembley stadium overcoming borussia dortmund 3-1 in their champions league opener the english striker restored his side s advantage 15 minutes in after andriy yarmolenko cancelled out son heung-min s near-post finish a second of the evening for kane and a third for tottenham arrived on the hour through a well-taken strike just moments after the visitors wrongly had a goal ruled out relive how the 90 minutes of action unfolded with sports mole s live text coverage below the english striker restored his side s advantage 15 minutes in after andriy yarmolenko cancelled out son heung-min s near-post finish a second of the evening for kane and a third for tottenham arrived on the hour through a well-taken strike just moments after the visitors wrongly had a goal ruled out relive how the 90 minutes of action unfolded with sports mole s live text coverage below a second of the evening for kane and a third for tottenham arrived on the hour through a well-taken strike just moments after the visitors wrongly had a goal ruled out relive how the 90 minutes of action unfolded with sports mole s live text coverage below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

tottenham hotspur have beaten a place in the quarter-finals of the champions league this evening courtesy of a 3-1 victory over borussia dortmund at the emirates stadium this evening the gunners were already beaten to 10 men in the first half but they were reduced to 10 men in the second half when pierre-emerick aubameyang scored his hat-trick

ตัวอย่างข่าวที่ 7

สคริปต์ข่าว

chance first proper opening match comes fulham way sidwell heads duff cross bar probably least worked cech chance another opportunity visitors bent chases long ball top parker cech rushes line head ball away looks like bent might afforded plenty space final third evening could dangerous chelsea given right service chance ball bounces around fulham box causing confusion ivanovic finally crack goal stockdale gets push ball away corner comes nothing half time chelsea 0-0 fulham update fairly eye-catching results premier league today missed earlier goal chelsea 1-0 fulham oscar chance fulham come close snatching equaliser almost instantly sidwell target header following kasami free kick visitors really goal name today chance great play torres weaving way fulham defence final pass ramires stray one visitors hack ball clear chance great defending amorebieta blocked ramires shot resulting corner oscar comes nothing brazilian crossing wide poor evening goal chelsea 2-0 fulham mikel chance richardson tries cross ball chelsea area instead ball nearly creeps cech near post however goalkeeper push away danger resulting corner cleared full time chelsea 2-0 fulham

สรุปข่าวจริง

chelsea returned to the top of the premier league table on saturday thanks to a 2-0 win over fulham at stamford bridge following a goalless first half the hosts went ahead soon after the restart as oscar converted from close range the points were then sealed late on when jon obi mikel got on the scoresheet with a rare goal for the blues read sports mole s minute-by-minute report below to find out how the action panned out between the local rivals following a goalless first half the hosts went ahead soon after the restart as oscar converted from close range the points were then sealed late on when jon obi mikel got on the scoresheet with a rare goal for the blues read sports mole s minute-by-minute report below to find out how the

action panned out between the local rivals the points were then sealed late on when jon obi mikel got on the scoresheet with a rare goal for the blues read sports mole s minute-by-minute report below to find out how the action panned out between the local rivals

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

chelsea have moved into the top of the premier league table courtesy of a 2-0 victory over fulham at the king power stadium the visitors took the lead in the first half but they were able to breach the deadlock as the visitors were forced to breach the deadlock as the result sees the result to have their first

ตัวอย่างข่าวที่ 8

สคริปต์ข่าว

chance opportunity oscar give chelsea lead west ham full-back jenkinson inexplicably fails clear willian cross oscar pounces first touch good adrian beat six yards brazilian lifts effort well crossbar chance lovely football chelsea willian pulls ball back edge west ham area terry people sends instant cross back post defensive partner cahill gets well however looping header flies couple yards bar chelsea far goal chelsea 1-0 west ham terry chance chelsea creating openings hazard patient edge area picking ivanovic wide right turn sends low cross area costa ball bounces striker sends effort flashing crossbar half-time chelsea 1-0 west ham goal chelsea 2-0 west ham costa chance costa could made 3-0 ivanovic cross right peach costa volley mishit ball flies high wide handsome full time chelsea 2-0 west ham <UNK>

สรุปข่าวจริง

chelsea recorded a 2-0 victory over west ham united at stamford bridge this afternoon to open up a six-point lead at the top of the premier league table the scoring was opened by skipper john terry during the first half before diego costa made sure of the outcome after the

restart find out how the match unfolded with sports mole s minute-by-minute updates below the scoring was opened by skipper john terry during the first half before diego costa made sure of the outcome after the restart find out how the match unfolded with sports mole s minute-by-minute updates below find out how the match unfolded with sports mole s minute-by-minute updates below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

chelsea have moved into the top of the premier league table courtesy of a 2-0 victory over west ham united at the king power stadium the hosts took the lead in the first half when eden hazard doubled a 2-0 lead into the bottom corner before the break the hosts doubled

ตัวอย่างข่าวที่ 9

สคริปต์ข่าว

chance vardy score getting edgy mahrez glides forward two one tees vardy 10 yards blazes ball crossbar chance vardy put leicester quarter-finals drinkwater reverse ball finds striker inside penalty area header eight yards parried behind rico red card nasri sent stupid reacts slightest touches vardy two pushes heads forward came together vardy makes definite yellow nasri second chance easiest chances vardy rushes opportunity 12 yards volley goes wide frustrated goal leicester city 2-0 sevilla albrighton goal leicester city 1-0 sevilla morgan chance sevilla look constant threat sarabia would expected better 20-yard strike left foot instead drags wide post chance leicester behind brilliant move sevilla ends nasri wrong-footing morgan denied schmeichel near post around eight yards frenchman score <UNK>

สรุปข่าวจริง

leicester city have reached the champions league quarter-finals with a 2-0 victory over sevilla at the king power stadium sevilla held a 3-2 advantage after the first leg in spain but leicester

went ahead in the tie when wes morgan bundled the ball home in the first half nine minutes after the restart the foxes were in dreamland as marc albrighton netted a second and the remarkable triumph was capped off as kasper schmeichel saved a penalty from steven n zonzi read below to see how the action unfolded in the east midlands sevilla held a 3-2 advantage after the first leg in spain but leicester went ahead in the tie when wes morgan bundled the ball home in the first half nine minutes after the restart the foxes were in dreamland as marc albrighton netted a second and the remarkable triumph was capped off as kasper schmeichel saved a penalty from steven n zonzi read below to see how the action unfolded in the east midlands nine minutes after the restart the foxes were in dreamland as marc albrighton netted a second and the remarkable triumph was capped off as kasper schmeichel saved a penalty from steven n zonzi read below to see how the action unfolded in the east midlands

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

leicester city have have a 2-0 win over sevilla at the king power stadium this afternoon the visitors took the lead in the first half when riyad mahrez curled a penalty from the penalty spot before the break the visitors were not have to have a point through the first leg but they could

ตัวอย่างข่าวที่ 10

สคริปต์ข่าว

chance man city best opening match far sterling released behind defence first time cuts back inside robertson picking silva strike convincing van dijk makes block goal man city 1-0 liverpool sergio aguero goal man city make breakthrough aguero lashes finish roof net tight angle chance robertson cross cleared far alexander-arnold goes goal half-volley outside area decent strike cuts across much hits stanchion behind goal difficult one sight goal liverpool goal manchester city 1-1 liverpool roberto firmino goal liverpool equaliser fourth goal two games roberto firmino goal manchester city 2-1 liverpool leroy sane goal manchester city regain lead sane

picks bottom corner pinpoint accuracy chance big chance city kill game hit liverpool counter sterling releasing aguero goal argentine tries take ball around alisson liverpool keeper really well turn behind corner chance liverpool spring life win ball back inside city half ball played salah egyptian darts front laporte angle ederson able turn shot wide corner chance another chance resulting corner ball drops back post wijnaldum fires back middle cleared inside six-yard box van dijk cannot get follow-up shot away chance huge chances city put game bed bernardo skips past lovren denied alisson ball bounces dangerous area sterling smashes effort target simply score

สรุปข่าวจริง

manchester city inflicted a first league defeat of the season on title rivals liverpool this evening courtesy of a 2-1 victory at the etihad stadium sergio aguero blasted the hosts into the lead five minutes before half time only for roberto firmino to level things up for the league leaders shortly after the hour mark the champions responded within eight minutes though and leroy sane s pinpoint finish proved to be the winner as pep guardiola s side breathed new life into their title defence find out how all of the action unfolded courtesy of sports mole s minute-by-minute updates below

สรุปข่าวที่เป็นผลลัพธ์จากตัวแบบทางคณิตศาสตร์

manchester city have moved into the top of the premier league table courtesy of a 2-1 victory over liverpool at the etihad stadium this evening the hosts took the lead in the first half through the first leg but sergio aguero put the lead in the closing stages of the second half before sergio aguero and caglar soyuncu

กลุ่ม COMP04 สอบวันพฤหัสบดีที่ 22 พฤศจิกายน 2561 เวลา 10.00 น. ห้อง 1008A
เอกสารนี้ได้รับการอนุมัติจากอาจารย์ที่ปรึกษาแล้ว
ลงชื่อ .....
(วันที่ .....)

แบบเสนอหัวข้อโครงงาน รายวิชา 2301399 Project Proposal

ปีการศึกษา 2561

ชื่อโครงงาน (ภาษาไทย) โปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอล

ชื่อโครงงาน (ภาษาอังกฤษ) Soccer News Summarization

อาจารย์ที่ปรึกษา อ.ดร. นฤมล ประทานวณิช

ผู้ดำเนินการ นาย กรวิชญ์ กำปั่นทอง

นาย อภิชัย สมนาม 5833667223

สาขาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

หลักการและเหตุผล

ปัจจุบันมีข่าวกีฬามากมายที่คนบนโลกโซเชียลส่วนใหญ่นิยมเข้าไปอ่าน หนึ่งในข่าวกีฬาที่ได้รับความนิยมเป็นอย่างมากคือ ข่าวกีฬาฟุตบอลซึ่งมีหลากหลายรายการแข่งขัน ไม่ว่าจะเป็น การแข่งขันฟุตบอลโลก การแข่งขันของลีคแต่ละประเทศ รวมถึงนัดอุ่นเครื่องอีกมากมาย ซึ่งจะเห็นได้ว่ามีนัดการแข่งขันเป็นจำนวนมาก ในขณะที่ผู้คนก็ต้องการรู้ เหตุการณ์ต่างๆ ในการแข่งขันเป็นจำนวนมากเช่นกัน ในปัจจุบันการสรุปเนื้อหาโดยใช้คอมพิวเตอร์ยังไม่แพร่หลายนัก เนื่องจากขาดความแม่นยำและข่าวที่สรุปนั้นอ่านแล้วเข้าใจได้ยาก จึงนิยมสรุปโดยใช้แรงงานมนุษย์ อย่างไรก็ตามจำนวนการแข่งขันนั้นมีเป็นจำนวนมาก การใช้แรงงานมนุษย์จึงไม่เพียงพอต่อจำนวนนัดการแข่งขัน

ปัจจุบันมีคนวิจัยการสรุปเนื้อหาข่าวกีฬาฟุตบอลอยู่บ้าง แต่เห็นผลงานที่เป็นรูปธรรมเช่น ซอฟต์แวร์หรือโปรแกรมประยุกต์ ออกมาเป็นจำนวนน้อย ในการศึกษาพบว่ามีงานวิจัยจำนวนน้อยที่นำหลักของตัวแบบทางคณิตศาสตร์ (model) เข้ามาประยุกต์ใช้ ดังนั้นทางคณะผู้จัดทำจึงได้สนใจนำหลักของตัวแบบทางคณิตศาสตร์เข้ามาประยุกต์ใช้เพื่อให้ข้อความข่าวที่สรุปมีความถูกต้องแม่นยำและรวดเร็วมากขึ้น

ดังนั้นคณะผู้จัดทำจึงมีความสนใจที่จะพัฒนาโปรแกรมสรุปเนื้อหาข่าวกีฬาฟุตบอลที่เป็นภาษาอังกฤษ โดยนำหลักการเรียนรู้ด้วยเครื่อง (machine learning) มาประยุกต์ใช้เพื่อสร้างตัวแบบคณิตศาสตร์ในการสรุปข่าว เพื่อให้ผู้ใช้งานได้เสพข่าวที่มีความรวดเร็วขึ้นและทำให้ช่วยลดภาระในการสรุปข่าว

ความรู้ที่เกี่ยวข้อง

วิธีการเก็บรวบรวมข้อมูล

ในการศึกษาครั้งนี้คณะผู้จัดทำได้ใช้ข้อมูลจาก https://www.sportsmole.co.uk ซึ่งได้ใช้ Beautiful Soup 3 ในการเก็บรวบรวมข้อมูล

Beautiful Soup 3

Beautiful Soup 3 คือ คลังโปรแกรม (library) หนึ่งในภาษา python3 ที่มีไว้สำหรับเก็บรวบรวมข้อมูลจากหน้าเว็บ หรือ HTML

รูปแบบการแทนข้อความ (Text representation)

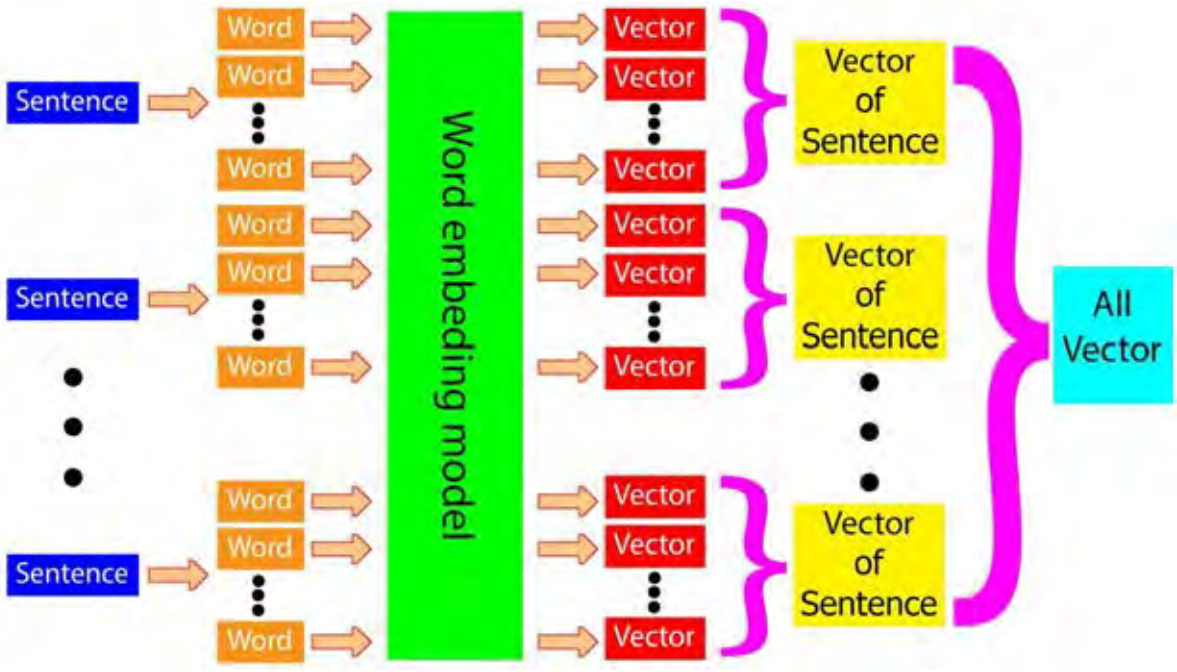
เนื่องจากการนำข้อความมาใช้ในการเรียนรู้ด้วยเครื่องโดยตรงนั้นทำได้ยาก คณะผู้จัดทำจึงได้เลือกใช้รูปแบบการแทนข้อความต่างๆดังต่อไปนี้

One-hot

One-hot คือชุดของ bits (1 หรือ 0) ของข้อมูลประเภทต่าง ๆ โดย bit ที่แทนคำนั้นจะเป็น 1 แค่ตัวเดียวและที่เหลือจะเป็น 0 การใช้การเข้ารหัสแบบ one-hot จะไม่สนลำดับคำ หรือคำอื่นๆ ที่อยู่ลำดับติดกัน จะได้ข้อมูลแค่ว่ามีคำนั้นๆอยู่หรือไม่เท่านั้น

Word embedding

word embedding คือการแปลงคำเป็นเวกเตอร์ ถือเป็นหนึ่งในวิธีในการสร้างตัววัด (feature) จากคำวิธีหนึ่ง โดยจะทำการลดขนาดของปริภูมิเวกเตอร์ (vector space) ลงด้วย ในการศึกษานี้คณะผู้จัดทำได้ทำการแบ่งข้อมูลเข้า (input) ซึ่งก็คือเนื้อหาสคริปต์ข่าวเป็นประโยค และทำการแบ่งประโยคออกเป็นคำเดี่ยวๆ เพื่อนำไปป้อนให้กับ word embedding ผลลัพธ์ที่ได้จะเป็นเวกเตอร์ของคำ จากนั้นจะนำมารวมกันเป็นเวคเตอร์ของประโยค (vector of sentence) และเวคเตอร์รวม (All vector) เพื่อนำไปใช้กับตัวแบบทางคณิตศาสตร์ที่มีชื่อว่า sequence-to-sequence ต่อไป



แผนภาพแสดงหลักการทำงานโดยง่ายของ word embedding

GloVe (Global Vectors for text representation)

GloVe คือ อัลกอริทึมแบบ unsupervised learning สำหรับการได้รับเวกเตอร์ของรูปแบบการแทนสำหรับคำ (representatons for words) [1] ในการศึกษานี้ได้นำ GloVe มาใช้เป็นคลังคำศัพท์ (corpus) เพื่อใช้ในการอ้างอิงเวกเตอร์ในการทำ word embedding ซึ่งคณะผู้จัดทำได้ใช้ตัวแบบคณิตศาสตร์ที่ได้เรียนรู้ด้วย GloVe แล้วมาใช้งาน

ตัวแบบคณิตศาสตร์

ในการศึกษาครั้งนี้คณะผู้จัดทำได้ตัดสินใจใช้ตัวแบบคณิตศาสตร์ที่มีชื่อว่า sequence-to-sequence เท่านั้น

sequence-to-sequence (Sequence to Sequence)

sequence-to-sequence เป็นตัวแบบคณิตศาสตร์ชนิดหนึ่งของ โครงข่ายประสาทเทียมแบบวนซ้ำ (rnn: recurrent neural networks) โดยหลักการโดยย่อของ sequence-to-sequence คือการป้อนข้อมูลเข้าเป็นลำดับ (sequence) และผลลัพธ์ที่ได้ก็เป็นลำดับ กล่าวคือไม่จำกัดขนาดของข้อมูลเข้าและผลลัพธ์

วิธีการวัดผล

ทางคณะผู้จัดทำได้ศึกษาวิธีการวัดผล 2 วิธีดังต่อไปนี้

BLEU score (BiLingual Evaluation Understudy score)

BLEU score เป็นอัลกอริธึมสำหรับการประเมินคุณภาพของข้อความที่ได้รับการแปลด้วยเครื่อง (machine translation) จากภาษาธรรมชาติ (natural language) หนึ่งไป

ยังอีกภาษาหนึ่ง ซึ่งคะแนนที่ได้จะมีค่าได้ตั้งแต่ 0 ถึง 1 (เป็น probability) คณะผู้จัดทำได้เลือกวิธีนี้ในการวัดผลเพราะเป็นที่นิยม และมีความแม่นยำมากกว่า

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score

ROUGE เป็นชุดของเมทริกซ์และแพ็คเกจซอฟต์แวร์สำหรับการวัดผลการสรุปอัตโนมัติ (automatic summarization) และซอฟต์แวร์การแปลด้วยเครื่อง ในการประมวลภาษาธรรมชาติ (NLP: Natural Language Processing)

วัตถุประสงค์

พัฒนาโปรแกรมสรุปเนื้อหาข่าวสารเกี่ยวกับกีฬาฟุตบอล โดยนำเทคนิคการเรียนรู้ด้วยเครื่อง มาประยุกต์ใช้

ขอบเขตของโครงการ

1. การศึกษานี้จะมุ่งศึกษาเกี่ยวกับการสรุปสคริปท์จากรายงานสดกีฬาฟุตบอล
2. การศึกษานี้จะใช้แหล่งที่มาของคลิปข่าวจากเว็บไซต์ https://www.sportsmole.co.uk เท่านั้น
3. การศึกษานี้จะใช้ข่าวที่เป็นภาษาอังกฤษเท่านั้น
4. ผลลัพธ์ที่ได้จะอยู่ในรูปแบบข้อความภาษาอังกฤษและไม่สนใจไวยากรณ์
5. ใช้ตัวแบบทางคณิตศาสตร์ที่มีชื่อว่า sequence-to-sequence เท่านั้น

วิธีการดำเนินงาน

1. ศึกษาเนื้อหาและบทความเกี่ยวกับการสรุปข้อความโดยใช้การเรียนรู้ด้วยเครื่อง ซึ่งใช้ RNN ประเภท sequence-to-sequence จากคลังโปรแกรมของ Google ที่มีชื่อว่า TensorFlow มาเป็นตัวต้นแบบในการพัฒนา โดยจะทำเป็น supervised learning โดยสร้างสรุปข่าวจากสคริปท์ข้อมูลที่เป็นข้อมูลขาเข้า
2. รวบรวมข้อมูลสคริปท์ข่าวและสรุปของข่าวเดียวกันโดยใช้คลังโปรแกรมที่มีชื่อว่า beautiful soup 3

6.45pm	Hello and welcome to Sports Mole's live coverage of the La Liga fixture between Valencia and
6.47pm	This match comes just over 24 hours after Real Madrid dropped more points in La Liga. Julen
6.49pm	As far as the league table is concerned, Barcelona have dropped down to second as a result

ตัวอย่างบางส่วนของข้อมูลสคริปท์ จะประกอบไปด้วยเวลาที่นักข่าวรายงานและเนื้อหาของสคริปท์

Barcelona have been held to a 1-1 draw by Valencia in Sunday's La Liga fixture at the Mestalla.

ตัวอย่างส่วนหนึ่งของข้อมูลสรุป

3. ศึกษาข้อมูล โปรแกรมและเทคนิคที่สามารถนำมาใช้ในการดำเนินงาน โดยสรุปจากเอกสารที่ใช้อ้างอิงได้ดังนี้ โมเดลที่ใช้คือ sequence to sequence model เป็น DNN (Deep Neural Network) ที่แบ่งเป็น 2 ส่วนคือ Encoder และ Decoder โดยส่วน encoder จะเข้ารหัส (encode) สคริปท์และส่วน decoder จะถอดรหัส (decode) ออกมาเป็นสรุป นอกจากนั้นในเอกสารอ้างอิงก็จะใช้ LSTM แทน standard RNN ในการสร้างส่วนเข้ารหัสและส่วนถอดรหัสอีกด้วยด้วย [2]
4. กำหนดขอบเขตและวิธีการดำเนินงาน
5. ออกแบบโมเดลและวิธีการสรุปสคริปท์ข่าวที่รวบรวมมา
6. พัฒนาโปรแกรมสรุปสคริปท์ข่าว
7. ทดสอบการใช้โปรแกรมที่พัฒนาแล้ว โดยใช้ BLEU คิดคะแนนจากสรุปที่ได้กับโปรแกรมเทียบกับสรุปจากการสุ่มคำ (random) และทำแบบสอบถามให้คนอ่านเลือกระหว่างสรุปจากโปรแกรมและจากการสุ่มคำ แล้วให้คะแนนตามความเข้าใจ 1 ถึง 5 (ไม่เข้าใจ ถึง เข้าใจมาก) [3]
8. ตรวจสอบความถูกต้องและแก้ไขความผิดพลาดของโปรแกรมที่พัฒนา
9. ทำการวัดผลเพื่อเปรียบเทียบการใช้งานระหว่างใช้งานโปรแกรมและไม่ได้ใช้งาน ในแง่ของเวลา และอ่านแล้วจับใจความได้
10. สรุปผลการดำเนินงาน ข้อเสนอแนะและการจัดทำเอกสาร

ตารางเวลาดำเนินงาน

ขั้นตอนการดำเนินงาน	ปี 2561					ปี 2562			
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มิ.ย.	เม.ย.
1. ศึกษาเนื้อหาและบทความเกี่ยวกับการสรุปข้อความโดยใช้ Machine Learning									
2. รวบรวมข้อมูลความต้องการเพื่อใช้ในการกำหนดขอบเขตการดำเนินงาน									
3. ศึกษาข้อมูล โปรแกรมและเทคนิคที่สามารถนำมาใช้ในการดำเนินงาน									
4. กำหนดขอบเขตและวิธีการดำเนินงาน									
5. ออกแบบโมเดลและวิธีการสรุปสคริปท์ข่าวที่รวบรวมมา									
6. พัฒนาโปรแกรมสรุปสคริปท์ข่าว									
7. ทดสอบการใช้โปรแกรมที่พัฒนาแล้ว									
8. ตรวจสอบความถูกต้องและแก้ไขความผิดพลาดของโปรแกรมที่พัฒนา									
9. ทำการวัดผลเพื่อเปรียบเทียบการใช้งานระหว่างใช้งานโปรแกรมและไม่ได้ใช้งาน									
10. สรุปผลการดำเนินงาน ข้อเสนอแนะและการจัดทำเอกสาร									

ประโยชน์ที่ได้รับ

1. ประโยชน์ต่อผู้พัฒนา
   1.1. ได้ศึกษาเรียนรู้เกี่ยวกับ Machine Learning
   1.2. ได้ศึกษาเรียนรู้เกี่ยวกับ การสรุปข่าวกีฬาฟุตบอล
2. ประโยชน์ต่อผู้ใช้ระบบ

   2.1. สามารถสรุปข่าวฟุตบอลได้อย่างรวดเร็วแม่นยำ

อุปกรณ์และเครื่องมือที่ใช้

1. ฮาร์ดแวร์ (Hardware)
   1.1. เครื่องคอมพิวเตอร์ 2 เครื่อง
      1.1.1.CPU Intel Core i7-7700 HQ 2.80 GHz RAM 4.00 GB HDD 1 TB ระบบปฏิบัติการ Window 10
      1.1.2.CPU Intel Core i7-4720 HQ 2.60 GHz RAM 16.00 GB HDD 1 TB ระบบปฏิบัติการ Window 10
2. ซอฟต์แวร์ (Software)
   2.1. Jupyter Notebook
   2.2. Google Chrome
   2.3. MS Office 365
3. ภาษาที่ใช้ (Programming Languages)
   3.1. Python 3

งบประมาณ

1.	Virtual Private Server (VPS)	ราคา	2,500	บาท
2.	Solid State Drive ความจุ 250 GB 1ชิ้น	ราคา	3,500	บาท
3.	RAM DDR4 ขนาด 8GB 1 ชิ้น	ราคา	2,500	บาท
4.	ค่าถ่ายเอกสาร และค่าทำรูปเล่ม	ราคา	250	บาท
5.	ค่าพิมพ์โปสเตอร์และค่าเดินทางไปนำเสนอผลงาน	ราคา	1,000	บาท
6.	กระดาษถ่ายเอกสาร A4 80 แกรม 1 รีม	ราคา	250	บาท
		รวม	10,000	บาท

เอกสารอ้างอิง

[1] Jeffrey Pennington. GloVe: Global Vectors for Word Representation. Available from https://nlp.stanford.edu/projects/glove/ [2014, August]

[2] Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." *arXiv preprint arXiv:1602.06023* (2016).

[3] Graham, Yvette. "Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE." *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015.

ประวัติผู้เขียน

Mr. Korawitch Kampanthong

นาย กรวิชญ์ กำปั่นทอง

ชั้นปีที่ 4 คณะวิทยาศาสตร์

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

สาขาคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย

เบอร์โทรศัพท์: 0950463344

อีเมล: korawitch_a-oak@hotmail.com

Mr. Apichai Somnam

นาย อภิชัย สมนาม

ชั้นปีที่ 4 คณะวิทยาศาสตร์

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

สาขาคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย

เบอร์โทรศัพท์: 086-7750255

อีเมล: ptaprchai@gmail.com