สาระสำคัญ

เนื่องจากในปัจจุบัน เอกสารความรู้ต่างๆ มากมาย ยังไม่ถูกนำมาจัดเก็บและจัดหมวดหมู่ให้เป็นระบบ ทำให้องค์ความรู้เหล่านั้นไม่สามารถนำมาใช้งานได้อย่างมีประสิทธิภาพ โดยในปัจจุบันนั้น การที่จะนำความรู้ที่มีอยู่เหล่านั้นมาจัดหมวดหมู่เพื่อใช้ในการสืบค้น จะต้องใช้มนุษย์เป็นผู้จัดการ ซึ่งจะเป็นการเปลืองทรัพยากรมนุษย์เป็นอย่างมาก โดยการพัฒนาการจัดการองค์ความรู้หรือ Knowledge management นั้น จะช่วยให้การนำความรู้ที่มีอยู่มาใช้ได้อย่างมีประสิทธิภาพมากขึ้น และช่วยส่งเสริมการพัฒนาประเทศชาติให้มุ่งไปสู่ความเป็นสังคมอุดมปัญญาได้ โดยที่ระบบ Knowledge management ที่มีอยู่ในปัจจุบันนั้น ยังไม่มีระบบที่สามารถนำเอาเอกสารไปทำการวิเคราะห์และ tag หมวดหมู่ของเนื้อหาได้โดยอัตโนมัติ โดยเฉพาะในส่วนของภาษาไทย ซึ่งยังไม่มีนักพัฒนารายใดพัฒนาเทคโนโลยีในลักษณะนี้ออกมา ดังนั้น เราจึงคิดที่จะทำระบบที่สามารถนำเอาเอกสารต่างๆ ที่มีอยู่ไปวิเคราะห์ข้อความและทำการสรุปว่า ข้อความนี้มีเนื้อหาที่เกี่ยวข้องกับเรื่องใดบ้าง และนำไปจัดเก็บลงไปยังฐานข้อมูล เพื่อให้สามารถทำการสืบค้นได้ โดยสิ่งที่ท้าทายกับการทำโครงงานชิ้นนี้ก็คือ การที่ภาษาไทยไม่มีรูปแบบประโยคที่ชัดเจน ทำให้การวิเคราะห์รูปประโยคมีความยาก และการทำ machine learning ให้ได้ความแม่นยำในระดับที่สามารถนำไปใช้ได้จริงนั้น จะต้องใช้การเลือกใช้ algorithm และการปรับแต่งที่เหมาะสมกับข้อมูลที่นำมาใช้ จึงทำให้โครงการนี้มีความท้าทายในการดำเนินการ และเป้าหมายในการทำโครงการนี้ จะเป็นการพัฒนา model ที่สามารถนำข้อความจากเอกสารมา tag และทำการจัดหมวดหมู่ได้ และพัฒนา web application ที่สามารถสืบค้นข้อมูลที่ได้จาก model ข้างต้น เพื่อนำมาเป็น Proof of Concept ของ model ที่พัฒนาขึ้น

หลักการและเหตุผล

ในยุคปัจจุบัน ที่มีการทำเอกสารในเรื่องต่างๆ ออกมาเป็นจำนวนมาก การทำ Knowledge management หรือการนำเอกสารข้อมูลเหล่านั้นมาจัดการให้เป็นระบบ นับเป็นเรื่องที่สำคัญมาก โดยเฉพาะในองค์กรหลายๆ แห่ง การมีระบบ knowledge management จะช่วยทำให้องค์กรนั้นๆ สามารถใช้งานองค์ความรู้ที่มีอยู่ได้อย่างมีประสิทธิภาพสูงสุด แต่ในปัจจุบัน เอกสารความรู้ต่างๆ ที่ถูกนำมาเผยแพร่อยู่นั้น มักจะอยู่ในรูปแบบของเอกสารในหน้ากระดาษ หรือเอกสารที่เป็นไฟล์ PDF ซึ่งยังไม่สามารถนำมาใช้งานได้อย่างมีประสิทธิภาพเท่าที่ควร เพราะว่า เอกสารเหล่านั้นมักจะมีข้อความอยู่มากมาย ที่เกี่ยวข้องกับเนื้อหาที่แตกต่างกัน แต่ว่าเมื่อผู้ที่ต้องการใช้งานความรู้เหล่านั้น ต้องการทำการหาเนื้อหาที่เฉพาะเจาะจงกับที่เขาสนใจในเอกสารนั้นๆ เขาก็ต้องทำการค้นหาด้วยตนเองโดยวิธีต่างๆ ไม่ว่าจะเป็นการไล่อ่านเนื้อหาทั้งหมดด้วยตนเอง ซึ่งใช้เวลามากในการอ่านและหาใจความสำคัญที่เขาต้องการ หรือใช้การค้นหา keyword ที่เขาต้องการด้วยวิธีต่างๆ เช่นการเปิดหาสารบัญ ซึ่งเอกสารบางฉบับก็ไม่มีสารบัญให้ หรือใช้การ search หา keyword ที่ต้องการ ซึ่งอาจจะเกิดการข้ามเนื้อหาในส่วนที่เกี่ยวข้องกับเรื่องที่ผู้ที่ค้นหาต้องการ แต่ไม่มี keyword ที่เขาใช้ค้นหาไปได้ ซึ่งสิ่งที่ได้กล่าวไปข้างต้นนั้น นับว่าเป็นปัญหาใหญ่ในการค้นคว้าหาข้อมูลเพื่อทำการศึกษาเป็นอย่างมาก เนื่องจากการที่ไม่มีระบบ knowledge management สำหรับเอกสารทั่วๆ ไปนั้น ทำให้แหล่งความรู้ที่สามารถนำมาสืบค้นได้นั้นลดลงเป็นอย่างมาก และทำให้ความรู้จำนวนมากถูกทิ้งร้างไว้ ไม่ได้ถูกนำมาใช้ให้เกิดประโยชน์ ดังนั้น ทางกลุ่มของเราจึงสนใจที่จะพัฒนา machine learning model ที่สามารถคัดแยกเนื้อหาในส่วนต่างๆ ในไฟล์เอกสาร และทำการ tag ข้อความเหล่านั้นได้โดยอัตโนมัติว่า เนื้อหาในส่วนนั้นๆ มีความเกี่ยวข้องกับเรื่องอะไรบ้าง และทำการจัดเก็บข้อมูลเหล่านั้นลงไปยังระบบฐานข้อมูล เพื่อให้สามารถทำการสืบค้นได้ง่ายและรวดเร็ว และทำให้การจัดการแหล่งความรู้ หรือ Knowledge management นั้น สามารถใช้งานกับเอกสารที่เป็นไฟล์ PDF ได้ ซึ่งส่งผลให้ความรู้ถูกนำไปใช้งานต่อ และเกิดการพัฒนาประเทศชาติในองค์รวมมากยิ่งขึ้น

วัตถุประสงค์

เพื่อสร้างเทคโนโลยีที่สามารถทำการวิเคราะห์ข้อความจากไฟล์ PDF และทำการจัดหมวดหมู่ของข้อความเหล่านั้นได้โดยอัตโนมัติ เพื่อให้เราสามารถนำเอกสารความรู้ต่างๆ ที่มีอยู่มากมายมาทำการจัดแบ่งกลุ่มและเก็บไว้ในฐานข้อมูลเพื่อใช้ในการสืบค้นได้ ซึ่งเทคโนโลยีนี้สามารถนำไปใช้งานในวงการการศึกษาได้เป็นอย่างดี โดยทำให้ผู้คนสามารถค้นหาความรู้จากแหล่งความรู้ที่กว้างขวางมากขึ้นโดยใช้เวลาที่น้อยลง และนำไปสู่การสร้างสังคมอุดมปัญญาต่อไปในอนาคต

ปัญหาหรือประโยชน์ที่เป็นเหตุผลให้ควรพัฒนาโปรแกรม

ในปัจจุบันนั้น มีโปรแกรมที่ถูกพัฒนาโดยมหาวิทยาลัยสแตนฟอร์ด ที่ชื่อว่า Stanford Deepdive ที่ทำการวิเคราะห์ข้อความและ tag เนื้อหาเหล่านั้นได้ แต่ว่าโปรแกรม Stanford Deepdive นั้น ถูกพัฒนาขึ้นสำหรับการวิเคราะห์เอกสารในภาษาอังกฤษเป็นหลัก ซึ่งภาษาไทยที่มีรูปแบบของประโยค การจัดเรียงคำ การวางตำแหน่งคำ,ตัวอักษร และอื่นๆ ที่แตกต่างจากภาษาอังกฤษเป็นอย่างมาก ทำให้การทำระบบ tag เอกสารอัตโนมัติสำหรับภาษาไทยนั้น ไม่สามารถใช้ Stanford Deepdive ได้ และการ tag หมวดหมู่ให้กับข้อความจำนวนมากโดยใช้มนุษย์ในการจัดการนั้น จะเป็นการเสียเวลาและทรัพยากรมนุษย์อันมีค่าไปเป็นจำนวนมาก ทำให้ทางกลุ่มของเราสนใจที่จะพัฒนาโปรแกรมในลักษณะคล้ายๆ กันกับ Stanford Deepdive ที่สามารถนำมาใช้กับภาษาไทยได้ เพื่อให้เอกสารต่างๆ ที่เป็นภาษาไทยนั้น ถูกนำมาใช้ประโยชน์ และนำมาศึกษาต่อได้อย่างมีประสิทธิภาพ

เป้าหมายและขอบเขตของโครงการ

ทำการพัฒนา machine learning model ที่สามารถนำมาใช้อ่านข้อความจากไฟล์ PDF ที่เป็นภาษาไทย และทำการ tag หมวดหมู่ให้กับแต่ละข้อความเหล่านั้น และทำการพัฒนา web application เพื่อที่จะใช้ในการสืบค้นข้อมูลที่ทำการ tag มาแล้วได้ โดยมีขอบเขตในการทำคือ เนื้อหาที่สามารถนำมาให้ตัว model ทำการจัดหมวดหมู่ได้นั้น ต้องเป็นเนื้อหาที่เกี่ยวข้องกับเรื่องที่ตัว model เคยเรียนรู้มาแล้ว