**ข้อเสนอโครงการ**

**โครงการระบบจัดหมวดหมู่องค์ความรู้แบบอัตโนมัติ**

**เพื่อพัฒนาการเรียนการสอน**

**1 สาระสำคัญ**

การศึกษา ถือเป็นรากฐานที่สำคัญในการพัฒนาทรัพยากรมนุษย์ เปรียบเสมือนเครื่องมือหลักในการช่วยพัฒนาตั้งแต่การวางรากฐาน ศักยภาพ และขีดความสามารถ จนก่อให้เกิดเป็นพลังสร้างสรรค์ในการพัฒนาประเทศอย่างยั่งยืน ในประเทศไทยรัฐบาลมีการให้ทุนสนับสนุนเพื่อการศึกษาค้นคว้า ทดลองหาวิธีการต่างๆ ที่จะช่วยเพิ่มประสิทธิภาพทั้งการสอนของคุณครู และการเรียนของเด็กนักเรียนให้สูงขึ้น โดยผลที่ได้จากการวิจัยหรือทดลองนี้ จะถูกเขียนออกมาเป็นรูปเล่มรายงานและเก็บในรูปแบบของไฟล์ PDF ที่ประกอบไปด้วย รายงานสำหรับผู้บริหารที่แสดงถึงตัวเลขสถิติต่างๆของการทดลอง และรายงานสำหรับการเรียนการสอน ที่อธิบายรายละเอียดแนวทางการเรียนการสอนเช่น เทคนิคการสอน การดึงความสนใจของนักเรียนในเรื่องต่างๆ หรือการนำสื่อการเรียนการสอนมาใช้ให้เป็นประโยชน์ เป็นต้น อย่างไรก็ตาม เอกสารรายงานเหล่านี้ไม่ได้มีการเผยแพร่ในวงกว้าง และถูกเก็บในรูปแบบไฟล์ PDF ที่ไม่สามารถสืบค้นได้ จึงไม่สามารถถูกนำมาใช้ประโยชน์ได้อย่างเต็มที่ การสร้าง knowledge sharing platform สำหรับคุณครู ผู้ปกครอง และผู้บริหาร ที่สามารถสืบค้นหาข้อมูลได้อย่างสะดวกจึงเป็นเรื่องจำเป็นเพื่อสร้างให้เกิดความเท่าเทียมของโรงเรียนในระดับประถมศึกษาและมัธยมศึกษาในพื้นที่ต่างๆ ดังนั้นข้อเสนอโครงการฉบับนี้จึงเสนอแนวคิดในการใช้เทคโนโลยีเพื่อรวบรวม คัดกรอง จัดหมวดหมู่ รวมถึงสกัดเนื้อหาส่วนที่สำคัญจากเอกสารอิเล็กทรอนิกส์ เพื่อให้เอกสารการทดลองด้านนวัตกรรมการเรียนการสอนสามารถถูกนำไปใช้ประโยชน์ และสร้างให้เกิดสังคมของการการแลกเปลี่ยนเรียนรู้ของคุณครูและผู้ปกครองได้

ทั้งนี้ ข้อมูลที่อยู่ในไฟล์ PDF หรือไฟล์ text นั้นเป็นข้อมูลขนาดใหญ่ที่ไม่มีโครงสร้าง (Schemaless) ไม่สามารถนำมาประมวลผลเพื่อวิเคราะห์ด้วยซอฟท์แวร์ทั่วๆไปได้ จึงจำเป็นต้องมีการพัฒนาระบบจัดหมวดหมู่องค์ความรู้แบบอัตโนมัติขึ้น โดนอาศัยเทคโนโลยีทางด้านการประมวลผลข้อมูล text หรือ Text Mining และเทคโนโลยีทางด้าน Machine Learning เข้ามาช่วยเพื่อให้สามารถใช้งานกับรายงานจำนวนมากจากทั่วประเทศได้อย่างมีประสทธิผล

ในการพัฒนาระบบในระยะเริ่มต้นนั้น จะมีคุณครูอาสาสมัครเข้ามาช่วยอ่านและทำการระบุข้อความส่วนที่เป็นเนื้อหาใจความสำคัญของรายงานนั้นๆ และสร้าง Tag เพื่อบ่งบอกหัวเรื่องของเนื้อหา เพื่อใช้สำหรับการจัดหมวดหมู่ โดยคุณครูจะช่วยวิเคราะห์รายงานเพียงส่วนน้อยเท่านั้น หลังจากนั้นเนื้อหาและ Tag ที่คุณครูสร้างขึ้นจะถูกนำมาพัฒนา Machine Learning Model โดย Model จะถูกสอนให้เรียนรู้คำต่างๆที่เกี่ยวข้องกับ tag ที่อยู่ในรายงาน เมื่อมีรายงานเล่มใหม่เข้ามาในระบบ ระบบจะทำการวิเคราะห์เนื้อหาและสามารถแสดงส่วนที่เป็นใจความสำคัญ รวมถึงจัดประเภทหมวดหมู่ของรายงานได้โดยอัตโนมัติ ซึ่งวิธีนี้ช่วยให้ประหยัดทั้งเวลาและจำนวนทรัพยากรบุคคลเป็นอย่างมาก เมื่อผู้ใช้งานเข้ามาใช้ระบบนี้ จะสามารถสืบหาข้อมูลที่เกี่ยวข้องกับการพัฒนาการเรียนการสอน ด้วยการใส่ Keyword จากนั้น web application จะให้ผลลัพธ์ออกมาเป็นข้อความที่เกี่ยวข้องพร้อมทั้งแนบลิงค์สำหรับดาวน์โหลดเอกสาร ผู้ใช้งานสามารถอ่านสรุปใจความสำคัญที่ระบบแสดงก่อน และหากตรงกับความสนใจสามารถดาวน์โหลดรายงานทั้งเล่มไปเพื่อศึกษารายละเอียดต่อไป

ผู้เสนอโครงการหวังว่าระบบดังกล่าวจะเป็นอีกหนึ่งแหล่งรวบรวมข้อมูลองค์ความรู้ที่จะช่วยให้คุณครูสามารถศึกษาค้นคว้าเพื่อพัฒนาการเรียนการสอนและช่วยผลักดันให้การศึกษาในประเทศไทยสามารถพัฒนาก้าวหน้าไปได้อีกขั้น

**2. หลักการและเหตุผล**

จากสถิติการศึกษาขั้นพื้นฐานของประเทศไทยในรายงานประจำปีของ World Economic Forum ปี 2014-2015 [1]พบว่าประเทศไทยอยู่ในลำดับที่ 90 จาก 144 ประเทศทั่วโลกที่ได้รับการจัดอันดับ ซึ่งถือได้ว่าอยู่ในลำดับค่อนข้างต่ำ ในขณะเดียวกันผลการวิเคราะห์ในรายงานของ International Institute of Management Development และ Pearson-The Economist Intelligence Unit พบว่าการศึกษาของไทยถูกจัดให้อยู่ในกลุ่มต่ำสุดเช่นเดียวกัน ซึ่งรายงานเหล่านี้ล้วนเป็นตัวบ่งชี้ให้เห็นว่าการศึกษาของไทยยังมีจุดบกพร่องอีกมาก ควรที่จะต้องได้รับการพัฒนาอย่างเร่งด่วน

หนึ่งในปัจจัยสำคัญที่มีผลต่อการพัฒนาการศึกษา คือการสอนของคุณครู เพราะคุณครูเปรียบเสมือนผู้ที่ถ่ายทอดความรู้ต่างๆและพัฒนาเด็กให้เติบโตไปเป็นทรัพยากรมนุษย์ที่มีคุณภาพ ดังนั้น รัฐบาลไทยจึงมีการให้ทุนสนับสนุนกับครูในการศึกษาค้นคว้า ทดลองหาวิธี ในการพัฒนาการเรียนการสอนให้มีประสิทธิภาพและเข้าถึงเด็กนักเรียนได้มากขึ้น โดยเฉพาะการศึกษาขั้นพื้นฐานในระดับประถามและมัธยมศึกษา ซึ่งในปัจจุบัน ได้มีเอกสารที่ถูกเขียนออกมาเพื่อรายงานผลการทดลอง และวิธีในการพัฒนาการเรียนการสอนที่ดี (best practice) ซึ่งรายงานเหล่านี้มักจะหนาและอยู่ในรูปแบบของไฟล์ PDF ทำให้ครูสืบค้นข้อมูลได้ยาก และต้องเสียเวลาในการอ่านหนังสือหลายร้อยหน้าจำนวนหลายเล่มเพราะ เอกสารไม่มีการรวบรวมและจัดเป็นหมวดหมู่ ทำให้ประสิทธิภาพในการสืบค้นข้อมูลนั้นไม่ดี อีกทั้งยังอาจได้ข้อมูลที่ไม่ครบถ้วน

ทางผู้จัดทำจึงจะทำการรวบรวมเอกสารรายงานเหล่านี้ เพื่อทำให้เกิดเป็น knowledge sharing platform ที่คุณครูสามารถเข้ามาสืบค้นหามูล และศึกษาค้นคว้าได้อย่างง่าย ระบบดังกล่าวจะช่วยรวบรวมและทำการจัดหมวดหมู่เอกสาร รวมทั้ง ทำการวิเคราะห์ คัดแยกเนื้อหาส่วนต่างๆ ในไฟล์เอกสาร และทำการ tag ข้อความสำคัญให้โดยอัตโนมัติ ว่าเนื้อหาในแต่ละส่วนมีความเกี่ยวข้องกับเรื่องอะไรบ้าง และทำการจัดเก็บข้อมูลเหล่าลงไปยังระบบฐานข้อมูล ความรู้จากเอกสารเหล่านี้จะได้ถูกนำไปพัฒนาการเรียนการสอน และพัฒนาให้การศึกษาของไทยก้าวไปสู่ในระดับต้นๆของโลกได้ในอนาคต

**3. วัตถุประสงค์**

* เพื่อสร้างฐานข้อมูลที่ก่อให้เกิดเป็น knowledge sharing platform ของประเทศไทย
* เพื่อให้คุณครู ผู้บริหาร และผู้ปกครองสามารถสืบค้นข้อมูลได้อย่างมีประสิทธิภาพและรวดเร็ว และ สามารถนำองค์ความรู้ไปพัฒนาการเรียนการสอนได้
* เพื่อศึกษาการทำ Document และ Text Clustering สำหรับการคัดแยกเนื้อหาและจัดหมวดหมู่เนื้อหาภายในเอกสาร
* เพื่อสร้าง Machine Learning Model สำหรับเรียนรู้เอกสารภาษาไทยทางด้านการศึกษา
* เพื่อสร้าง Web Application สำหรับการจัดการองค์ความรู้

**4. เป้าหมายและขอบเขตของโครงการ**

เป้าหมายของโครงการโครงการสร้างระบบจัดหมวดหมู่องค์ความรู้แบบอัตโนมัติด้วยเทคโนโลยี Text Mining และ Machine Learning คือการสร้าง knowledge sharing platform ที่จะช่วยรวบรวมข้อมูลรายงานตัวอย่างการเรียนการสอนที่ดี (best practice) มาทำการคัดแยก จัดหมวดหมู่ เพื่อให้คุณครูสามารถสืบค้นข้อมูลได้อย่างมีประสิทธิภาพและรวดเร็ว สามารถนำความรู้ไปพัฒนาและปรับใช้ตามเพื่อให้เกิดการเรียนการสอนที่ดีที่จะช่วยพัฒนาศักยภาพของเด็กนักเรียนได้ โดย platform มีขอบเขต ดังนี้

* 1. สร้างเครื่องมือสำหรับรับไฟล์ PDF/text และดึงข้อความภาษาไทยในแต่ละย่อหน้าออกมา โดยข้อความจะถูกประมวลผล Text และใส่เข้าใน Machine Learning Model เพื่อติด Tag ที่เกี่ยวข้องกับเนื้อหาต่อไป
  2. สร้าง Machine learning model ที่สามารถรับข้อมูลจากไฟล์ text ภาษาไทยที่ได้จากขั้นตอนที่ 4.1 และทำการติด tag สำหรับแต่ละย่อหน้า โดยวิธีการ supervised classification ซึ่งจะแบ่งเป็น 2 ขั้นตอนคือ
  + เครื่องมือสำหรับการ train model โดยจะรับ text และ tag ของย่อหน้าต่างจากผู้เชี่ยวชาญและนำมาปรับจนได้โมเดลทางคณิตศาสตร์ที่เหมาะสม
  + เครื่องมือสำหรับการทำนายหรือการติด tag อัตโนมัติด้วยโมเดลทางคณิตศาสตร์ที่พัฒนาขึ้นจะรับข้อความภาษาไทยเป็นอินพุธและทำการติด tag ให้ย่อหน้าต่างๆที่อธิบายประเด็นสำคัญของหนังสือ/รายงานแต่ละเล่ม ผลลัพธ์จะถูกเก็บลงในฐานข้อมูลเพื่อการสืบค้นต่อไป
  1. สร้าง Web application โดยมีคุณสมบัติหลักคือ
  + สามารถรับไฟล์รายงานในรูปแบบไฟล์ text หรือ pdf เพิ่มเติมจากผู้ใช้ และนำไปประมวลผลโดยใช้ Machine Learning Model ในหัวข้อที่ 4.2 ซึ่งจะทำการติด tag โดยอัตโนมัติและเก็บลงฐานข้อมูล
  + สามารถสืบค้นข้อมูลตัวอย่างการเรียนการสอนที่ดี โดยผ่านการพิมพ์ข้อความหรือคีย์เวิร์ดลงในช่อง Search หรือผ่านการคลิกเลือกจากเมนู Advance Filter หรือจากการคลิกบนหน้าจอแผนที่ได้
  + สามารถแสดงผลการสืบค้นในหลายระดับ คือประเภทเอกสาร หมวดหมู่ของ tag และตัวอย่างข้อความที่สำคัญ โดยการแสดงผลสามารถแสดงในรูปแบบแผนที่หรือ Formatted Text ได้ นอกจากนี้ผู้ใช้ระบบจะสามารถดาวน์โหลดไฟล์ต้นฉบับออกจากระบบได้หากต้องการ
  + ผู้ใช้ระบบสามารถให้ review และ กด Like ข้อความและแชร์บทความที่น่าสนใจบน Facebook ได้ ซึ่งระบบจะนำข้อมูล Review/Like/Share ไปใช้ในการจัดอันดับข้อความและรายงานต่อไป

\*รายละเอียดเชิงเทคนิคดูได้จากเอกสารแนบ 2

**5. ข้อจำกัดของโปรแกรมที่พัฒนา**

* เนื่องจากภาษาไทยเป็นภาษาที่มีความซับซ้อนสูง ทั้งทางด้านตัวอักษร ที่มีสระบน-ล่าง และทางด้านรูปประโยคที่ไม่มีความแน่นอน ทำให้การเขียนโปรแกรมที่สามารถประมวลผลภาษาไทยได้อย่างสมบูรณ์แบบจึงเป็นเรื่องยาก ดังนั้นความแม่นยำในการ tag อาจจะต่ำกว่าการใช้งานกับภาษาอังกฤษที่มีรูปประโยคและการตัดคำที่แน่นอนกว่า
* ข้อมูลที่จะนำไปเข้าระบบ Machine Learning เพื่อให้ระบบทำการเรียนรู้ด้วยตนเองนั้น จะต้องใช้มนุษย์เป็นตัวช่วยในการกำหนดข้อมูลก่อนในเบื้องต้น (Supervised learning) เพราะฉะนั้น ถ้าเราต้องการให้ระบบเรียนรู้เนื้อหาเรื่องใหม่ๆ จะต้องมีการใช้ผู้เชี่ยวชาญที่เกี่ยวข้องกับเรื่องที่จะให้ระบบเรียนรู้มาช่วยทำการ Tag ย่อหน้าก่อนที่จะนำข้อมูลเข้าไปในระบบ ความเข้าใจและการทำงานของผู้เชี่ยวชาญในการ Train Machine จึงมีความสำคัญเป็นอย่างยิ่ง
* การระบุย่อหน้าจาก PDF นั้นสามารถทำได้ยาก เนื่องจากการระบุย่อหน้าจาก PDF จำเป็นต้องใช้ค่าตำแหน่งของตัวอักษรต่างๆ เพื่อระบุว่าย่อหน้าควรจะอยู่ตำแหน่งไหน ซึ่ง PDF ที่ได้รับมานั้น มีรูปแบบการจัดหน้าและ font ที่แตกต่างกันรวมถึงรูปแบบคำภาษาไทยและภาษาอังกฤษในเอกสาร จะทำให้ตำแหน่งของคำเกิดการคลาดเคลื่อนซึ่งจะส่งผลให้ย่อหน้าที่ได้ออกมาอาจเกิดความผิดพลาดได้

**6. วิธีดำเนินงาน**

|  |  |  |
| --- | --- | --- |
| ขั้นตอนการดำเนินงาน | กรอบเวลา  (สัปดาห์ที่) | สิ่งที่ส่งมอบ |
| 1. มจธ. ศึกษาเทคนิค อัลกอริธึ่ม และเทคโนโลยีที่เกี่ยวข้องเพื่อเตรียมความพร้อมในการพัฒนาระบบ | 1-3 |  |
| 2. มสส. เลือกเอกสารที่เหมาะสมที่จัดเป็นตัวอย่างที่ดีทางด้านการศึกษา และเลือกผู้เชี่ยวชาญทางด้านการศึกษา | 1-3 |  |
| 3. มจธ. และ มสส.ร่วมจัด workshop เพื่อทำความเข้าใจกับผู้เชี่ยวชาญในการระบุคำสำคัญในเอกสารและวิธีการติด Tag ให้กับข้อความ เพื่อใช้ในการทำ Supervised Learning | 3 |  |
| 4. มจธ. ออกแบบและพัฒนาเครื่องมือในการสะกัดข้อความจากไฟล์ pdf | 1-3 | ระบบสำหรับดึงข้อความออกจากไฟล์ pdf |
| 5. กลุ่มผู้เชี่ยวชาญโดยการกำกับของ มสส. สะกัดใจตวามสำคัญจากรายงานและติด Tag (เอกสาร ประมาณ 100 เล่ม) | 4-7 |  |
| 6. มจธ. สร้าง Machine Learning Model สำหรับสอนให้ระบบซอฟท์แวร์เข้าใจกระบวนการคิดของผู้เชี่ยวชาญในการเลือกข้อความที่เป็นประเด็นสำคัญและการติด Tag ที่เหมาะสม | 7-14 |  |
| 7. มจธ. และ มสส. ร่วมกัน ทดสอบ Machine Learning Model และปรับให้เหมาะสม | 13-16 | รายงานอธิบายโมเดลคณิตศาสตร์ที่พัฒนาขึ้นและวิธีการใช้งาน |
| 8. มจธ. ออกแบบหน้าจอทั้งหมดของ Web Application | 16-17 |  |
| 9. มจธ. พัฒนา Web Application ตามคุณสมบัติที่อธิบายในหัวข้อที่ 4.3 | 17-20 |  |
| 10. มจธ. และ มสส. ทดสอบระบบ และปรับแก้ตามความเหมาะสม | 18-22 | ระบบ Web Application ตามที่ระบุในหัวข้อที่ 4.3  รายงานแสดงวิธีการใช้งานระบบ web Application |
| 11. มสส. จัดประชุมผู้เชี่ยวชาญและกลุ่มผู้สนใจเพื่อแนะนำระบบ | 22 |  |
| 12. มจธ. เขียนรายงานฉบับสมบูรณ์ | 23-24 | รายงานฉบับสมบูรณ์ |

หมายเหตุ เพื่อให้การดำเนินการเป็นไปอย่างราบรื่นทีม มจธ. และ มสส. จะมีการประชุมร่วมกันทุกเดือน