**2.1 สาระสำคัญ**

การศึกษา ถือเป็นรากฐานที่สำคัญในการพัฒนาทรัพยากรมนุษย์ เปรียบเสมือนเครื่องมือหลักในการช่วยพัฒนาตั้งแต่การวางรากฐาน ศักยภาพ และขีดความสามารถ จนก่อให้เกิดเป็นพลังสร้างสรรค์ในการพัฒนาประเทศอย่างยั่งยืน ดังนั้น การศึกษาจึงเป็นอีกหนึ่งปัจจัยสำคัญที่มีความจําเป็นต่อการพัฒนาประเทศ ซึ่งปัจจุบันการศึกษาถูกใช้เป็นนโยบายหลักสำหรับพัฒนาประเทศในหลายประเทศ

ในประเทศไทย การพัฒนาการศึกษาในระดับประถมศึกษาและมัธยมศึกษา ถือเป็นหนึ่งปัญหาสำคัญที่เกิดขึ้น และรอแนวทางที่จะเข้ามาช่วยพัฒนาและแก้ไข ซึ่งในปัจจุบันรัฐบาลได้มีการให้ทุนสนับสนุนสำหรับการศึกษาค้นคว้า ทดลองหาวิธีการต่างๆ ที่จะช่วยเพิ่มประสิทธิภาพทั้งการสอนของคุณครู และการเรียนของเด็กนักเรียนให้สูงขึ้น โดยผลที่ได้จากการวิจัยหรือทดลองนี้ จะถูกเขียนออกมาเป็นรูปเล่มรายงานอย่างหนา ในรูปแบบของไฟล์ PDF ที่ประกอบไปด้วย รายงานสำหรับผู้บริหารที่แสดงถึงตัวเลขสถิติต่างๆของการทดลองนั้น และรายงานสำหรับการเรียนการสอน ที่อธิบายรายละเอียดแนวทางการเรียนการสอนในการทดลองนั้นๆ เช่น เทคนิคการสอน การดึงความสนใจของนักเรียนในเรื่องต่างๆ หรือการนำสื่อการเรียนการสอนมาใช้ให้เป็นประโยชน์ เป็นต้น อย่างไรก็ตาม เอกสารรายงานเหล่านี้ไม่ได้มีการเผยแพร่ และกระจัดกระจายอยู่ตามโรงเรียนหรือหน่วยงานต่างๆ ส่งผลให้การใช้ประโยชน์จากเอกสารกเหล่านั้นเป็นไปได้ยากมาก เนื่องจากในปัจจุบันยังไม่มี knowledge sharing platform สำหรับคุณครู ที่จะสามารถเข้าไปสืบค้นหาข้อมูลได้ อีกทั้งจำนวนหน้าของรายงานที่ค่อนข้างมาก ทำให้การกรองข้อมูลต่างๆเพื่อที่จะศึกษาเรื่องใดเรื่องหนึ่งนั้นทำได้ช้ามาก

ทั้งนี้ ทางผู้จัดทำได้เล็งเห็นถึงความสำคัญในการรวบรวมและคัดกรองข้อมูล ซึ่งการที่จะทำให้ข้อมูลเหล่านี้สามารถนำไปใช้ให้เกิดประโยชน์สูงสุดในเชิงของ knowledge sharing ได้นั้น จะต้องอาศัยการสรุปใจความในส่วนที่สำคัญของรายงานแต่ละเล่ม เพื่อให้คุณครูสามารถสืบค้น เรียนรู้จากตัวอย่างการสอนที่ดี (best practice) และสามารถใช้ประโยชน์จากข้อมูลในรายงานเหล่านั้นได้ในเวลาอันรวดเร็ว ดังนั้น จึงได้ออกแบบระบบ Building a knowledge sharing platform with Text mining and knowledge extraction ขึ้น ภายใต้ความร่วมมือกับมูลนิธิ… เพื่อรวบรวมรายงานเคล็ดลับตัวอย่างวิธีการสอนที่ดีเหล่านี้ รวมถึงวิเคราะห์ข้อมูลในรายงาน สรุปใจความสำคัญ และแบ่งออกเป็นหมวดหมู่ เพื่อให้ง่ายต่อการสืบค้นและศึกษาเรียนรู้ อย่างไรก็ตาม ข้อมูลที่อยู่ในไฟล์ PDF หรือไฟล์ text นั้นเป็นข้อมูลขนาดใหญ่ที่ไม่มีโครงสร้าง (Schemaless) จึงไม่สามารถนำมาประมวลผลเพื่อวิเคราะห์ด้วยซอฟท์แวร์ทั่วๆไปได้ ดังนั้น เพื่อให้ข้อมูลเหล่านี้สามารถนำมาใช้ให้เกิดประโยชน์ได้ จำเป็นต้องมีการพัฒนาระบบบริหารจัดการข้อมูลชนิดใหม่ที่สามารถดึงองค์ความรู้ออกจากข้อมูล text และนำไปบูรณาการกับข้อมูลอื่นๆ เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์เชิงสถิติ และการวิเคราะห์เชิงทำนาย (Predictive Analytics) ได้ โดยการประมวลผลข้อมูล text หรือ Text Mining นี้ ต้องอาศัยเทคโนโลยีทางด้าน Machine Learning มาช่วยในการทำงาน

ในการพัฒนาระบบในระยะเริ่มต้นนั้น จะมีคุณครูอาสาสมัครจากมูลนิธิ เข้ามาช่วยอ่านและทำการเน้นข้อความ หรือพารากราฟ ในส่วนที่เป็นเนื้อหาใจความสำคัญของรายงานนั้นๆ อีกทั้งสร้าง tag เพื่อบ่งบอกหัวเรื่องของเนื้อหานั้นด้วย เพื่อใช้สำหรับการจัดหมวดหมู่ เช่น สอนยังไงให้เด็กเข้าใจง่ายที่สุด เทคนิคการดึงดูดความสนใจของนักเรียน เป็นต้น โดยคุณครูจะช่วยทำการวิเคราะห์รายงานนี้ทำแค่บางส่วนเท่านั้น หลังจากนั้น ข้อมูลที่คุณครูวิเคราะห์เหล่านี้ จะถูกนำมาพัฒนา Machine Learning Model โดย model จะถูกสอนให้เรียนรู้คำต่างๆ (train) และ tag ของพารากราฟนั้นๆ ที่อยู่ในรายงาน เมื่อมีรายงานเล่มใหม่เข้ามาในระบบ ระบบจะทำการวิเคราะห์เนื้อหา (Topic discovery) และสามารถแสดงส่วนที่เป็นใจความสำคัญ รวมถึงจัดประเภทหมวดหมู่ (Document classification) ของรายงานได้อัตโนมัติ โดยไม่ต้องอาศัยคุณครูมานั่งอ่านหมดทั้งเล่ม ซึ่งวิธีนี้ช่วยให้ประหยัดทั้งเวลาและจำนวนทรัพยากรบุคคลเป็นอย่างมาก เมื่อผู้ใช้งานเข้ามาใช้ระบบนี้ จะสามารถสืบหาข้อมูลที่เกี่ยวข้องกับการพัฒนาการเรียนการสอน ด้วยการใส่คำ (tag) ที่ต้องการค้นหา จากนั้น web application จะให้ผลลัพธ์ออกมาเป็นย่อหน้าพารากราฟที่เกี่ยวข้องกับ tag ที่ผู้ใช้งานใส่ พร้อมทั้งแนบลิงค์สำหรับดาวน์โหลดเอกสาร ผู้ใช้งานสามารถอ่านพารากราฟใจความสำคัญที่ระบบแสดงก่อนได้ หากตรงกับความสนใจสามารถดาวน์โหลดรายงานทั้งเล่มไปเพื่อศึกษารายละเอียดต่อไป

สำหรับความท้าทายของการพัฒนาระบบนี้คือ รูปแบบประโยคของภาษาไทย เนื่องจากภาษาไทยไม่มีการแบ่งคำที่ชัดเจนเหมือนภาษาอังกฤษที่จะมีการใช้เว้นวรรคคั่นระหว่างคำ ทำให้การวิเคราะห์รูปประโยคการแบ่งพารากราฟต่างๆในรูปแบบไฟล์ PDF ค่อนข้างยาก จึงต้องมีการใช้ Image processing รวมกับ Text analytic algorithm เข้ามาช่วยในการจำแนกคำต่างๆออกมากจากไฟล์ PDF อีกทั้งยังต้องทดลองหา Model ที่เหมาะสมสำหรับการเรียนรู้ tag ของแต่ละพารากราฟในแต่ละเอกสารอีกด้วย

ทางผู้จัดทำคาดหวังว่าระบบที่กล่าวมานี้ จะเป็นอีกหนึ่งแหล่งรวบรวมข้อมูลความรู้ ที่จะช่วยให้คุณครูสามารถศึกษาค้นคว้าเพื่อพัฒนาการเรียนการสอน และช่วยผลักดันให้การศึกษาในประเทศไทยสามารถพัฒนาก้าวหน้าไปได้อีกขั้น

**2.2 คำสำคัญ**

**Machine Learning, Clustering, Latent Semantic Analysis, Classification, One vs Rest, Big data, Spark, Impala, PDF to Text, Keyword extraction, Text Analysis, Tag Box Extraction from PDF**

**3. หลักการและเหตุผล**

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

หนึ่งในปัจจัยสำคัญที่มีผลต่อการพัฒนาการศึกษา คือการสอนของคุณครู เพราะคุณครูเปรียบเสมือนผู้ที่ถ่ายทอดความรู้ต่างๆและพัฒนาเด็กให้เติบโตไปเป็นทรัพยากรมนุษย์ที่มีคุณภาพ ดังนั้น รัฐบาลไทยจึงมีการให้ทุนสนับสนุนกับครูในการศึกษาค้นคว้า ทดลองหาวิธี ในการพัฒนาการเรียนการสอนให้มีประสิทธิภาพมากยิ่งขึ้น และเข้าถึงเด็กนักเรียนได้มากขึ้น โดยเฉพาะการศึกษาขั้นพื้นฐานในระดับประถามและมัธยมศึกษา ซึ่งในปัจจุบัน ได้มีเอกสารที่ถูกเขียนออกมาเพื่อรายงานผลการทดลอง และวิธีในการพัฒนาการเรียนการสอนที่ดี (best practice) ซึ่งรายงานเหล่านี้มักจะประกอบไปด้วยจำนวนหน้าที่มาก และอยู่ในรูปแบบของไฟล์ PDF ทำให้ครูสามารถสืบค้นข้อมูลได้ยาก และต้องเสียเวลาในการอ่านหนังสือหลายร้อยหน้าจำนวนหลายเล่มเพราะ เอกสารไม่มีการรวบรวมและจัดเป็นหมวดหมู่ ทำให้ประสิทธิภาพในการสืบค้นข้อมูลนั้นไม่ดี อีกทั้งยังอาจได้ข้อมูลที่ไม่ครบถ้วน

ทางผู้จัดทำจึงจะทำการรวบรวมเอกสารรายงานเหล่านี้ เพื่อทำให้เกิดเป็น knowledge sharing platform ที่คุณครูสามารถเข้ามาสืบค้นหามูล และศึกษาค้นคว้าได้อย่างง่าย ภายใต้โครงการ Building a knowledge sharing platform with Text mining and knowledge extraction โดยได้รับความร่วมมือกับมูลนิธิ….. ซึ่งระบบนี้ จะช่วยรวบรวมและทำการจัดระเบียบเอกสารให้มีความเป็นระบบมากขึ้น รวมทั้ง ทำการวิเคราะห์ คัดแยกเนื้อหาส่วนต่างๆ ในไฟล์เอกสาร และทำการ tag ข้อความเหล่านั้นให้โดยอัตโนมัติว่า เนื้อหาในส่วนนั้นๆ มีความเกี่ยวข้องกับเรื่องอะไรบ้าง และทำการจัดเก็บข้อมูลเหล่าลงไปยังระบบฐานข้อมูล ซึ่งเปรียบเสมือนเครื่องมือที่ช่วยคัดกรองเนื้อหาขั้นต้นในเรื่องที่คุณครูสนใจศึกษา เพื่อให้สามารถทำการสืบค้นได้ง่ายและรวดเร็วยิ่งขึ้น สามารถนำความรู้เหล่านี้ไปพัฒนาการเรียนการสอน และพัฒนาให้การศึกษาของไทยก้าวไปสู่ในระดับต้นๆของโลกได้

**4. วัตถุประสงค์**

* เพื่อสร้างฐานข้อมูลที่ก่อให้เกิดเป็น knowledge sharing platform ของประเทศไทยได้
* เพื่อให้คุณครูสามารถสืบค้นข้อมูลได้อย่างมีประสิทธิภาพและรวดเร็ว สามารถนำไปพัฒนาการเรียนการสอน เพื่อช่วยให้การศึกษาของไทยดีขึ้น
* เพื่อศึกษาการทำ Document clustering และ Topic discovery สำหรับการคัดแยกเนื้อหา และจัดหมวดหมู่เอกสาร
* เพื่อสร้าง Machine Learning Model สำหรับเรียนรู้เอกสารภาษาไทยและ tag ที่กำหนดเพื่อใช้สำหรับการประมวลผลออกมาเป็นพารากราฟที่สำคัญ และ tag ที่เกี่ยวข้องจากเอกสาร
* เพื่อสร้าง Web Application สำหรับค้นหา tag ที่สนใจ และแสดงผลลัพธ์ออกมาเป็น Paragraph และ Tag ที่เกี่ยวข้องพร้อมเอกสารฉบับสมบูรณ์ในรูปแบบ PDF

**5. ปัญหาหรือประโยชน์ที่เป็นเหตุผลให้ควรพัฒนาโปรแกรม**

ปัจจุบันการศึกษาค้นคว้าหาข้อมูลเพื่อที่จะนำมาพัฒนาการเรียนการสอนของคุณครูนั้นทำได้ยาก เนื่องจากเอกสารข้อมูลต่างๆถูกเก็บอย่างกระจัดกระจาย ทำให้ถูกนำมาใช้ประโยชน์ได้เท่าที่ควร และในขณะ เดียวกันเอกสารรายงานเหล่านี้มักจะประกอบไปด้วยจำนวนหน้าที่มาก และอยู่ในรูปแบบของไฟล์ PDF ทำให้ครูสามารถสืบค้นข้อมูลได้ยาก และต้องเสียเวลาในการอ่านเพื่อหาใจความสำคัญเฉพาะในส่วนที่ต้องการเรียน รู้เพื่อนำไปพัฒนา ตามที่ได้กล่าวมาแล้วในข้อที่ 3

ดังนั้น โครงการ Building a knowledge sharing platform with Text mining and knowledge extraction จะช่วยสร้างฐานข้อมูลรวบรวมความรู้ที่เกี่ยวกับการพัฒนาการเรียนการสอน และก่อให้เกิดเป็น knowledge sharing platform ของประเทศไทย เพื่อให้คุณครูสามารถเข้ามาค้นคว้าหาความรู้ได้ง่ายยิ่งขึ้น ซึ่งระบบนี้จะช่วยคัดแยก จัดหมวดหมู่เอกสารเป็นประเภท รวมถึงช่วยคัดเลือกส่วนที่เป็นความสำคัญของเนื้อหาในรายงานแต่ละเล่ม เพื่อมาแสดงให้เห็นสาระสำคัญของรายงานเล่มนั้นๆ ซึ่งเปรียบเสทือนเป็นเครื่องมือที่ช่วยคัดกรองเนื้อหาที่คุณครูสนใจได้อีกทางหนึ่ง ซึ่งจะช่วยให้คุณครูสามารถเข้าถึงองค์ความรู้เหล่านี้ได้อย่างครบถ้วน สามารถค้นหาข้อมูลได้อย่างรวดเร็ว และนำความรู้ไปพัฒนาการเรียนการสอนเพื่อให้การศึกษาของเด็กไทยก้าวหน้าต่อไป

ทั้งนี้ หลังจากโครงการพัฒนาสำเร็จ ระบบนี้ยังสามารถนำไปประยุกต์ใช้กับเรื่องอื่นๆได้ เช่น วิธีการเพาะปลูกให้ได้ผลิตผลที่ดี เป็นต้น โดยการเปลี่ยนข้อมูลของการ train model (ไฟล์ PDF และข้อมูล tag)

**6. เป้าหมายและขอบเขตของโครงการ**

เป้าหมายของโครงการนี้ คือการสร้าง knowledge sharing platform ที่จะช่วยรวบรวมข้อมูลรายงานตัวอย่างการเรียนการสอนที่ดี (best practice) มาทำการคัดแยก จัดหมวดหมู่ เพื่อให้คุณครูสามารถสืบค้นข้อมูลได้อย่างมีประสิทธิภาพและรวดเร็ว สามารถนำความรู้ไปพัฒนาและปรับใช้ตาม เพื่อให้เกิดการเรียนการสอนที่ดีที่จะช่วยพัฒนาศักยภาพของเด็กนักเรียนได้ โดย platform มีขอบเขต ดังนี้

* สามารถรับไฟล์ PDF ภาษาไทย เพื่อทำการระบุและแบ่งพารากราฟต่างๆจากไฟล์นั้น และทำการดึงข้อความภาษาไทยในแต่ละพารากราฟออกมาผ่านกระบวนการ Text processing เพื่อให้ได้ผลลัพธ์สุดท้ายเป็นคำที่สำคัญต่างๆ สำหรับนำไปใช้ในการทำ Machine Learning
* สร้าง Machine learning model ที่สามารถรับข้อมูล text file ภาษาไทยที่ได้จากขั้นตอนที่ 1 และทำการติด tag ของแต่ละพารากราฟในไฟล์นั้นๆ โดยวิธีการ supervised classification ซึ่งจะแบ่งเป็น 2 ขั้นตอนคือ
  + การทำ training model จะรับ text และ tag ของ paragraph
  + การทำ prediction จาก model โดยจะรับ text เป็น paragraph และแสดง tag เป็นผลลัพธ์และทำการเก็บลงใน Database
* Web application เพื่อที่จะใช้ในการสืบค้นข้อมูลที่ทำการ tag มาแล้วจากขั้นตอน prediction จากใน Database

**บรรณานุกรม**

[1]World Economic Forum, **Thailand Report** [Online]**,** Available: http://www3.weforum   
 .org/docs/GCR2014-15/THA.pdf [2016, October 18]