**สาระสำคัญ**

เนื่องจากในปัจจุบัน เอกสารความรู้ต่างๆ มากมาย ยังไม่ถูกนำมาจัดเก็บและจัดหมวดหมู่ให้เป็นระบบ ทำให้องค์ความรู้เหล่านั้นไม่สามารถนำมาใช้งานได้อย่างมีประสิทธิภาพ โดยในปัจจุบันนั้น การที่จะนำความรู้ที่มีอยู่เหล่านั้นมาจัดหมวดหมู่เพื่อใช้ในการสืบค้น จะต้องใช้มนุษย์เป็นผู้จัดการ ซึ่งจำเป็นต้องใช้แรงงานคนในการมาจำแนกหมวดหมู่เหล่านั้น โดยการพัฒนาการจัดการองค์ความรู้หรือ Knowledge management นี้ จะช่วยให้การนำความรู้ที่มีอยู่มาใช้ได้อย่างมีประสิทธิภาพมากขึ้น และช่วยส่งเสริมการพัฒนาประเทศชาติให้มุ่งไปสู่ความเป็นสังคมอุดมปัญญาได้ โดยที่ระบบ Knowledge management ที่มีอยู่ในปัจจุบันนั้น ยังไม่มีระบบที่สามารถนำเอาเอกสารไปทำการวิเคราะห์และ tag หมวดหมู่ของเนื้อหาได้โดยอัตโนมัติ โดยเฉพาะอย่างยิ่งในส่วนของภาษาไทย ซึ่งยังไม่มีนักพัฒนารายใดพัฒนาเทคโนโลยีในลักษณะนี้ออกมา ดังนั้น เราจึงคิดที่จะทำระบบที่สามารถนำเอาเอกสารหรือบทความต่างๆที่ถูกจัดเก็บไว้ในรูปของ PDF นำไปวิเคราะห์ข้อความและทำการสรุปว่า ข้อความนี้มีเนื้อหาที่เกี่ยวข้องกับเรื่องใดบ้าง และนำไปจัดเก็บลงไปยังฐานข้อมูล เพื่อให้สามารถทำการสืบค้นได้ โดยสิ่งที่ท้าทายสำหรับการทำโครงงานชิ้นนี้ก็คือ การที่ภาษาไทยไม่มีการแบ่งคำที่ที่ชัดเจนเหมือนภาษาอังกฤษที่มีการใช้ space คั่น 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

**คำสำคัญ**

**Machine Learning, Clustering: LDA(Latent Dirichlet allocation), Classification: One vs Rest, Big data, Spark, PDF to Text, Keyword extraction, Text Analysis**

**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 เนื้อหาเหล่านั้นได้โดยการใช้ Machine Learning แต่ว่าโปรแกรม Stanford Deepdive ที่กล่าวมานั้น ถูกพัฒนาขึ้นสำหรับการวิเคราะห์เอกสารในภาษาอังกฤษเป็นหลัก ซึ่งภาษาไทยที่มีรูปแบบของประโยค การจัดเรียงคำ การวางตำแหน่งคำ,ตัวอักษร และอื่นๆ ที่แตกต่างจากภาษาอังกฤษเป็นอย่างมาก ทำให้การทำระบบ tag เอกสารอัตโนมัติสำหรับภาษาไทยนั้น ไม่สามารถใช้ Stanford Deepdive ได้ และการ tag หมวดหมู่ให้กับข้อความจำนวนมากโดยใช้มนุษย์ในการจัดการนั้น จะเป็นการเสียเวลาไปเป็นจำนวนมาก ทำให้ทางกลุ่มของเราสนใจที่จะพัฒนาโปรแกรมในลักษณะคล้ายกันกับ Stanford Deepdive ที่สามารถนำมาใช้กับภาษาไทยได้ เพื่อให้เอกสารต่างๆ ที่เป็นภาษาไทยนั้น ถูกนำมาใช้ประโยชน์ และนำมาศึกษาต่อได้อย่างมีประสิทธิภาพ

**เป้าหมายและขอบเขตของโครงการ**

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

**รายละเอียดของการพัฒนา**

**Story board**

สิ่งที่เราจะพัฒนาคือ machine learning model ที่ใช้รับ PDF file ที่เป็นภาษาไทยมาประมวลผลและทำการ tag แต่ละบทความใน PDF file นั้นๆ ว่าบทความแต่ละอันมีความเกี่ยวข้องกับเรื่องใดบ้าง โดยจะแบ่งขั้นตอนการทำงานเป็น 3 ส่วนหลักๆ ได้แก่ 1. การแปลง PDF เป็น text file และ preprocess 2. การทำ data clustering เพื่อดึง keyword คำ 3. การทำ classification ซึ่งการทำ classification นั้นจะแบ่งเป็น 2 ส่วนคือ 3.1 training data โดยจะนำ keyword ที่มีอยู่จากส่วนที่ 2 มาแก้ไขและติด label เพื่อนำไป train ใน classifier 3.2 testing data จะนำ keyword จากส่วนที่ 2 มาทำการ classify จาก training model เพื่อให้ได้ผลลัพธ์เป็น tag ของ document เหล่านั้น และมีการทำ web application เพื่อเรียกใช้งาน model ที่ทำขึ้น โดยการทำระบบค้นหาข้อมูลจาก tag ต่างๆ แล้วตัว web application จะทำการดึง paragraph ที่เกี่ยวข้องกับ tag ที่ค้นหาขึ้นมาแสดงผลที่หน้าเว็บ และมีลิงค์ให้ดาวน์โหลดไฟล์ PDF ที่มีข้อความนั้นอยู่

**เทคนิคหรือเทคโนโลยีที่ใช้**

* Hadoop เป็น platform สำหรับทำการจัดเก็บ และ ประมวลผลข้อมูลขนาดใหญ่ (Big Data) ซึ่งสามารถรองรับการขยายตัวของข้อมูล และมีความน่าเชื่อถือสูงเนื่องจากสามารถทำการประมวลผลแบบกระจายโดยผ่านเครื่อง computer มากมายที่อยู่ใน cluster ได้ง่าย
* Apache Spark เป็น engine สำหรับการทำการประมวลผลข้อมูลขนาดใหญ่ (Big data processing) ที่สามารถทำงานได้อย่างรวดเร็ว เนื่องจากใช้การประมวลผลในหน่วยความจำหลัก (In-memory processing) ทำให้การเข้าถึงข้อมูลทำให้รวดเร็วมากขึ้น
* Natural Language Processing เป็นอัลกอริทึมที่ใช้เพื่อวิเคราะห์ภาษาธรรมชาติของมนุษย์ ด้วยการรับค่าที่เป็นประโยค จากนั้นทำการวิเคราะห์ทั้งในเชิงโครงสร้าง และความหมายทางภาาา โดยการทำงานร่วมกับฐานความรู้ และทำการคำนวณและประเมินค่าเพื่อหาคำตอบ แล้วแทนความหมายสร้างเป็นผลลัพท์ซึ่งเป็นภาษาธรรมชาติออกมา
* SQL หรือ Structured Query Language เป็นภาษาที่ใช้ในการติดต่อกับฐานข้อมูลหรือพูดอีกอย่างก็คือ เป็นภาษาที่ใช้ในการสั่งให้ฐานฐานข้อมูลกระทำการใด ๆ ตามคำสั่งที่เราสั่ง

**เครื่องมือที่ใช้ในการพัฒนา**

* โปรแกรม PDF2TXT ใช้สำหรับการแปลงไฟล์ในรูปแบบ PDF ให้เป็น text file
* โปรแกรม Spark ML เป็นโปรแกรมที่ใช้ทำ Machine Learning โดยที่สามารถทำงานแบบขนาน (Parallel programming) ได้
* โปรแกรม LexTo ใช้สำหรับการตัดคำในภาษาไทย
* โปรแกรม Apache Impala เป็นโปรแกรมจัดการฐานข้อมูลที่ใช้งานกับ Hadoop system
* ภาษา Python สำหรับการเขียนโปรแกรมที่ใช้งานกับ Spark ML.

**รายละเอียดโปรแกรมที่จะพัฒนา**

**Input/Output Specification**

* Input เป็นเอกสารไฟล์ PDF ภาษาไทย ที่มีเนื้อหาเกี่ยวข้องกับสิ่งที่ machine learning model ได้เรียนรู้ไว้แล้ว
* Output เป็น tag ของแต่ละบทความ โดย tag ที่ได้จะมาจากการ label ที่กำหนดไว้

**Functional Specification**

**Text Processing**

* ทำการแปลง PDF ให้เป็น text fileโดยการใช้ PDF2MINER
* ทำการ preprocess text file ที่ได้ โดยการตัด tagged PDF ที่ติดมาหลังจากการทำการแปลง และทำการแก้ภาษาไทยที่ผิดพลาด เช่น สระ อำ (อ า เป็น อำ)
* นำเอา text file ที่ได้มาทำการตัดคำ (word segmentation) โดยใช้ Lexitron
* ทำการนำ stop word ออก เพื่อลดคำที่ไม่จำเป็นออก
* ทำการ word count และทำ word to vector ของแต่ละ document

**Clustering**

* นำ vector ของ document มาทำการ matrix reduction เพื่อให้ได้ keyword ที่สำคัญออกมา โดยจะใช้ LDA (Latent Dirichlet allocation) หรือ LSA (Latent Semantic Analysis) เพื่อให้ได้ keyword ออกมา

**Classification: Training model**

* นำ keyword ที่ได้จากการทำ Clustering มา label
* นำ keyword และ label ที่ได้มา training classifier model
* ใช้ Multiclass and multilabel algorithms (One-Vs-The-Rest)

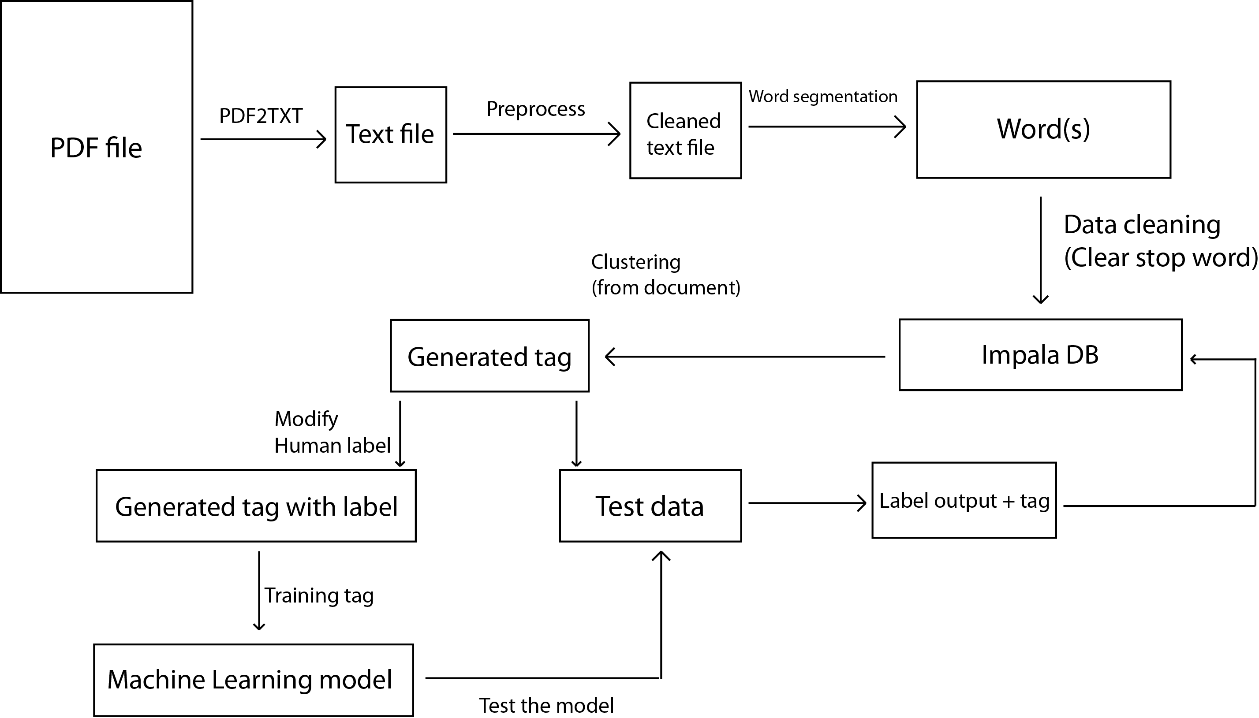
**Classification: Test model**

* นำ keyword ที่ได้จากการ Clustering ของ test document มาทำการ Classification โดยใช้ training model ที่ได้จากขั้นตอนการทำ Training model เพื่อบอก tag ของ output ออกมา
* ใช้ Multiclass and multilabel algorithms (One-Vs-The-Rest)

**Database** แบ่งเป็น 3 ส่วนหลักๆ คือ

* Keyword database เก็บข้อมูลคำที่เป็น keyword ต่างๆ โดยเก็บคำและความถี่ที่ปรากฏของคำนั้นๆ
* Document database เก็บข้อมูลเอกสาร โดยเก็บชื่อเอกสาร เนื้อหา และ tag
* Label database เก็บข้อมูลว่าแต่ละ label มีคำอะไรบ้าง

**Design**

****

**บรรณานุกรม**

* มาเริ่มเรียนรู้ Hadoop กันหน่อย, <http://www.somkiat.cc/start-with-hadoop/> (Accessed 2016-9-23)
* Apache Spark, <http://spark.apache.org> (Accessed 2016-9-23)
* SQL Database & table, <http://www.thaicreate.com/tutorial/sql-database-introduction.html> (Accessed 2016-9-23)
* Pdfminer, <http://euske.github.io/pdfminer/index.html>
* One-vs-Rest classifier, <https://spark.apache.org/docs/latest/ml-classification-regression.html#one-vs-rest-classifier-aka-one-vs-all>
* Lexto , http://www.sansarn.com/lexto/
* Latent Dirichlet allocation(LDA), https://spark.apache.org/docs/1.6.0/ml-clustering.html#latent-dirichlet-allocation-lda