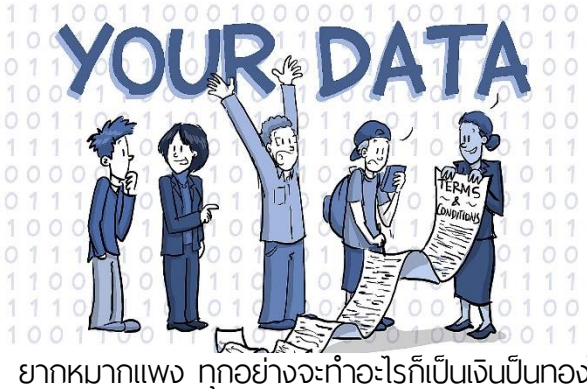
THAI TEXT ANALYTICS

“ความยากของการพัฒนา Text Analytics ภาษาไทย”

ระบบการประมวลผลภาษาธรรมชาติ (NLP) ค่อนข้างมีความยากต่อการพัฒนาด้วยเหตุผลต่างๆ ดังนั้นบทความนี้จึงขอนำเสนอความท้าทายของการทำ Thai Text Analytics ให้ทุกคนได้เห็นและเข้าใจ บางคนเห็นหัวข้อนี้แล้วอาจกำลังเลิกคิดที่จะทำ Thai Text Analytics ไปเลย ตั้งสติก่อนครับ ทุกปัญหาย่อมมีแสงสว่างที่ปลายอุโมงค์เสมอ ก็ขอให้ข้อมูลในบทความนี้เป็นแนวทางในการแก้ปัญหาและรับมือ (หรือเตรียมใจ) กับความท้าทายในการพัฒนา Thai Text Analytics แล้วกันครับ ผมจะแบ่งเป็น 2 หัวข้อดังนี้

1. ข้อมูล (Data Source)



ในเรื่องของข้อมูล การไม่มีข้อมูลให้ใช้เป็นวัตถุดิบพัฒนาเทคโนโลยีวิเคราะห์ข้อความนับเป็นปัญหาใหญ่เป็นอันดับต้นต้นๆ ของวงการ Data scientist หรือ Big Data เลยก็ว่าได้ ขึ้นชื่อว่าทำงานด้าน Data แล้ว ถ้าไม่มี Data ให้เล่นนี้มันก็แปลกๆนะครับ (55555555) ความจริงแล้วภาษาไทยเราคนทั่วไปอาจจะมองว่าข้อมูลมีเยอะแยะไปทำไมคนถึงบ่นว่า Data น้อย นั่นก็เพราะเราอยู่ในยุคข้าวยากหมากแพง ทุกอย่างจะทำอะไรก็เป็นเงินเป็นทองไปซะหมดทุกอย่างดังนั้นของที่เอามาสร้างมูลค่าได้ ก็ย่อมแพงและหายากอยู่แล้ว ข้อมูลก็เหมือนกัน เป็นอะไรที่มีประโยชน์มาก ถึงมากที่สุดสำหรับคนที่อยู่ในวงการ ข้อมูล กว่าจะได้เอามาวิเคราะห์แต่ละทีนั้นยากเย็นแสนเข็ญมากครับ และถ้าเสียเงินราคาก็ใช่ว่าจะถูกๆซะเมื่อไหร่ แต่ถ้าถามหาของฟรี!!! ล่ะมีไหม? มีครับ แต่มันน้อยจริงๆ บางทีอาจไม่เพียงพอ แต่ก็เอามาวิเคราะห์ได้ครับผม ดังนั้นมันเป็นเรื่องยากมากสำหรับการหาข้อมูลเพื่อมาวิเคราะห์

“จะทำวิเคราะห์ข้อมูล แต่ดันไม่มีข้อมูลหรือมีน้อยก็ จบ!! สิครับ”

2. ภาษาไทย

ง่ายๆ สั้นๆ เลยนะครับ *“ภาษาไทย”* ยากจริงอะไรจริง เป็นภาษาที่ *อัลไล* ก็ *มะรุ* 55555 กำลังคิดว่าผมเขียนผิดใช่ไหมครับ ใช่แล้วครับ ผมตั้งใจเขียนผิด เพราะอยากให้เห็นว่าภาษาเป็นอะไรเปลี่ยนแปลไปตามยุคสมัยจริงๆ แล้วยิ่งภาษาไทยยิ่งเปลี่ยนด้วยอัตราความเร็วแบบชัวโมงต่อชั่วโมงเลยก็ว่าได้ นักวิจัยก็วิ่งตามแทบไม่ทัน มีคำพูดดาราหรือคำพูดติดปาก ประโยคเด็ด ก็จะถูกยกมาเป็นกระแสดังในช่วงนั้นหรือ อาจจะติดปากไปเรื่อยๆ ซ้ำยังเป็นภาษาที่เพิ่งเกิดขึ้นใหม่และยังใช้คำผิดอีก นักวิจัยอย่างเราๆ ท่านๆ ก็ปวดหัวสิครับ ยิ่งนักวิจัยที่ทำ Text Analytics แล้วยิ่งเอามือกุมขมับกันเป็นแถวๆ เพราะโมเดลตัดคำที่มีอยู่ก็ตัดคำพวกนี้ไม่ได้อีก จะแก้ใหม่ให้ทันตามภาษาที่เกิดขึ้นใหม่ก็เกรงว่าจะต้องมานั่งแก้ทุกอาทิตย์น่ะสิ เห็นแบบนี้ก็จะไม่ให้ปวดหัวหายใจไม่ทั่วท้องก็คงไม่ไหวล่ะครับ สำหรับการพัฒนาเทคโนโลยีอย่าง Thai Text Analytics

“ภาษาไทย ยากมากจิงจิ้งงงงงงงงงง(งองู ล้านตัว) #เสียงสูง ”