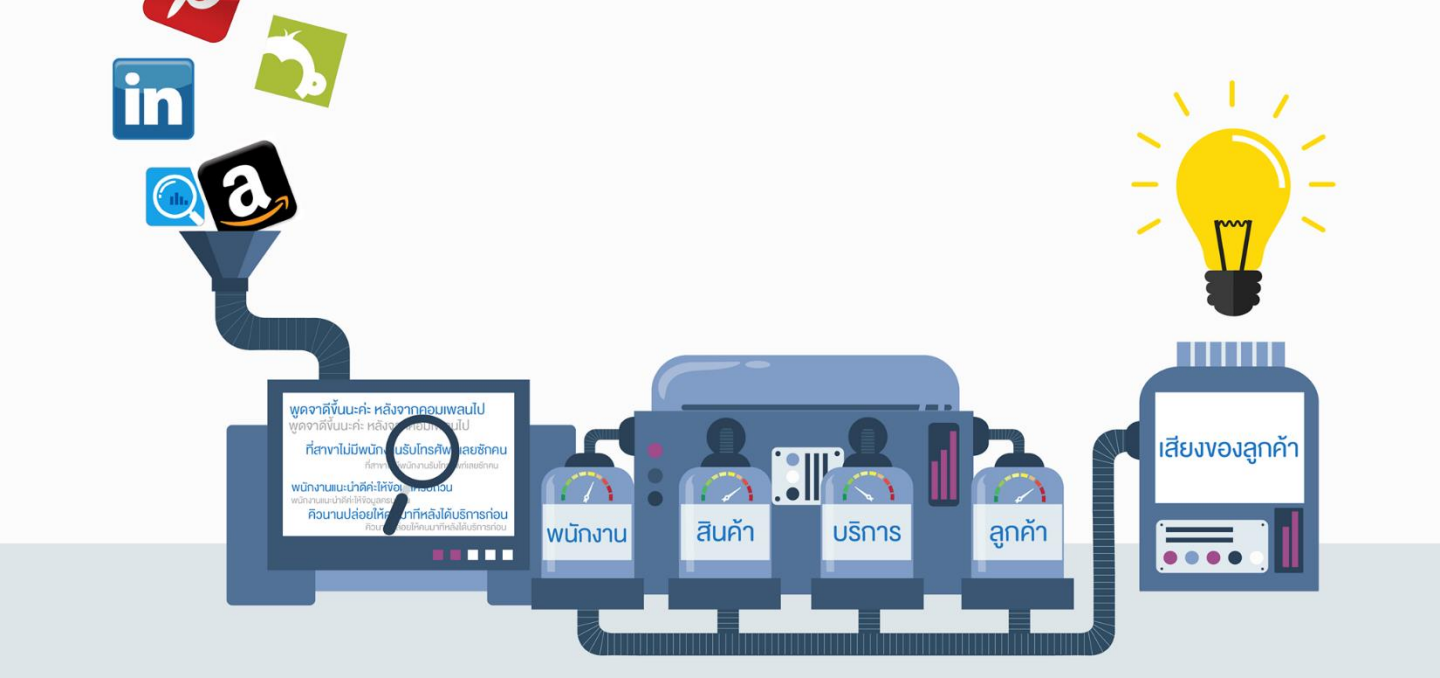

THAI TEXT ANALYTICS

"ความยากของการทำ Text Analytics ภาษาไทย"

ระบบการประมวลผลภาษาไทย (NLP) ค่อนข้างมีความยาก ที่ประกอบด้วยเหตุผลต่างๆรวมกันมากมาย ดังนั้นจึงนำเสนอเรื่องที่เกี่ยวกับการทำ Thai Text Analytics ให้ ทุกคนได้เห็นถึงปัญหาความยากของมัน แต่ไม่ใช่ว่ามันจะไม่มีทางออกหรือบางคนเห็นหัวข้อนี้แล้วเลิกคิดที่จะทำ Thai Text Analytics ไปเลย ตั้งสติก่อนครับ อย่างน้อยมันก็มีมุมดีๆบ้างเหละครับ ก็เอาไปเป็นแนวทางในการแก้ปัญหาและรับมือกับมัน ความยากของการทำ Thai Text Analytics ผมจะแบ่งเป็น 2 หัวข้อดังนี้ครับ

1. ข้อมูล (Data Source)

ในเรื่องของข้อมูลก็เป็นปัญหาใหญ่เป็นอันดับต้นของวงการ Data scientist หรือ Big Data เลยก็ว่าได้ ขึ้นชื่อว่าทำงาน Data แล้ว ไม่มี Data ให้เล่นนี้มันแปลกๆนะครับ ความจริงแล้วภาษาไทยเราอาจจะมองว่าข้อมูลมีเยอะแยะ แต่ทำไมคนถึงบนว่า Data น้อย สืบเนื่องมาจากเราอยู่ในข้าวยากหมากแพง ทุกอย่างขะทำอะไรก็เป็นเงินป็นท้องไปซะหมดทุกอย่างดังนั้นของที่เอามาสร้างมูลค้าได้ ก็ย่อมแพงและหายากอยู่แล้ว ข้อมูลก็เหมือนกัน เป็นอะไรทีมีประโยชน์มาก ถึงมากที่สุดสำหรับคนที่อยู่ในวงการ ข้อมูลกว่าจะได้เอามาวิเคราะห์แต่ละทียากเย็นแสนเข็นมากครับ และถ้าเสียเงินราคาก็ใช่ว่าจะถูกๆซะเท่าไหร่ แต่ถ้าถามหาของฟรี!!! หล่ะมีไหม มีครับ แต่มันน้อยจริงๆ บางทีอาจไม่เพียงพอ แต่ก็เอามาวิเคราะห์ได้ครับผม ดังนั้นมันเป็นเรื่องยากมากสำหรับการหาข้อมูลเพื่อมาวิเคราะห์

"จะทำวิเคราะห์ข้อมูล ที่ข้อมูลไม่มีหรือมีน้อยก็ จบ!! สิครับ"

2. ภาษาไทย

ง่ายๆ สั้นๆ *"ภาษาไทย"* อยากจริงอะไรจริง เป็นภาษาที่ *อัลไล* ก็ *มะรุ* 55555 กำลังคิดว่าผมเขียนผิดใช่ไหมครับ ใช่แล้วครับเขียนเขียนผิด แต่ตั้งใจเขียนผิด ภาษาเป็นอะไรเปลี่ยนแปลไปตามยุคสมัยจริงๆ แล้วยิ่งภาษาไทยยิ่งเปลี่ยนตาม ชัวโมงเลยก็ว่าได้ นักวิจัยแทบตามไม่ทัน มีคำพูดดาราหรือคำพูดติดปาก ประโยคเด็ด ก็จะถูกยกมาเป็นกระแสดังในช่วงนั้นหรือ อาจจจะติดปากไปเรื่อยๆ ซ้ำยังเป็นภาษาที่เพิ่งเกิดขึ้นใหม่ และยังใช้คำผิดอีก นักวิจัยอย่างเราๆก็ปวดหัวสิครับ และยิ่งนักวิจัยที่ทำ Text Analytics แล้วยิ่งเอามือกุมขมับกันเป็นแถวๆ โมเดลตัดคำที่มีอยู่ก็ตัดคำพวกนี้ไม่ได้อีก จะแก้ใหม่ให้ทันตามภาษาที่เกิดขึ้นใหม่ เกรงว่าจะต้องมานั่งแก้ทุกอาทิตย์หล่ะ สิ เห็นแบบนี้ก็ปวดหัวกันเลยทีเดียว สำหรับการทำ Thai Text Analytics ใช่ไหมหล่ะครับ

"ภาษาไทย ยากมากจิงจิ๊งงงงงงงงงงง(งอง ล้านตัว) #เสียงสูง "