การวิเคราะห์ความนิยมของโพสต์บนทวิตเตอร์โดยการใช้การวิเคราะห์หัวข้อคำ
Analyzing Twitter Post Popularity Using Topic Analysis

พุดธิดา นิลสมบูรณ์, สุวดี เจียมจิรังกร และ อ.ดร.ศิริสรรพ เหล่าหะเกียรติ
หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์

ABSTRACT

At present, most people are interested in social media, and Twitter is one of the most popular social networks. We analyzed Twitter's information by focusing on popular users that are news channels which post regularly. We summarized the characteristics of the sets of topic words that cause Twitter's posts to be popular using topic analysis. We acquired posts from Twitter and classified posts into popular and unpopular posts based on the number of responses in each post. Then preprocessing steps were performed and Latent Dirichlet Allocation (LDA) was subsequently applied to find the topic words that are outstanding for popular as well as unpopular posts. These topics words can indicate the current interest of the audience of each channel.

องค์ความรู้ที่เกี่ยวข้อง

Data Cleansing [1]

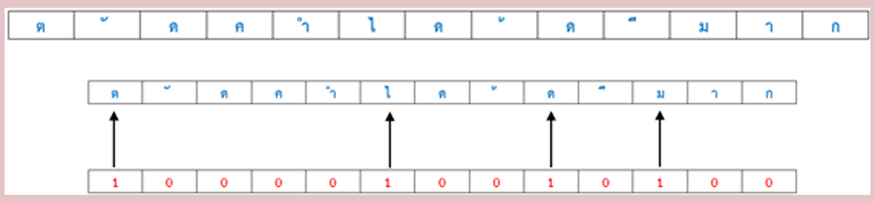
เป็นกระบวนการตรวจสอบและการแก้ไข (หรือลบ) รายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูล โดยงานวิจัยนี้จะใช้ Tokenization และ Remove Stop words เพื่อมา Cleansingข้อมูล

Tokenization หรือการกำหนดขอบเขตของคำต่างๆ ให้คอมพิวเตอร์เข้าใจ ส่วนใหญ่แล้วเราสามารถใช้คำที่ถูกแบ่งคำเป็นคำๆ ให้เป็น token ขอคำนั้นๆได้เลย

Stop words คือคำที่จะไม่เอามาประมวลผล ซึ่งคำพวกนี้ไม่มีความสำคัญ อย่างเช่น is am are ไม่มีความหมายแต่ปรากฏบ่อย

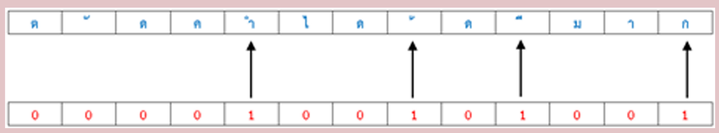
ระบบตัดคำแบบ Deep Learning [2]

deepcut ระบบตัดคำแบบ deep learning โดยพัฒนาด้วย Keras หลักการทำงาน เมื่อ AI เห็นข้อความก็จะรู้ว่าประกอบด้วย



โดย AI จะมอง 1 ประโยค เป็นอาร์เรย์ของชุดตัวอักษร จากนั้น AI ก็จะบอกว่าตัวอักษรไหนเป็นตัวเริ่มต้นของคำ ถ้าใช่ก็จะให้คำตอบว่า 1

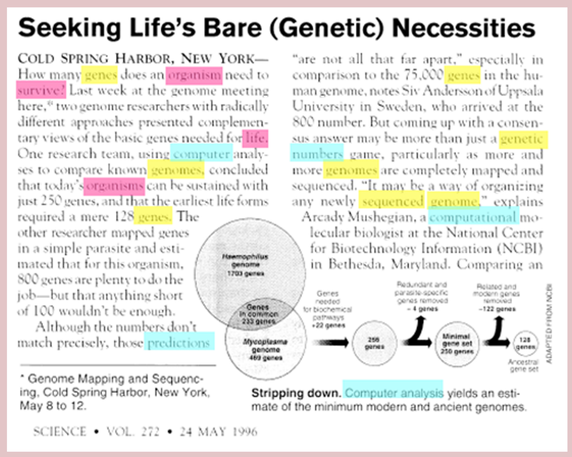
แต่เพื่อความสะดวก deepcut.tokenize ฟังก์ชันก็จะตัดแบ่งคำออกมาให้เลย อัตโนมัติโดยเบื้องหลังการทำงานจะใช้เทคนิคเลื่อนเลข 1 ในอาร์เรย์ไปทางซ้าย ส่วนอักษรตัวสุดของประโยคก็จะระบุว่าเป็น 1 ไปเลย



จากนั้นภายในไลบรารีจะเริ่มไล่ตัวอักษร ตั้งแต่หมายเลข 0 กำกับ จนมาหยุดตรงตัวอักษรเลข 1 ก็จะถือว่าเป็น 1 คำไทย ด้วยเหตุนี้เมื่อใช้คำสั่ง deepcut.tokenize ก็จะให้ค่าออกมาเป็นอาร์เรย์ดังต่อไปนี้

```
['ตัด','คำ','ได้','ดี','มาก']
```

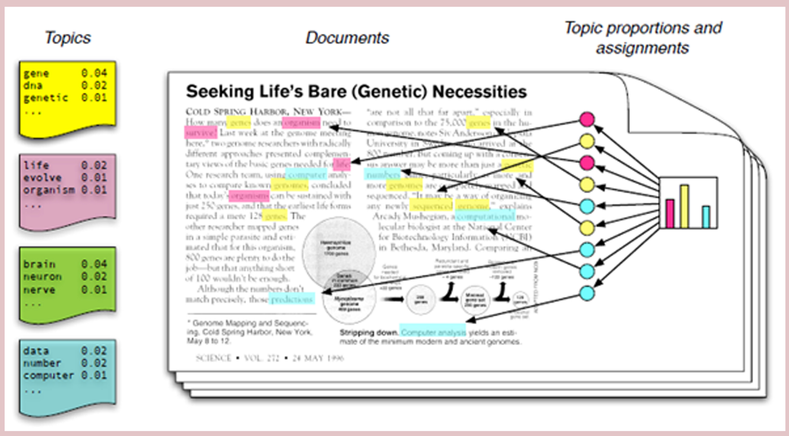
Latent Dirichlet Allocation (LDA) [3]



จากข้างบนอธิบายได้ว่าสำหรับ LDA นั้นมีสมมติฐานเกี่ยวกับโมเดล 3 ข้อคือ

1. Each topic is a distribution over words
2. Each document is a mixture of topics
3. Each words in document are picked or randomly from the above topics

สมมติถ้าเรามีเอกสารที่ชื่อว่า Seeking Life's Bare (Genetic) Necessities สิ่งที่เราจะได้คือ

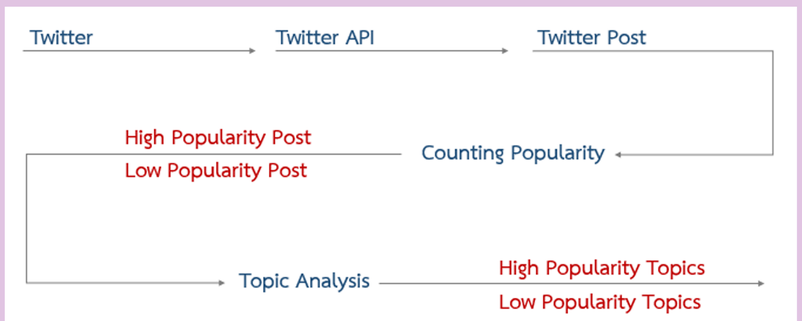


ต่อเนื่อง (discrete) เช่น เอกสารในรูปข้างบนนี้ ประกอบด้วย topic 1, topic 2 และ topic 3 ด้วยความน่าจะเป็น 0.35, 0.5 และ 0.15 ตามลำดับ

สมมติว่าเรากำหนดไว้ให้เอกสารนี้มี 4 Topics แล้วในแต่ละ Topics จะมีคำ ๆ นึงอยู่ด้วยความน่าจะเป็นเท่าไร พอได้ output แล้ว เราก็สามารถมาทำการ Post - Processing ต่อว่า

- Topic สีแดง มีคำ gene, dna, genetics อยู่ อาจจะเกี่ยวกับ Genetics
- Topic สีฟ้ามีคำว่า data, number, computer น่าจะเกี่ยวกับ Computer Science

วิธีการดำเนินงาน



สรุปผล

การวิเคราะห์ความนิยมของโพสต์บนทวิตเตอร์โดยการใช้การ วิเคราะห์หัวข้อคำ การนำข้อมูลมาวิเคราะห์นั้น จะทำให้สามารถทราบได้ว่าหัวข้อ หรือ Topic แบบใดเป็นที่นิยมอยู่และไม่ได้เป็นที่นิยมอยู่ในช่วงเวลานั้นๆ