**NLP**

**(Natural Language Processing)**

**การประมวลผลภาษาธรรมชาติหรือภาษามนุษย์ ทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ รวมไปถึงการประมวลผลที่ไม่ใช่แค่ทำให้คอมพิวเตอร์ แต่รวมถึงไปการวิเคราะห์ทางด้านภาษาศาสตร์ การตีความจากข้อความ โดยในการทำงานที่เกี่ยวกับ NLP จะต้องใช้ความรู้ในด้านหลายด้าน เช่น Mathematics, Linguistics, Psychology หรือ Philoshphy ดังนั้นเพื่อให้เห็นภาพรวมของงานด้าน NLP ได้ครอบคลุมที่สุด เป็น Field หนึ่งของ Computer science และ Artificial Intelligence ที่ช่วยให้ให้มนุษย์ปฏิสัมพันธ์กับคอมพิวเตอร์ด้วยภาษาธรรมชาติ หรือ ช่วยให้คอมพิวเตอร์เข้าใจภาษาธรรมชาติของมนุษย์**

**NLP ใช้ machine learning algorithms**กับข้อมูลประเภท text, speech หรือแม้แต่ภาพ ซึ่งสามารถประยุกต์ใช้ในงานทางด้าน NLP เช่น  
**document summarization** การสรุปความในภาษา ที่เป็นการลดความซับซ้อนและขนาดของเอกสารข้อความโดยไม่ทำให้ความหมายหรือสาระสำคัญของข้อมูลเอกสารสูญเสียไป

**machine translation เ**ครื่องมือที่ใช้สำหรับการแปลข้อความจำนวนมาก ๆ จากภาษาหนึ่งไปยังอีกภาษาหนึ่ง เช่น google translate

**spam detection**

**speech recognition**แปลงเสียงพูด (Audio File) เป็นข้อความตัวอักษร (Text) โดยสามารถแจกแจงคำพูดต่างๆ  
**question answering** การถามตอบ

**autocomplete**เวลาพิมพ์ตัวอักษรไปในช่องรับข้อมูลแล้วจะแสดงข้อมูลที่เคยพิมพ์ขึ้นมา  
**named entity recognition**การสกัดนิพจน์เฉพาะหรือชื่อเฉพาะในประโยค  
**predictive typing** การคาดเดาคำต่อไปเมื่อพิมพ์

TF-IDF หรือ Term Frequency-Inverse Document Frequency — เป็นหนึ่งในวิธีหา “คำ(term)” ที่สำคัญ ใน “เอกสาร(document)” โดยดูจากเนื้อหาของเอกสารทั้งหมด มักจะใช้ในงานพวก Information-retrieval หรือ Text mining

TF-IDF เกิดจากผลคูณของสองค่า คือ TF(Term Frequency) กับ IDF(Inverse Document Frequency) *Term ในที่นี้ ผมจะกำหนดให้เป็น “คำ” นะครับ แต่จริง ๆ แล้ว มันไม่ใช่*

Term Frequency

หรือ **ความถี่ของ คำ(term)**และเนื่องจากว่าแต่ละ เอกสาร(document) อาจจะมีความยาวที่แตกต่างกัน ดังนั้นการคำนวณ Term Frequency จึงมักจะหารด้วยความยาวของ document (จำนวนคำทั้งหมดใน document นั้น)

**วิธีการคำนวณ Term Frequency มีหลายวิธี เช่น**

* Raw counts
* Log normalization
* Double normalization

**max\_df** is used for removing terms that appear **too frequently**, also known as "corpus-specific stop words". For example:

* max\_df = 0.50 means "ignore terms that appear in **more than 50% of the documents**".
* max\_df = 25 means "ignore terms that appear in **more than 25 documents**".

The default max\_df is 1.0, which means "ignore terms that appear in **more than 100% of the documents**". Thus, the default setting does not ignore any terms.

**use\_idf*bool, default=True***

Enable inverse-document-frequency reweighting.

**smooth\_idf*bool, default=True***

Smooth idf weights by adding one to document frequencies, as if an extra document was seen containing every term in the collection exactly once. Prevents zero divisions.

**norm*{‘l1’, ‘l2’}, default=’l2’***

Each output row will have unit norm, either: \* ‘l2’: Sum of squares of vector elements is 1. The cosine similarity between two vectors is their dot product when l2 norm has been applied. \* ‘l1’: Sum of absolute values of vector elements is 1. See preprocessing.normalize.