

รายงานความก้าวหน้า 240-401 โครงงานวิศวกรรมคอมพิวเตอร์ 1 ครั้งที่ 1/2564

การประยุกต์ใช้การประมวลผลภาษาธรรมชาติเพื่อการวิเคราะห์ข่าวตลาด forex gold spot

An Application of Natural Language Processing on forex gold spot News Analysis

นายชวลชัย อภิชาตฐิติวรณ์

รหัสนักศึกษา 6210110646

อาจารย์ที่ปรึกษาโครงงาน

.......................................

(รศ.ดร.มนตรี กาญจนะเดชะ)

รายงานความก้าวหน้าโครงงานวิศวกรรมคอมพิวเตอร์นี้เป็นส่วนหนึ่งของการศึกษา

ตามหลักสูตรปริญญาวิศวกรรมศาสตร์บัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์

**บทที่ 1**

* 1. **ที่มาและความสำคัญ**

ในปัจจุบันสังคมมีการลงทุนในส่วนของหุ้น เทรดหุ้น forex หรือ cryptocurrency มากขึ้นและมีอาชีพเกิดขึ้นมากมายในวงการของการเล่นหุ้น หรือ Stock และมีการใช้ AI ต่าง ๆ เพื่อวิเคราะห์หาแนวโน้มหรือคาดการณ์ล่วงหน้าของกราฟหุ้น ซึ่งบางครั้งย่อมเกิดปัญหา AI วิเคราะห์ได้แค่ทฤษฎีของกราฟและเครื่องมือต่าง ๆ ที่ใช้กัน และการวิเคราะห์ข่าวนั้นย่อมคาดการณ์ได้ยากเพราะข่าวสารที่มากมายย่อมส่งผลกระทบต่อหุ้นนั้น ๆ

ดังนั้นการลงทุนในส่วนของหุ้นเทรดหุ้น หรือcryptocurrency ควรต้องมีการใช้เครื่องมือวิเคราะห์ข่าวเข้ามาเป็นส่วนหนึ่งของการคาดการณ์จากการเกร็งกำไร หรือ ลงทุน ซึ่งเห็นได้ชัดว่า AI ที่มีอยู่แล้วเช่น EA เป็นตัวช่วยให้เราไม่ต้องมาเทรดเองโดยเป็นการเซทค่าจาก เครื่องมือต่าง ๆตามเทคนิคคออยู่ของเราเอง ซึ่งในบางครั้งเรื่องของเทคนิคก็ผิดพลาดเพราะบางหลักทรัพย์มีความผันผวนสูง ทำให้เทคนิคที่เราเซทไว้อาจะเกิดการผิดพลาดได้สูง และส่งผลให้เราขาดทุน

ในการทำ AI วิเคราะห์ข่าวหุ้นได้มีการนำ Machine Learning NLP มาวิเคราะห์ ซึ่งเป็นตัวช่วยทำให้คอมพิวเตอร์วิเคราะห์ข่าวสารหรือข้อความได้อย่างง่ายดาย NLP เป็นสาขาหนึ่งในการเรียนรู้ของเครื่องด้วยความสามารถของคอมพิวเตอร์ในการทำความเข้าใจ วิเคราะห์ จัดการ และสร้างภาษามนุษย์ได้ ในปัจจุบัน เทคโนโลยี Machine Learning NLP เป็นที่นิยมในการเอามาทำ AI เช่น การดึงข้อมูล การแปลภาษา การทำให้ข้อความง่ายขึ้น การวิเคราะห์ให้ความรู้สึกของผู้ใช้ การสรุปข้อความ ตัวกรองสแปม คาดการณ์ผลการค้นหาของผู้ใช้ แก้ไขขำผิดอัตโนมัติ เป็นต้น Natural Language Processing(NLP) for Machine Learning หรือการประมวลผลภาษาธรรมชาติด้วย Python ซึงภาษา Python เป็นภาษาที่รวดเร็วและในการทำ NLP จะใช้ **Natural Language Toolkit** (NLTK) เป็นLibrary Opensource ยอดนิยมใน Python

* 1. **วัตถุประสงค์ของโครงงาน**

1. พัฒนาระบบที่ใช้หลักการของ NLP
2. วิเคราะห์ข่าวต่าง ๆ ที่เกี่ยวกับตลาด forex gold spot ซึ่งเป็นข่าวที่อยู่ในรูปแบบออนไลน์ เพื่อใช้เป็นข้อมูลสำหรับระบบ Robot Trader
   1. **ขอบเขตโครงงาน**
3. สร้าง Machine Learning AI วิเคราะห์ข่าว โดยใช้ Language Processing(NLP)
4. เฉพาะ forex gold spot
5. วิเคราะห์แนวโน้มขึ้นหรือลง แสดงเป็น 1 0 และ -1
6. วิเคราห์เฉพาะข่าวสำคัญที่อยู่ในตารางปฏิทิน
7. วิเคราห์เฉพาะตลาดอเมริกา ช่วงเวลา 7.00 pm – 3.00 am ตามเวลาประเทศไทย
8. จัดทำให้แสดงข้อมูลบนเว็บไซต์
   1. **แผนการดำเนินงาน**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| เวลา  กิจกรรม | PROJECT 1 | | | | | | | | | | PROJECT 2 | | | | | | | |
| มิ.ย. 65 | | ก.ค. 65 | | ส.ค. 65 | | ก.ย. 65 | | ต.ค. 65 | | ธ.ค. 65 | | ม.ค. 66 | | ก.พ. 66 | | มี.ค. 66 | |
| 1. พัฒนาอัลกอริทึม |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. พัฒนาเขียนโปรแกรมด้วย Python |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. เทรนสมองกล |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. ทดสอบระบบ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. ทดลองนำไปใช้จริง กับบัญชี Forex demo |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. ปรับปรุงแก้ไขอุปกรณ์ให้สมบูรณ์ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. จัดทำเว๊บไซต์ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. จัดทำรายงานให้สมบูรณ์ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**บทที่ 2**

**ทฤษฎีและความรู้พื้นฐาน**

**2.1 Forex gold spot**

Forex คือ ตลาดแลกเปลี่ยนเงินตราต่างประเทศ (หรือที่เรียกว่า forex หรือ FX) หมายถึงตลาดที่ซื้อขายกันโดยตรง (OTC) ระดับโลกซึ่งเทรดเดอร์ นักลงทุน สถาบัน และธนาคารจะแลกเปลี่ยน เก็งกำไร ซื้อและขายสกุลเงินของโลกการเทรดจะทำขึ้นใน ‘ตลาดระหว่างธนาคาร’ ซึ่งเป็นช่องทางทางออนไลน์ที่มีการเทรดสกุลเงิน 24 ชั่วโมงต่อวัน ห้าวันต่อสัปดาห์ Forex เป็นหนึ่งในตลาดการเทรดที่ใหญ่ที่สุดโดยมีเงินหมุนเวียนทั่วโลกในแต่ละวันโดยประมาณมากกว่า 5 ล้านล้านดอลลาร์สหรัฐฯ

Gold Spot คือตลาดสากลในการซื้อขายทองคำทั่วโลก เป็นตลาดที่มี Volume สูงมากเพราะเป็นการซื้อขายทองคำจากทั่วโลก มักเรียกกันในอีกชื่อหนึ่งว่า “การเทรดทองคำในตลาดโลก” โดยจะเป็นการซื้อขายในรูปแบบสัญญาหรือใบรับประกัน ไม่ได้มีการจัดส่งทองคำแท่งให้ผู้ซื้อ

**2.1.1 ปัจจัยที่ส่งผลกระทบต่อราคาทองคำ**

ในปัจจุบัน ราคาทองคำมีความผันผวนค่อนข้างต่ำในระยะยาวจึงเป็นสินทรัพย์ที่ปลอดภัยถึงแม้ว่าทองคำจะเป็นสินทรัพย์ที่ปลอดภัย แต่ก็จะมีความเสี่ยงและมีปัจจัยที่ต้องพิจารณาด้วย เช่นปัจจัยที่จะมีผลต่อทิศทางการเคลื่อนไหวของราคาทองคำและส่งผลต่อกำไรที่นักลงทุนจะได้รับ ในปัจจุบันทองคำยังได้รับความนิยมอยู่ เนื่องจากเป็นสินทรัพย์ที่มีความสามารถในการป้องกันความเสี่ยงในรูปแบบต่าง ๆได้ เช่น ความเสี่ยงจากภาวะเงินเฟ้อ ความผันผวนของอัตราแลกเปลี่ยนเงินตราต่างประเทศ ภาวะเศรษฐกิจหดตัว ไปจนถึงการเปลี่ยนแปลงทางการเมือง เพราะทองคำเป็นสิ่งที่มีคุณค่าในตัวเองอยู่ตลอดเวลา จึงทำให้การลงทุนในทองคำสามารถกระจายความเสี่ยงของพอร์ตการลงทุน และยังนำไปใช้สร้างผลกำไรหากจับจังหวะซื้อขายได้ถูกทางอย่างไรก็ตาม ทองคำก็มีความเสี่ยงจึงต้องทำการวิเคราะห์ถึงปัจจัยที่มีผลต่อทิศทางราคาทองคำ โดยหลัก ๆ แล้ว ควรพิจารณาปัจจัยใน 3 สัญญาณ ได้แก่ สัญญาณระยาว สัญญาณระกลาง และสัญญาณระสั้น และอีกหนึ่งอย่างที่สำคัญคือข่าวสาร

**2.1.1.1 สัญญาณระยะยาว**

นักลงทุนควรพิจารณาราคาทองคำย้อนหลังในอดีตไปประมาณ 7-10 ปี เพื่อเห็นภาพทิศทางราคาทองคำที่แม่นยำมากขึ้น ตัวอย่างเช่น สถิติราคาทองคำ 7 ปีย้อนหลัง ตั้งแต่ปี 2558 จนถึงต้นปี 2564 พบว่าราคาทองคำ (Gold Spot) มีระดับต่ำสุดของแต่ละปีสูงขึ้นเรื่อย ๆ (ภาษาในตลาดทองคำเรียกว่า การยกฐานราคาในระดับต่ำสุดขึ้น) ซึ่งเหตุการณ์นี้ทำให้นักวิเคราะห์ทองคำทั่วโลกประเมินว่าทิศทางราคาทองคำยังเป็นขาขึ้นในระยะยาวค่อนข้างชัดเจน

โดยสภาทองคำโลก ได้อธิบายถึงปัจจัยที่ทำให้ทิศทางราคาทองคำจะยังคงเป็นขาขึ้นต่อไป นั่นคือ ในช่วงวิกฤติ COVID-19 ที่ผ่านมา ทองคำเป็นสินทรัพย์ที่มีราคาผันผวนน้อย ขณะเดียวกันก็ให้ผลตอบแทนค่อนข้างสม่ำเสมอ ประกอบกับนักลงทุนยังคงมองว่าการลงทุนในสินทรัพย์อื่นยังคงมีความเสี่ยงสูง ขณะที่ อัตราดอกเบี้ยทั่วโลกยังคงอยู่ในระดับต่ำและเศรษฐกิจโลกอยู่ในภาวะชะลอตัว จึงทำให้กระแสเงินลงทุนไหลเข้าสู่ตลาดทองคำอย่างต่อเนื่อง โดยเฉพาะความต้องการจากผู้บริโภคชาวจีนแลคำะอินเดีย ซึ่งล้วนแล้วแต่เป็นปัจจัยบวกต่อทิศทางราคาทองคำทั้งสิ้น

สำหรับนักลงทุนที่กำลังตัดสินใจลงทุนหรือว่าถือทองคำอยู่แล้วและเน้นลงทุนระยะยาว ยังสามารถลงทุนและถือต่อไปได้ โดยพฤติกรรมการลงทุนทองคำในระยะยาวจะลงทุนตั้งแต่ 6 เดือนขึ้นไป (มากกว่า 2 ไตรมาส) หรืออาจถือข้ามปี ซึ่งกลยุทธ์ในการลงทุน ก็คือ **ทยอยซื้อในช่วงต้นปีหรือช่วงตรุษจีน จากนั้นให้ถือและรอจังหวะทยอยขายในช่วงปลายไตรมาส 3 หรือ ก่อนสิ้นปี** นอกจากนี้ ยังมีกลุ่มนักลงทุนทองคำที่ทยอยลงทุนไปเรื่อย ๆ และถือเป็นระยะเวลาหลายปีหรือสะสมเพื่อเป็นมรดก เพราะเชื่อว่าการถือทองคำเกิน 10 ปี จะมีแต่กำไร

**2.1.1.2 สัญญาณระยะปานกลาง**

นักลงทุนควรพิจารณาราคาทองคำเป็นรายไตรมาส โดยสถิติในช่วง 3 ปีที่ผ่านมา พบว่าราคาทองคำมักปรับขึ้นสู่ระดับสูงสุดของปี ในช่วงไตรมาส 3 และราคาจะปรับลดลงเมื่อเข้าสู่ไตรมาส 4 และไตรมาส 1 ของปีถัดไป เช่น ราคาปรับขึ้นไปที่ระดับ 2,075 เหรียญสหรัฐต่อออนซ์ในช่วงเดือนสิงหาคม ปี 2563 หลังจากนั้นราคาเริ่มอ่อนตัวลง และล่าสุดไตรมาส 1 ปี 2564 ราคาทองคำอ่อนตัวลงสู่ระดับ 1,767 เหรียญสหรัฐต่อออนซ์ และปรับลดลงสู่ระดับต่ำสุดที่บริเวณ 1,676 เหรียญสหรัฐต่อออนซ์ และเมื่อเข้าสู่ไตรมาส 2 ราคาก็จะมีสัญญาณค่อย ๆ ปรับขึ้น โดยประเมินว่าในไตรมาส 3 ปีนี้ ราคาทองคำก็จะปรับขึ้นไปทำระดับสูงสุดของปีเหมือนภาพในอดีต

หากมองจากปัจจัยพื้นฐานจะพบว่า ช่วงต้นปีนักลงทุนเริ่มคาดการณ์ว่าเศรษฐกิจโลกจะฟื้นตัวอย่างชัดเจนหลังจากวัคซีน COVID-19 มีประสิทธิภาพ จึงเห็นการปรับประมาณการการเติบโตทางเศรษฐกิจ ถือเป็นปัจจัยกดดันให้ราคาทองคำ ซึ่งเป็นสินทรัพย์ปลอดภัย (Safe Haven) อ่อนตัวลง

อย่างไรก็ตาม หลังจากการแพร่ระบาด COVID-19 รอบล่าสุด นักลงทุนประเมินว่าเศรษฐกิจโลกในปีนี้จะฟื้นตัวในลักษณะค่อยเป็นค่อยไป ประกอบกับ มาตรการผ่อนคลายทางการเงินของธนาคารกลางต่าง ๆ ทั่วโลก จะไม่เปลี่ยนแปลงในช่วงครึ่งปีหลัง จึงเริ่มเห็นการเปลี่ยนทิศทางของราคาทองคำที่มีสัญญาณฟื้นตัวขึ้น จากแนวโน้มที่จะยังคงมีเม็ดเงินถูกอัดฉีดเข้ามาเพื่อกระตุ้นเศรษฐกิจ ส่งผลผลักดันให้ราคาทองคำปรับตัวขึ้น

สำหรับนักลงทุนทองคำในระยะปานกลางจะเน้นลงทุนเป็นรายเดือนและไม่เกิน 3 เดือน (1 ไตรมาส) โดยพยายามจับจังหวะการแกว่งตัวของราคาทองคำเพื่อหาจังหวะซื้อและขายเพื่อทำกำไร ซึ่งกลยุทธ์ในการลงทุน ก็คือ **รอจังหวะลงทุนเมื่อเห็นราคาทองคำอ่อนตัวลง และรอขายทำกำไรเมื่อราคาเริ่มปรับขึ้น**

**2.1.1.3 สัญญาณระยะสั้น**

นักลงทุนจะวิเคราะห์ราคาทองคำเป็นรายวันด้วยการพิจารณาราคาสินทรัพย์อื่น ๆ ประกอบ เพื่อดูทองคำกับสินทรัพย์อื่น ๆ ว่ามีความเคลื่อนไหวด้านราคาอย่างไร โดยจะพิจารณาใน 2 ปัจจัย ได้แก่ ปัจจัยที่ส่งผลไปในทิศทางตรงข้ามกับราคาทองคำ และปัจจัยที่ส่งผลไปในทิศทางเดียวกันกับราคาทองคำ เช่น ตลาดหุ้นสหรัฐอเมริกา อัตราผลตอบแทนพันธ์บัตรรัฐบาล และค่าเงินสกุลดอลลาร์สหรัฐ

**2.1.1.3.1** ตลาดหุ้นสหรัฐอเมริกา

หากตลาดหุ้นสหรัฐฯ ปรับขึ้น ราคาทองคำมีแนวโน้มปรับตัวลดลง เนื่องจากตลาดหุ้นเป็นสินทรัพย์เสี่ยง ขณะที่ทองคำเป็นสินทรัพย์ปลอดภัย ราคาจึงเคลื่อนไหวในทิศทางตรงกันข้าม

**2.1.1.3.2 อัตราผลตอบแทนพันธบัตรรัฐบาล**

จะพิจารณาจากอัตราผลตอบแทนพันธบัตรรัฐบาลสหรัฐอเมริกาอายุ 10 ปี ซึ่งถือเป็นปัจจัยสำคัญที่สะท้อนภาวะเศรษฐกิจและทิศทางอัตราดอกเบี้ยในตลาดเงินและตลาดทุนของสหรัฐฯ หากอัตราผลตอบแทนพันธบัตรรัฐบาลสหรัฐอเมริกาอายุ 10 ปีปรับขึ้น ส่วนใหญ่ค่าเงินดอลลาร์จะแข็งค่าขึ้นตามไปด้วย รวมทั้งนักลงทุนประเมินว่าเศรษฐกิจสหรัฐอเมริกาจะเติบโตและอัตราดอกเบี้ยมีแนวโน้มเป็นขาขึ้น ก็จะส่งผลให้ราคาทองคำปรับลดลง เนื่องจากทองคำไม่มีผลตอบแทนอยู่ในรูปของดอกเบี้ย ดังนั้น เมื่ออัตราดอกเบี้ยปรับตัวขึ้น การลงทุนในทองคำจึงถูกลดความน่าสนใจลง

**2.1.1.3.3 ค่าเงินสกุลดอลลาร์สหรัฐ**

จะมีความสัมพันธ์ในเชิงลบกับราคาทองคำโลก กล่าวคือ ถ้าค่าเงินดอลลาร์สหรัฐอ่อนลงเมื่อเทียบกับเงินสกุลสำคัญของโลก เช่น เงินยูโร เงินเยน หรือพิจารณาจาก US Dollar Index ก็ได้เช่นกัน ราคาทองคำโลกจะสูงขึ้น เพราะราคาทองคำซื้อขายเป็นสกุลเงินดอลลาร์สหรัฐ เมื่อค่าเงินดอลลาร์อ่อนลง ทองคำจะมีราคาถูกลงเมื่อเทียบกับเงินสกุลอื่นที่นักลงทุนถือไว้ จึงสร้างแรงซื้อเข้ามาดันให้ราคาทองคำปรับตัวเพิ่มสูงขึ้น

**2.1.1.4 ข่าวสาร**

ข่าวสารเป็นสิ่งสำคัญอีกอย่างที่ราคาทองคำจะมีความผันผวนเนื่องจากข่าวสารจะมีทั้งข่าวดีที่ส่งผลดีกับทองคำแล้ว แต่ก็ยังมีข่าวสารที่ไม่ดีส่งผลต่อทองคำในทิศทางลงส่วนใหญ่ข่าวสารที่กระทบถึงทองคำ เช่น ข่าวสารเกี่ยวกับน้ำมัน ข่าวสารเกี่ยวกับเศษฐกิจ อัตราการว่างงาน หรือโรงงานต่าง ภาวะเงินเฟ้อ หรือ เงินฝืด และข่าวสารที่สำคัญมากสำหรับทองคำจะเป็นข่าวสารจากธนาคารโลก หรือ ธนาคารกลางต่าง ๆ

|  |  |  |
| --- | --- | --- |
| ข่าวสาร | ข่าวประเภท | อ้างอิงค์ |
| 1. Census Bureau | กระทรวงพาณิชย์สหรัฐ | [7] |
| 1. Us Department of labo | กระทรวงแรงงาน | [8] |
| 1. Energy information Administration | ข้อมูล[ด้านพลังงาน](https://en.wikipedia.org/wiki/Energy) | [9] |
| 1. กองทุน SPDR | การซื้อขายทองคำของกองทุน SPDR | [10] |
| 1. Federal Reserve | ธนาคารกลางสหรัฐ | [11] |
| 1. Bloomberg | รายงานข่าวทั่วไปรอบโลก | [12] |
| 1. Twitter | ข่าวทั่วไป | [13] |

รูปที่ 1 แหล่งข่าวสาร

ในปัจจุบัน เทคโนโลยี Machine Learning NLP เป็นที่นิยมในการเอามาทำ AI เช่น การดึงข้อมูล การแปลภาษา การทำให้ข้อความง่ายขึ้น การวิเคราะห์ให้ความรู้สึกของผู้ใช้ การสรุปข้อความ ตัวกรองสแปม คาดการณ์ผลการค้นหาของผู้ใช้ แก้ไขขำผิดอัตโนมัติ เป็นต้น Natural Language Processing(NLP) for Machine Learning หรือการประมวลผลภาษาธรรมชาติด้วย Python ซึงภาษา Python เป็นภาษาที่รวดเร็วและในการทำ NLP จะใช้ **Natural Language Toolkit** (NLTK) เป็น Library Opensource ยอดนิยมใน Python ซึ่งมีรายละเอียดดังนี้

* 1. **Natural Language Processing NLP**

การประมวลผลภาษาธรรมชาติ (NLP) คือความสามารถของโปรแกรมคอมพิวเตอร์ในการทำความเข้าใจภาษามนุษย์ในขณะที่พูดและเขียน ซึ่งเรียกว่าภาษาธรรมชาติ เป็นส่วนประกอบของปัญญาประดิษฐ์

### **2.2.1 How does natural language processing work**

NLP ช่วยให้คอมพิวเตอร์เข้าใจภาษาธรรมชาติเหมือนที่มนุษย์เข้าใจ ไม่ว่าจะเป็นภาษาพูดหรือเขียน การประมวลผลภาษาธรรมชาติใช้ปัญญาประดิษฐ์เพื่อป้อนข้อมูลในโลกแห่งความเป็นจริง ประมวลผล และทำความเข้าใจในลักษณะที่คอมพิวเตอร์สามารถเข้าใจได้ เช่นเดียวกับที่มนุษย์มีเซ็นเซอร์ที่แตกต่างกัน เช่น หูสำหรับได้ยินและตาที่มองเห็น คอมพิวเตอร์มีโปรแกรมสำหรับอ่านและไมโครโฟนเพื่อรวบรวมเสียง และเช่นเดียวกับที่มนุษย์มีสมองในการประมวลผลอินพุตนั้น คอมพิวเตอร์ก็มีโปรแกรมสำหรับประมวลผลอินพุตที่เกี่ยวข้อง ในบางจุดของการประมวลผล อินพุตจะถูกแปลงเป็นโค้ดที่คอมพิวเตอร์สามารถเข้าใจได้  
  
มีสองขั้นตอนหลักในการประมวลผลภาษาธรรมชาติ: การประมวลผล[ข้อมูลล่วงหน้า](https://searchsqlserver.techtarget.com/definition/data-preprocessing)และการพัฒนาอัลกอริธึม

การประมวลผลข้อมูลล่วงหน้าเกี่ยวข้องกับการเตรียมและ "การล้าง" ข้อมูลข้อความสำหรับเครื่องเพื่อให้สามารถวิเคราะห์ได้ การประมวลผลล่วงหน้าทำให้ข้อมูลอยู่ในรูปแบบที่ใช้การได้และเน้นคุณลักษณะในข้อความที่อัลกอริธึมสามารถทำงานได้ สามารถทำได้หลายวิธี ได้แก่:

[**Tokenization**](https://www.techtarget.com/searchsecurity/definition/tokenization) คือเมื่อข้อความถูกแบ่งออกเป็นหน่วยย่อยเพื่อใช้งาน

[**Stop word**](https://www.techtarget.com/whatis/definition/stop-word) คือเมื่อคำทั่วไปถูกลบออกจากข้อความเพื่อให้คำที่ไม่ซ้ำกันซึ่งให้ข้อมูลส่วนใหญ่เกี่ยวกับข้อความยังคงอยู่

[**Lemmatization**](https://www.techtarget.com/searchenterpriseai/definition/lemmatization)**and stemming** คือเวลาที่คำถูกลดขนาดให้อยู่ในรูปแบบรูทเพื่อประมวลผล

**Part-of-speech tagging** อการทำเครื่องหมายคำตามส่วนของคำพูด เช่น คำนาม กริยา และคำคุณศัพท์

เมื่อข้อมูลได้รับการประมวลผลล่วงหน้า อัลกอริทึมจะได้รับการพัฒนาเพื่อประมวลผล มีอัลกอริธึมการประมวลผลภาษาธรรมชาติที่แตกต่างกันมากมาย แต่โดยทั่วไปจะใช้สองประเภทหลัก:

**Rules-based system.** ระบบนี้ใช้กฎทางภาษาที่ออกแบบมาอย่างดี แนวทางนี้ใช้ในช่วงเริ่มต้นในการพัฒนาการประมวลผลภาษาธรรมชาติ และยังคงใช้อยู่

**Machine learning-based system.** อัลกอริธึมการเรียนรู้ของเครื่องใช้วิธีทางสถิติ พวกเขาเรียนรู้ที่จะดำเนินการตามข้อมูลการฝึกอบรมที่ได้รับ และปรับวิธีการของตนเมื่อมีการประมวลผลข้อมูลมากขึ้น อัลกอริธึมการประมวลผลภาษาธรรมชาติใช้การผสมผสานระหว่างแมชชีนเลิร์นนิง การเรียนรู้เชิงลึกและ[โครงข่ายประสาทเทียม ผ่านการประมวลผลและการเรียนรู้ซ้ำๆ](https://www.techtarget.com/searchenterpriseai/definition/neural-network)

### **2.2.2 เหตุใดการประมวลผลภาษาธรรมชาติจึงมีความสำคัญ**

ธุรกิจใช้ข้อมูลที่ไม่มีโครงสร้างและมีข้อความจำนวนมากและต้องการวิธีในการประมวลผลอย่างมีประสิทธิภาพ ข้อมูลจำนวนมากที่สร้างขึ้นทางออนไลน์และจัดเก็บไว้ในฐานข้อมูลเป็นภาษามนุษย์ตามธรรมชาติ และจนกระทั่งเมื่อไม่นานมานี้ ธุรกิจต่างๆ ก็ไม่สามารถวิเคราะห์ข้อมูลนี้ได้อย่างมีประสิทธิภาพ นี่คือจุดที่การประมวลผลภาษาธรรมชาติมีประโยชน์

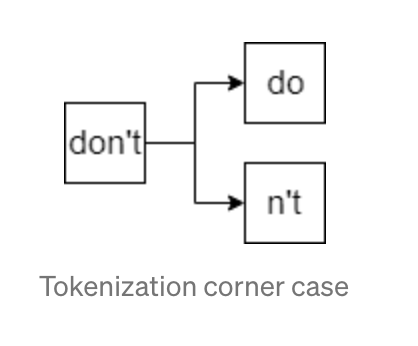
ประโยชน์ของการประมวลผลภาษาธรรมชาติสามารถเห็นได้เมื่อพิจารณาจากสองข้อความต่อไปนี้: "การประกันการประมวลผลแบบคลาวด์ควรเป็นส่วนหนึ่งของข้อตกลงระดับบริการทุกฉบับ" และ " [SLA](https://www.techtarget.com/searchitchannel/definition/service-level-agreement) ที่ดี ช่วยให้นอนหลับสบายตลอดคืน แม้จะอยู่ในระบบคลาวด์" หากผู้ใช้อาศัยการประมวลผลภาษาธรรมชาติในการค้นหา โปรแกรมจะรับรู้ว่าการประมวลผลแบบคลาวด์เป็นเอนทิตีคลาวด์นั้นเป็นรูปแบบย่อของคลาวด์คอมพิวติ้ง และSLAนั้นเป็นคำย่อของอุตสาหกรรมสำหรับข้อตกลงระดับบริการ

เหล่านี้เป็นประเภทขององค์ประกอบที่คลุมเครือซึ่งมักปรากฏในภาษามนุษย์และอัลกอริธึมการเรียนรู้ด้วยเครื่องมีการตีความที่ไม่ดีในอดีต ด้วยการปรับปรุงวิธีการเรียนรู้เชิงลึกและแมชชีนเลิร์นนิง อัลกอริธึมสามารถตีความได้อย่างมีประสิทธิภาพ การปรับปรุงเหล่านี้ช่วยขยายความกว้างและความลึกของข้อมูลที่สามารถวิเคราะห์ได้

### **2.2.3 เทคนิคและวิธีการประมวลผลภาษาธรรมชาติ**

ทั่วไปสำหรับข้อความก่อนการประมวลผลประกอบด้วย 5 อย่าง

2.2.3.1 Sentence segmentation ในขั้นตอนแรกของการเตรียมประโยคข้อความจะถูกแบ่งออกเป็นประโยคข้อความที่ใช้ เช่น ภาษาอังกฤษ เครื่องหมายวรรคตอน โดยเฉพาะอักขระหยุด เครืองหมายอัศเจรีย์และเครื่องหมายคำถามสามารถใช้ระบุจุดสิ้นสุดของประโยคได้อย่างไรก็ตาม อักขระจุดยังสามารถมใช้เป็นตัวย่อของข้อความได้ เช่น Ms. หรือ UK ซึ่งในกรณีนี้ อักขระหยุดไม่ได้หมายถึงจุดสิ้นสุดของประโยค ในกรณีเหล่านี้ใช้อักษรย่อเพื่อหลีกเลี่ยงการแบ่งประเภทขอบเขตประโยคที่ไม่ถูกต้อง เมื่อข้อความมีคำศัพท์เฉพาะ เราจะต้องสร้างพจนานุกรมคำย่อเพิ่มเติมเพื่อหลีกเลี่ยงการทำเครื่องหมายผิดหลักธรรมชาติ ตัวอย่างการทำให้เป็นมาตรฐาน



รูปที่ 2 เป็นการทำ Tokenization [6]

2.2.3.2 Tokenization คือ การแบ่งข้อความออกเป็นคำและเครื่องหมายวรรคตอนที่เป็นเครื่องหมายเช่น เดียวกับการแบ่งประโยคเครื่องหมายวรรคตอน ตัวอย่างเช่น U.K. ควรจะเป็นเครื่องหมายและ don’t ควรแบ่งออกเป็นสองเครื่องหมาย do และ n’t

2.2.3.3 Stemming และ lemmatization เป็นส่วนสำคัญของกระบวนการทำให้เป็นมามาตรฐาน การทำให้เป็นมาตรฐานประกอบด้วยการสกัดคำที่ต้องระบุต้นคำโดนการลบต่อท้าย เช่น -ed และ -ing ไม่จำเป็นต้องเป็นคำ ในทำนองเดียวกัน lemmatization เกี่ยวข้องกับการลบคำนำหน้าและส่วนต่อท้ายความแตกต่างที่สำคัญคือผลที่ได้คือภาษา ผลที่ได้นี้เรียกว่าการอ้างอิง ตัวอย่างของ Stemming และ lemmatization ดังรูปที่ 2

รูปภาพประกอบด้วย โต๊ะ

คำอธิบายที่สร้างโดยอัตโนมัติ

รูปที่ 3 รูปแบบของการใช้งาน Stemming และ lemmatization [6]

ทั้ง 2 เทคนิคที่กล่าวมาจะช่วยในการลดสัญญาณรบกวนในข้อความโดยแปลงคำให้อยู่ในรูปแบบพื้นฐาน เช่น การประเภทข้อความหรือการจัดกลุ่มเอกสาร ซึ่งการรักษาความหมายของคำเป็นสิ่งสำคัญ ควรใช้ lemmatization มากกว่าการวิเคราะห์ ตัวอย่างเช่น คำนามและคำกิริยา ซึ่งทำให้สูญเสียความหมายดั้งเดิมไป เทคนิคการทำให้เป็นมาตรฐานอื่น ๆ ได้แก่ การขยายคำย่อ การลบตัวเลขและเครื่องหมายวรรคตอน การแก้ไขคำผิดพลาดทางไวยากรณ์ การดำเนินการเหล่านี้ส่วนใหญ่สามารถทำได้โดยใช้นิพจน์ทั่วไป

2.2.3.4 Part of speech tagging ขั้นตอนนี้จะเป็นการแบ่งเครื่องหมายเป็น part of speech (POS) หรือที่เรียกว่าคำศัพท์ หรือ หมวดหมู่คำศัพท์ คำที่ประกอบด้วยคำนาม,คำกิริยา,คำบุพบท,คำกิริยาวิเศษณ์ ดังตารางต่อไปนี้จะแสดงคำและตัวอย่าง ส่วยสัญลักษณ์ จะใช้ lemmatization ซึ่งเป็นสิ่งจำเป็นสำหรับการตั้งชื่อ บุคคล

รูปภาพประกอบด้วย โต๊ะ

คำอธิบายที่สร้างโดยอัตโนมัติ

รูปที่ 4 ตัวอย่างการติดแทรกจัดกลุ่มข้อความ POS-taggers [6]

POS-taggers มี 3 ประเภท ได้แก่ ตามกฎสถิติและตามการเรียนรู้เชิงลึก กฎตามเครื่องหมายขึ้นอยู่กับว่ากฎที่ชัดเจนเพื่อทำเครื่องหมาย เช่น บทความต้องตามด้วยคำนาม เพื่อกำหนดเครื่องหมาย ตามกฎสถิติใช้แบบจำลองความน่าจะเป็นในการมาร์คแต่ละคำหรือลำดับของคำ กฎตามแท๊กตามกฎนั้นแม่นยำมาก แต่ก็ยังขึ้นอยู่กับภาษาด้วย การขยาย tagger เพื่อรองรับภาษาอื่น ๆ ตัวติดแท๊กภาษาอังกฤษนั้นสร้างได้ง่ายกว่าและไม่ขึ้นกับภาษา และมีการใช้วิธีผสมผสานของแบบจำลองตามกฎและแบบจะลองทางสถิติ โดยที่แบบจำลองจะได้รับการฝึกอบรมเกี่ยวกับชุดประโยคที่ติดแท็กล่วงหน้า วิธีการแบบไฮบริดและการเรียนรู้เชิงลึกจะสามารถปรับปรุงการติดแท็กได้ตามบริบท

2.2.3.5 Named entity Recognition คือการแบ่งกลุ่มของเครื่องหมาย การแบ่งกลุ่มหมายถึงการติดแท็ก หนึ่งในกลุ่มคำที่ใช้มากที่สุด คือกลุ่มตำนามที่ประกอบด้วยตัวกำหนด คำคุณศัพท์ และ คำนาม เช่น a happy unicorn ประโยค He found a happy unicorn ประกอบด้วยสองส่วน he และ a happy unicorn

รูปภาพประกอบด้วย ข้อความ

คำอธิบายที่สร้างโดยอัตโนมัติ

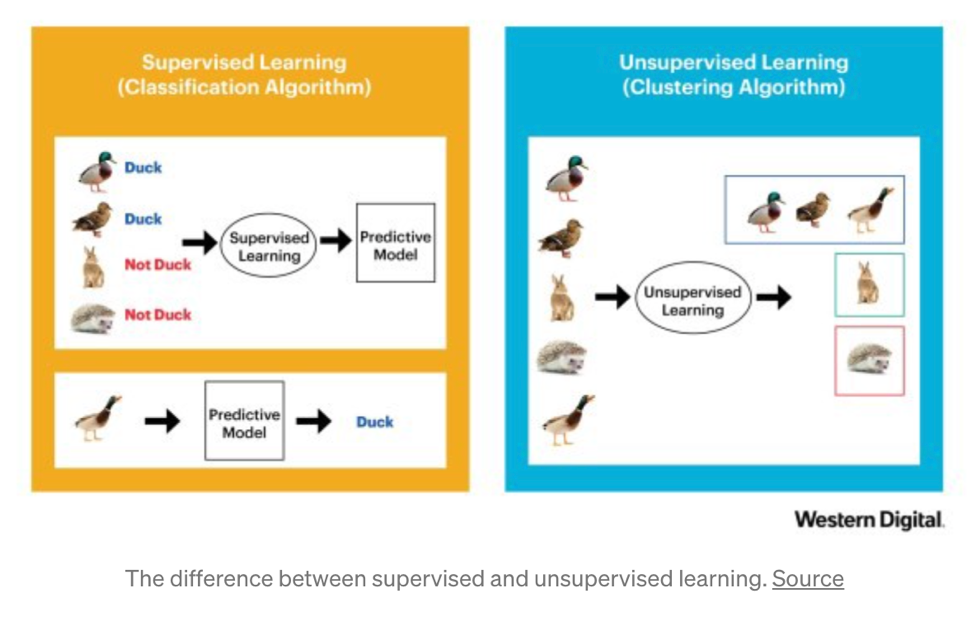
รูป 5 การติดแทรกที่อ้างอิงถึงวัตถุเฉพาะ Named entity [6]

Named entity เป็นคำนามที่อ้างอิงถึงวัตถุเฉพาะ เช่น บุคคล องค์กร สถานที่ วันที่ และหน่วยทางภูมิศาสตร์เป้าหมายของขั้นตอน Named entity Recognition คือการระบุชื่อนิติบุคคลที่กล่าวถึงในข้อความ

**2.4 Machine Learning**

As Brink และคนอื่น ๆ ให้คำจำกัดความไว้ว่า Machine Learning(ML) คือการใช้ประโยชน์จากรูปแบบของข้อมูลในอดีตเพื่อตัดสินใจเกี่ยวกับข้อมูลใหม่ หรือเป็นทาง Google หัวหน้านักวิทยาศาสตร์ด้านการตัดสินใจ หรือ Cassie Kozyrkov ชี้ให้เห็นว่า Machine Learning เป็นเพียงตัวติดฉลาก อธิบายไว้เกี่ยวกับบางสิ่งบางอย่างและบอกให้รู้ว่าควรได้รับฉลากอะไร การใช้เทคนิค ML มีประโยชน์เมื่อปัญหานั้นซับซ้อนเกินกว่าจะแก้ไขด้วยการเขียนโปรแกรม เช่น แยกแยะสายพันธุ์แมวต่าง บนรูปภาพ หรือโซลูชันจำเป็นต้องปรับเปลี่ยนเมื่อเวลาผ่านไป เช่น การจดจำข้อความที่เขียนด้วยลายมือ

โดยทั่วไปแล้ว Machine Learning จะแบ่งออกเป็น Machine Learning ที่จะต้องดูแล และ ไม่ต้องดูแล เราสามารถการเรียนรู้ภายใต้การดูแลเมื่อข้อมูลการฝึกอบรมในอดีตของเรามีป้ายกำกับ (เช่น duck และ no duck ในรูปตัวอย่างด้านล่าง) ในทางกลับกันการเรียนรู้แบบไม่มีผู้ดูแลจะถูกนำมาใช้เมื่อไม่มีป้ายกำกับในข้อมูล วิธีการเรียนรู้ของเครื่องที่ไม่ได้รับการดูแลมีเป้าหมายเพื่อสรุปหรือบีบอัดข้อมูลการฝึกอบรมพร้อมป้ายกับกับ สแปม/ไม่สแปม ในกรณีหลัง เราจะต้องตรวจหาอีเมลผิดปกติตามชุดการฝึกอบรมของอีเมล



รูปที่ 6 ความแตกต่างระหว่างการเรียนรู้แบบมีผู้ดูแล และ ไม่มีผู้ดูแล [6]

**2.5 Extracting features from text**

อัลกอริทึม ของ Machine Learning ทั้งหมดต้องการข้อมูลดิจิตอลเป็นอินพุต ซึ่งหมายความว่าข้อมูลและข้อความจะต้องถูกแปลงเป็นตัวเลข ขั้นตอนการแยกคุณลักษณะของ NLP

**2.5.1 Count-based strategies**

เป็นวิธีที่ง่ายที่สุดในการแปลงข้อความเป็นเวกเตอร์ตัวเลขคือการใช้วิธี Bag-of-Words (BoW) หลักการของ BoW คือการแยกคำที่ไม่ซ้ำกันทั้งหมดจากข้อความและสร้างคลังข้อความที่เรียกว่าคำศัพท์ การใช้คำศัพท์แต่ละประโยคสามารถแสดงเวเกตอร์ประกอบด้วย 1 และ 0 ขึ้นอยู่กับว่ามีคำศัพท์อยุ่ในประโยคหรือไม่ รูปด้านล่างแสดงตัวอย่างของเมทริกที่สร้างขึ้นโดยใช้วิธี BOW ในประโยคห้าประโยคที่ทำให้เป็นมาตรฐาน

**รูปภาพประกอบด้วย ข้อความ

คำอธิบายที่สร้างโดยอัตโนมัติ**

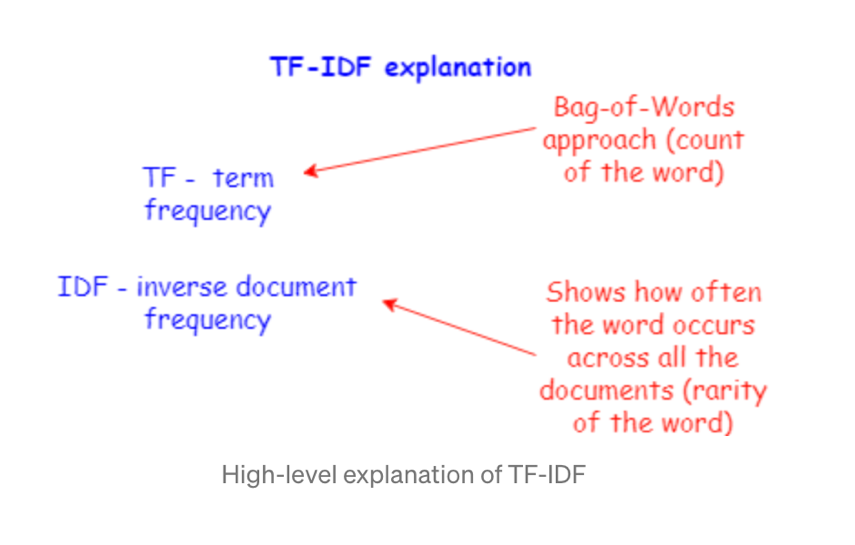
รูปที่ 7 เป็นขั้นตอนการเคลียข้อความ ตัดอักขระออก Sentence [6]

**รูปภาพประกอบด้วย โต๊ะ

คำอธิบายที่สร้างโดยอัตโนมัติ**

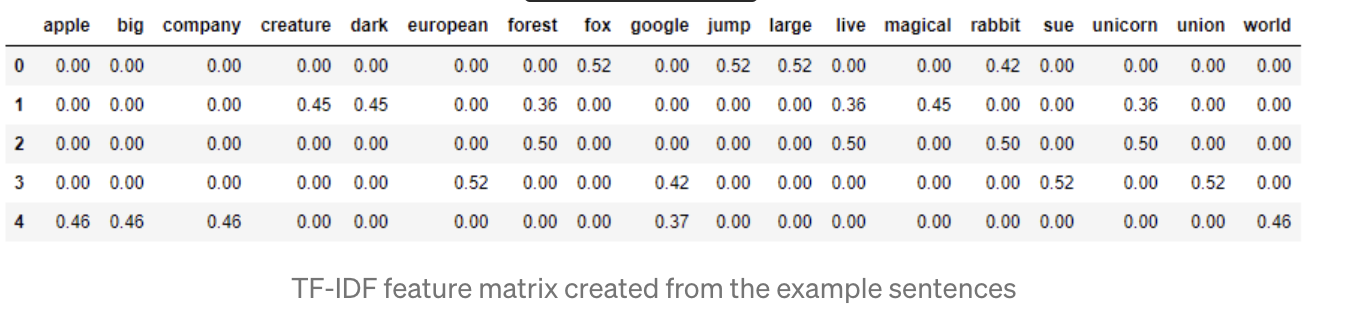
รูปที่ 8 เป็นสร้างครังคำศัพท์เพื่อแยกว่าคำศัพท์ตัวไหนมีการใช้บ่อย หลักการของ BoW [6]

จะสามารถจัดกลุ่มแท๊กเข้าด้วยกันเพื่อเพิ่มบริบทเพิ่มเติมไปยังคำศัพท์ วิธีนี้เรียกว่า N-gram วิธี N-gram คือลำดับของเครื่องหมาย N เช่น 2-gram คำลำดับของคำสองคำ ในขณะที่ trigram คือลำดับของสามเมื่อเลือกคำศัพท์แล้ว ไม่ว่าจะเป็น 1-, 2- หรือ 3- gram จะต้องนับจำนวน gram เราสามารถใช้วิธ BoW ได้ ข้อเสียของแนวทางนี้คือคำที่นิยมมีความสำคัญเกินไป ดังนั้น วิธีที่นิยมใช้กันมากที่สุดจึงเรียกว่า term frequency - inverse document frequency (TF-IDF)



รูปที่ 9 เป็นการจัดความสำคัญของคำเทียบกับความยาวประโยค [6]

TF-IDF ประกอบด้วย term frequency (TF) เป็นการจัดความสำคัญของคำเทียบกับความยาวประโยคและ inverse document frequency (IDF) ซึ่งจัดจำนวนแถวเอกสารที่ gram เกิดเมื่อเทียบกับจำนวนของแถวในเอสารเพื่อเน้นความหายากของคำ ตามที่คิดไว้ คำที่ปรากฎอยู่บ่อยในเอกสารแต่ไม่ค่อยปรากฎในเอกสารทั้งหมด คำหนึ่งจะมีคะแนน TF-IDF สูงกว่า หากพบบ่อยในเอการแต่จะไม่พบพบ่อยในชุดเอกสารทั้งหมด ดังรูปที่ 9 ตัวอย่างของเมทริกซ์ที่สร้างขึ้นโดยใช้วิธี TF-RDF ในประโยคตัวอย่างที่เห็นก่อนหน้านี้ สังเกตว่าคะแนนของคำว่า fox แตกต่างจากคะแนนที่กำหนดให้กับ rabbit

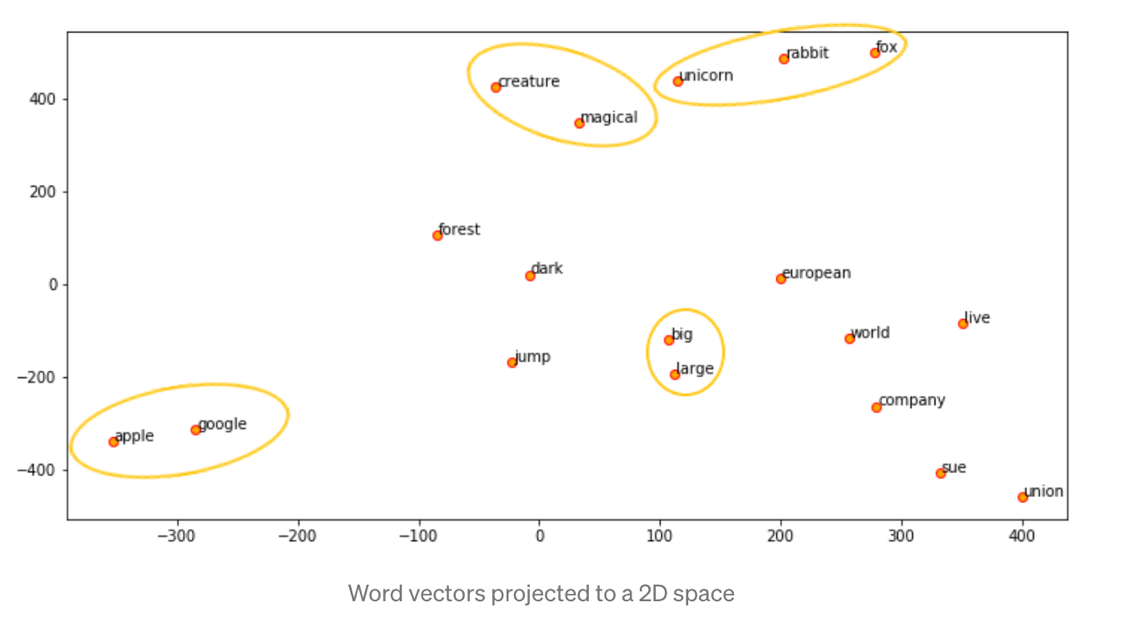


รูปที่ 10 แสดงถึงความสำคัญของแต่ละประโยค [6]

**2.5.2 Advanced strategies** วิธีการ count-based แม้สว่าสามารถใช้เพื่อจัดลำดับของคำ (N-grams ) แต่ก็ไม่ได้จัดบริบททางความหมายของคำซึ่งเป็นแกนหลักของแอปพริเคชัน NLP จำนวนมาก เทคนิคการฝั่งคำใช้เพื่อแก้ปัญหานี้ การใช้การฝั่งคำคำศัพท์จะถูกแปลงเป็นเวกเตอร์เพื่อให้คำที่มีบริบทคล้ายกันอยู่ใกล้กัน

Word2Vex เป็นเฟรมเวิร์คจาก Google ที่ใช้โครงข่ายประสามเทียมแบบตื้นเพื่อฝึกโมเดลฝั่งคำสั่ง อัลกอรธึม โดย Word2Vac มี 2 ประเภท ประเภทที่ 1 Skip-gram ซึ่งใช้เพื่อทำนายบริบทรอบ ๆ คำที่กำหนด ในขณะที่โทเดล Continuous Bag of Words ( CBOW) ใช้เพื่อทำนายคำถัดไปตามบริบทที่กำหนด

วิธีที่2 GloVe วิธี Global Vector ใช้สถิติการเกิดขึ้นร่วมเพื่อสร้างช่องว่างเวกเตอร์ วิธีนี้เป็นส่วนขยายต่อมาจาก Word2Vec ที่มีแนวโน้มว่าจะให้การฝั่งคำที่ดีกว่า

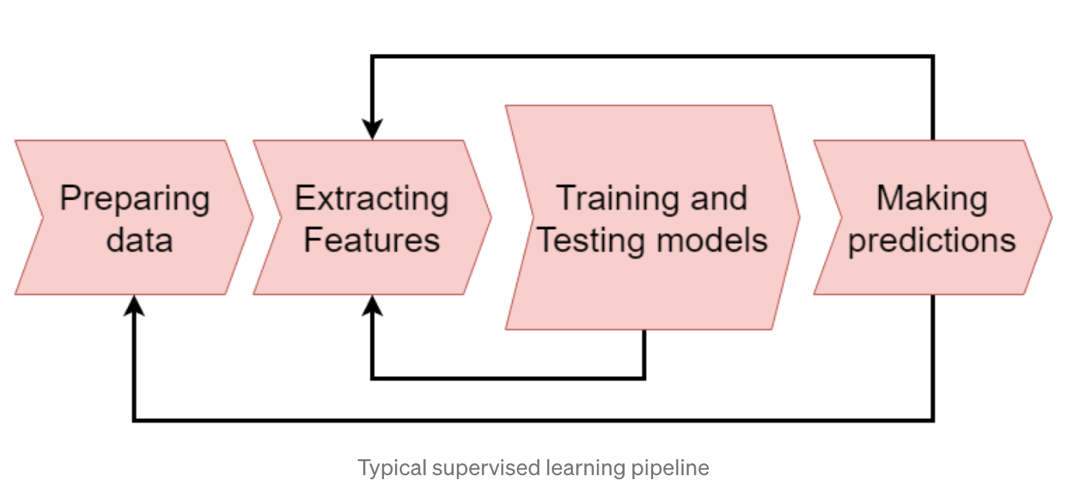


รูปที่ 11 เป็นการแสดงบริบทที่คล้ายกันจัดอยู่ในกลุ่มเดียวกัน [6]

**2.6 Supervised learning on text**

**2.6.1 Supervised learning**

การดูแลเครื่องการเรียนรู้งานแบ่งออกเป็นสองส่วน ตามรูปแบบของป้ายกำกับ (เรียกอีกอย่างว่าเป้าหมาย) หากเป้าหมายคือการจำแนกค่า (cat/dog) แสดงว่าเป็นปัญหาการจัดหมวดหมู่ ในทางกลับกัน หากเป้าหมายเป็นตัวเลข (ราคาของบ้าน) แสดงว่าปํญหาการทดถอย เมื่อต้องจัดการกับข้อความ ปัญหาที่ตามมาจะเป็นการจำแนกประเภท



รูปที่ 12 workflow supervised [6]

จากรูปด้านบนแสดง workflow ทั่วไปของระบบการจัดการประเภทข้อความ เราเริ่มต้นด้วยการแบ่งข้อมูลออกเป็นชุดฝึกอบรม และชุดทดสอบ ข้อมูลชุดฝึกและข้อมูลทดสอบต้องได้รับการประมวลผลล่วงหน้าและทำให้เป็นมาตรฐาน หลังจากนั้นจึงจะสามารถดึงคุณลักษณะออก มาได้ เทคนิคการแยกคุณลักษณะยอดนิยมสำหรับข้อมูลประเภทข้อความครอบคลุมอยู่ในส่วนก่อนหน้านี้ เมื่อข้อมูลข้อความถูกแปลเป็นรูปแบบตัวเลขแล้ว สามารถใช้อัลกอริธึมการเรียนรู้ของเครื่องได้ กระบวนการนี้เรียกว่าฝึกโมเดล โมเดลเรียนรู้รูปแบจากคุณสมบัติต่าง ๆ เพื่อนทำนายฉลาก โมเดลสามารถปรับให้เหมาะสมเพื่อประสิทธิภานพที่ดีขึ้นโดยการใช้ พารามิเตอร์โมเดลผ่านกระบวนการที่เรียกว่าการปรับแต่งไฮเปอร์พารามิเตอร์ แบบจะลองผลธ์จะถูกประเมินบนข้อมูลการทดสอบที่มองไม่เห็นก่อนหน้า ประสิทธิภาพของโมเดลวัดโดยใช้เมตริกต่าง ๆ เช่นความแม่นยำ การเรียกคืนคะแนนF1และอื่นๆ อัลกอริธึมที่ใช้สำหรับการจัดประเภทข้อความเช่น

2.6.1.1 Multinomial Naive Bayes อยู่ในตระกูลของอัลกอริทึม Naïve Bayes ซึ่งสร้างขึ้นจากการใช้ทฤษฎี ของ Bayes โดยใช้สมมติฐานที่มีป้ายกำหับต่างกันมากกว่าสองป้ายที่แตกต่างกัน

2.6.1.2 Logistic Regression เป็นอัลกอริทึมที่ใช้ฟังก์ชัน Simoid เพื่อทำนายค่าการจำแนก แพคเกจซอฟแวร์ที่ได้รับความนิยม SKLearn จะอนุญาติให้ปรับพารามิเตอร์ของโมเดลในลักษณะที่อัลกอริทึมสามารถใช้สำหรับการจำแนกประเภทหลายป้ายกำกับได้เช่นกัน

2.6.1.3 [Support Vector Machines](https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72) (SVM) อัลกอริธึมที่ใช้เส้นหรือไฮเปอร์เพลน (ในกรณีที่มีคุณสมบัติมากกว่าสองอย่าง จึงสร้างพื้นที่หลายมิติ) เพื่อแยกคลาส

2.6.1.4 Random Forest การบูรณาการวิธีการฝึกต้นไม้ตัดสินใจหลายขนานในเซตย่อยข้อมูลที่แตกต่างกัน

2.6.1.5 Gradient Boosting Machine (GBM) ชุดของวิธีการแบบบูรณาการที่ใช้ในการฝึกชุดของผู้เรียนที่อ่อนแอเช่นต้นไม้การตัดสินใจที่จะได้รับผลลัพธ์ที่ถูกต้อง XGboost เป็นหนึ่งในการใช้งานที่นิยมมากที่สุดของชุดนี้

Random Forest และ Random Forest เป็นอัลกอริทึมการจัดหมวดหมู่เป็นวิธีการแบบบูรณาการซึ่งใช้ขั้นตอนวิธีการพยากรณ์หลายเพื่อให้บรรลุทั่วไปดีกว่า ผลของวิธีการตั้งค่ามักจะเฉลี่ยมากกว่ารุ่นเดียวและมีประสิทธิภาพมากขึ้นในชุดข้อมูลที่มีขนาดใหญ่กว่า อย่างไรก็ตามเป็น sarkar ได้พิสูจน์แล้วว่าใน วิธีการบูรณาการไม่จําเป็นต้องจัดการกับข้อมูลข้อความดีกว่า

* 1. **Naive Bayes Classification**

เป็นการจัดหมวดหมู่โดยใช้หลักความน่าจะเป็นเข้ามาช่วยคำนวณ

**รูปภาพประกอบด้วย ข้อความ

คำอธิบายที่สร้างโดยอัตโนมัติ**

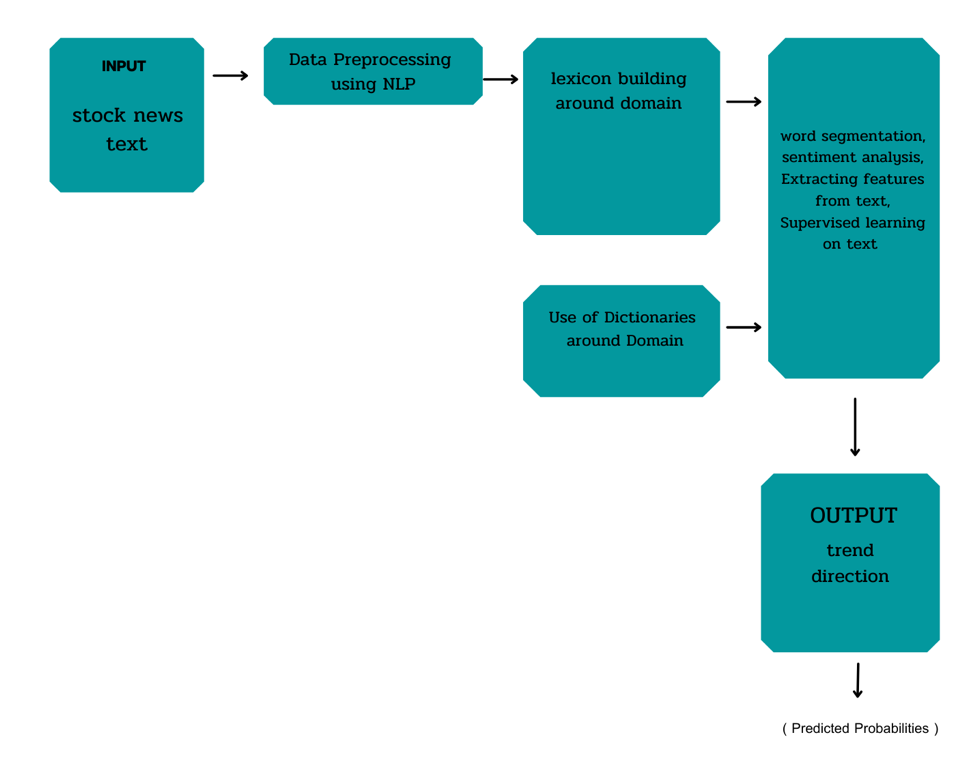
อธิบายสมการ แทนตัวแปร 3 ตัว คือ  
c คือ Class  
x คือ Attribute  
P คือ Probability (ความน่าจะเป็น)

P(c|x) **Posterior probability คือ**ความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น x จะมีคลาส C  
P(x|c) **Likelihood**คือ ความน่าจะเป็นที่ข้อมูลที่มีคลาส C และมีแอตทริบิวต์ x  
P(c) **Prior** **probability**คือ จำนวน Classที่อาจจะเกิดขึ้น/จำนวนClass ทั้งหมด  
หรือความน่าจะเป็นของ Class C  
P(x) **Predictor Prior probability**คือ จำนวน Attribute ทั้งหมด

**บทที่ 3**

**รายละเอียดการดำเนินงาน**

1. **ขั้นตอนและวิธีการดำเนินงาน**
   1. **ขั้นตอนการทำงานของระบบ**



รูปที่ 13 เป็นการแสดงขั้นตอนการทำงานของ NLP

* + 1. **ขั้นตอน INPUT (แหล่งข้อมูลข่าวสาร)**
       1. ดังรูปที่ 1
    2. **ขั้นตอน Processing**
       1. การแยกตัวย่อและคำจำกัดความ
       2. แยกหน่วยงาน (เช่น คน บริษัท ผลิตภัณฑ์ จำนวนเงิน สถานที่ ฯลฯ)
       3. ดึงข้อมูลอ้างอิงไปยังเอกสารอื่น ๆ
       4. การแยกอารมณ์ความรู้สึก (ข่าวเชิงบวก/เชิงลบและการอ้างอิง)
       5. ดึงคำพูดจากบุคคลที่มีการอ้างอิงถึงผู้เขียน
       6. สกัดเงื่อนไขสัญญา
       7. เลือกอัลกอริทึมที่จะมาใช้ในการทำนาย
       8. เทรนนิ่ง
    3. **ขั้นตอน Output**

9.3.1.1 สรุปข้อความ

9.3.1.2 คาดการ์ณล่วงหน้า

9.3.1.3 แสดงทิศทางแนวโน้มราคาขึ้นหรือลง

9.3.1.3 แสดงขึ้นบนเว็บไซต์

* 1. **ขั้นตอนการดำเนินงาน**
     1. พัฒนาอัลกอริทึม
     2. พัฒนาเขียนโปรแกรมด้วย Python
     3. เทรนสมองกล
     4. ทดสอบระบบ
     5. ทดลองนำไปใช้จริง กับบัญชี Forex demo
     6. ปรับปรุงแก้ไขอุปกรณ์ให้สมบูรณ์
     7. จัดทำเว็บไซต์
     8. จัดทำรายงานให้สมบูรณ์

**บทที่ 4**

**ความก้าวหน้าการดำเนินงาน**

* 1. **ความก้าวหน้า 1 ศึกษา Library Python NLP และทดสอบ**
     1. Library Python NLP

# NLTK Natural Language Toolkit

อินเทอร์เฟซที่ใช้งาน

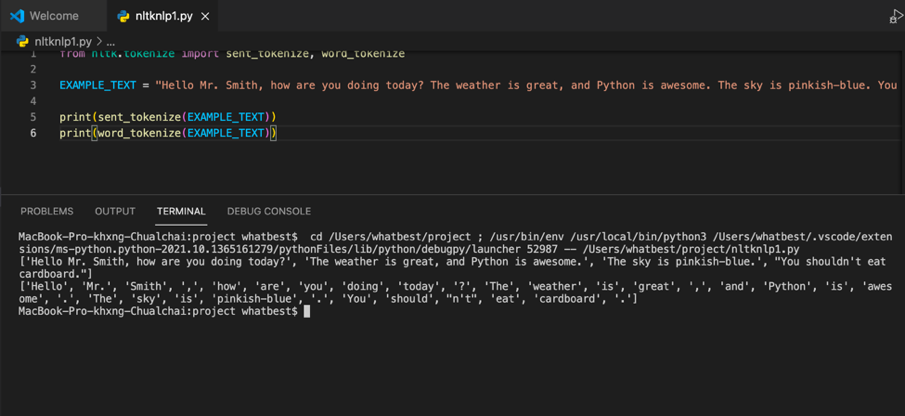
1. Tokenisation
2. Stemming
3. Tagging
4. Parsing
5. Semantic reasoning
6. Wrappers for industrial-strength NLP libraries

WordNet

# TextBlob ศึกเป็น Library ของภาษา Python

# อินเทอร์เฟซที่ใช้งาน

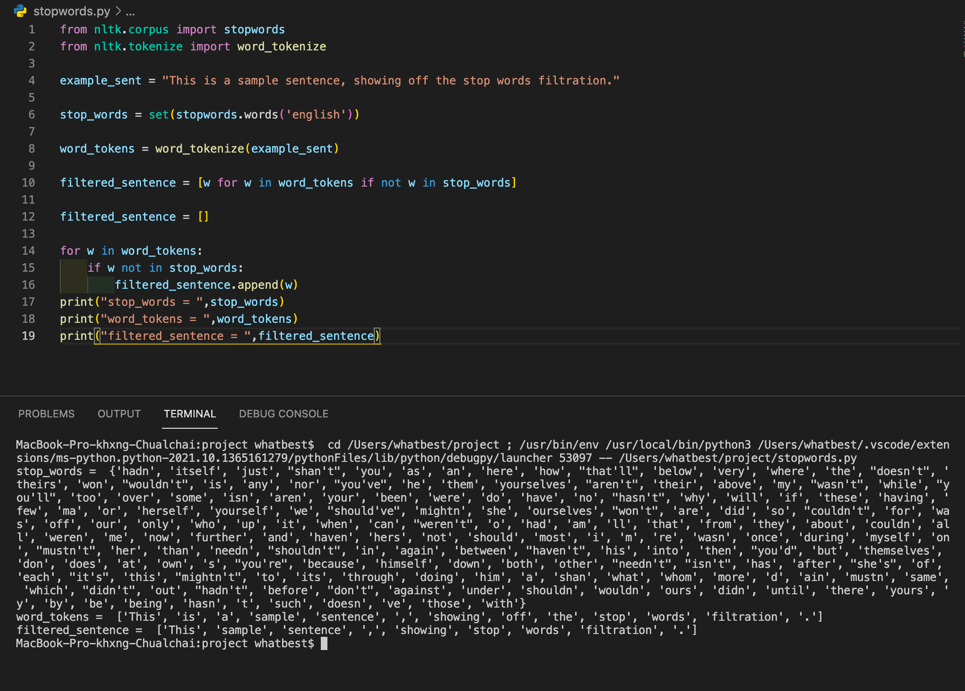
1. Tagging
2. noun phrase extraction
3. sentiment analysis
4. classification
5. language translation
6. word inflection, parsing
7. n-grams
8. WordNet integration.
   * 1. รายละเอียดการทดลอง
        1. Tokenizing Words and Sentences with NLTK



Sentences \_Tokenizing – การแยกประโยคและคำออกจากเนื้อความของข้อความ

Words\_ Tokenizing เป็นการแยกคำออกมาจากประโยค

* + - 1. Stop words with NLTK



เป็นการใช้ คำหยุด โดยอ้างอิงค์จากคลังข้อมูลของภาษาอังกฤษ คำหยุด คือ คำที่เป็นคำวลีหรือเป็นคำที่เสริมทำให้ประโยคดีขึ้น แต่ในเชิงคอมพิวเตอร์ คำหยุด คือคำที่ไม่มีความหมาย เหตุผลที่ต้องลบคำหยุดเนื่องจากประโยคที่ยาวขึ้นจะกินทรัพยากรและเวลาในการประมวลผล

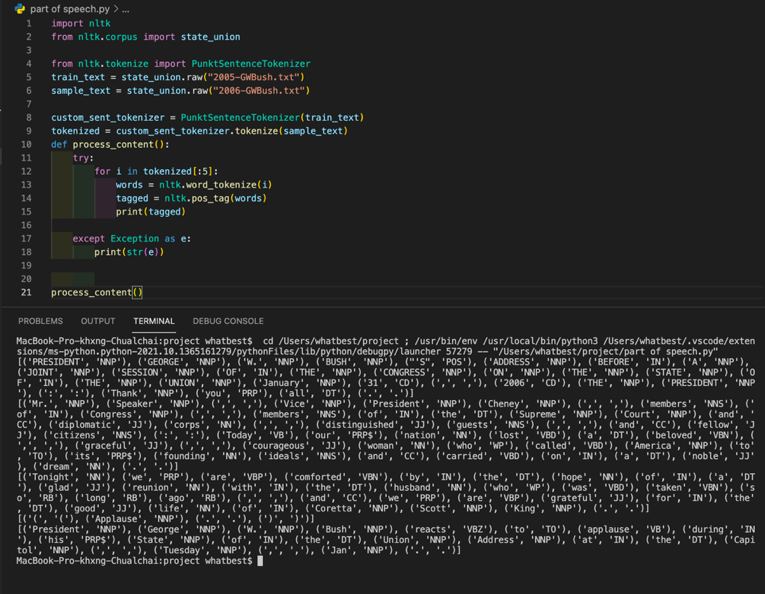
* 1. **ความก้าวหน้า 2 ทดสอบ Library NLTK** 
     1. รายละเอียดการทดลอง
        1. Stemming words with NLTK

รูปภาพประกอบด้วย ข้อความ

คำอธิบายที่สร้างโดยอัตโนมัติ

Stemming หรือที่เรียกว่า suffix stripping เป็นเทคนิคที่ใช้ในการลดขนาดข้อความ **ต้นกำเนิดยังเป็นประเภทของข้อความที่ทำให้เป็นมาตรฐานที่ช่วยให้สร้างมาตรฐานของคำบางคำให้เป็นนิพจน์เฉพาะ ข้อเสีย** ในไพพ์ไลน์การทำเหมืองข้อความหลายๆ แบบ มีตัวเลือกมากมายที่เกี่ยวข้องซึ่งอาจทำให้ข้อมูลสูญหายได้

* + - 1. Part of Speech Tagging with NLTK



**POS Tagging** (ส่วนหนึ่งของการแท็กคำพูด) เป็นกระบวนการในการทำเครื่องหมายคำในรูปแบบข้อความสำหรับส่วนใดส่วนหนึ่งของคำพูดตามคำจำกัดความและบริบท มีหน้าที่รับผิดชอบในการอ่านข้อความในภาษาและกำหนดโทเค็นเฉพาะ (Parts of Speech) ให้กับแต่ละคำ เรียกอีกอย่างว่าการติดแท็กทางไวยากรณ์

**ตัวอย่างแท็ก NLTK POS มีดังนี้:**

|  |  |
| --- | --- |
| **Abbreviation** | **Meaning** |
| CC | coordinating conjunction |
| CD | cardinal digit |
| DT | determiner |
| EX | existential there |
| FW | foreign word |
| IN | preposition/subordinating conjunction |
| JJ | This NLTK POS Tag is an adjective (large) |
| JJR | adjective, comparative (larger) |
| JJS | adjective, superlative (largest) |
| LS | list market |
| MD | modal (could, will) |
| NN | noun, singular (cat, tree) |
| NNS | noun plural (desks) |
| NNP | proper noun, singular (sarah) |
| NNPS | proper noun, plural (indians or americans) |
| PDT | predeterminer (all, both, half) |
| POS | possessive ending (parent\ ‘s) |
| PRP | personal pronoun (hers, herself, him, himself) |
| PRP$ | possessive pronoun (her, his, mine, my, our ) |
| RB | adverb (occasionally, swiftly) |
| RBR | adverb, comparative (greater) |
| RBS | adverb, superlative (biggest) |
| RP | particle (about) |
| TO | infinite marker (to) |
| UH | interjection (goodbye) |
| VB | verb (ask) |
| VBG | verb gerund (judging) |
| VBD | verb past tense (pleaded) |
| VBN | verb past participle (reunified) |
| VBP | verb, present tense not 3rd person singular(wrap) |
| VBZ | verb, present tense with 3rd person singular (bases) |
| WDT | wh-determiner (that, what) |
| WP | wh- pronoun (who) |
| WRB | wh- adverb (how) |

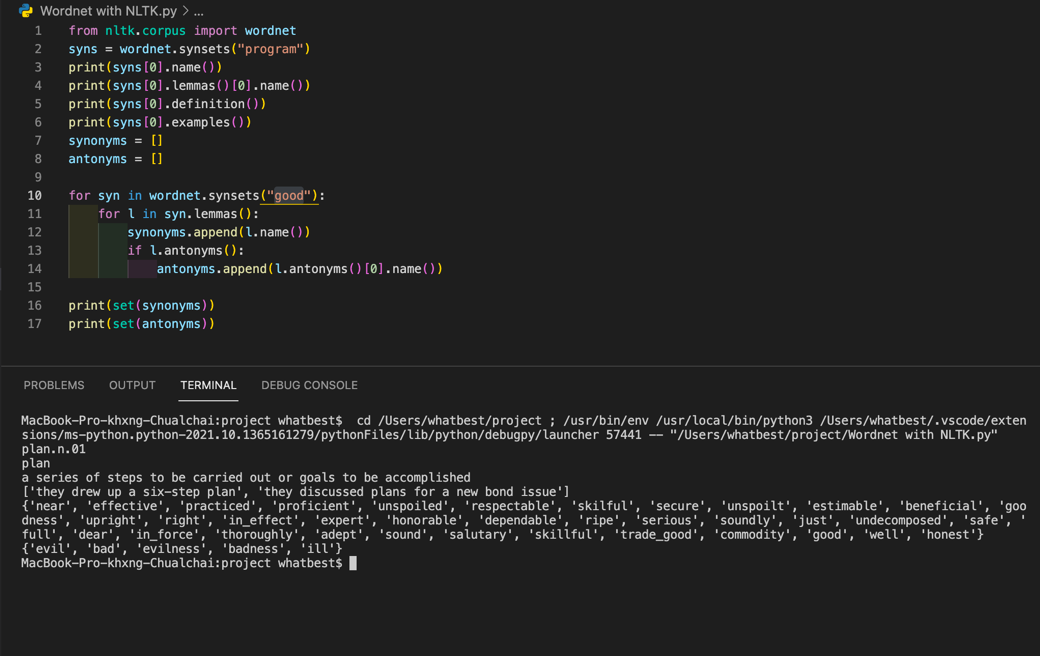
* + - 1. **Lemmatizing with NLTK**

รูปภาพประกอบด้วย ข้อความ

คำอธิบายที่สร้างโดยอัตโนมัติ

Lemmatization เป็นกระบวนการของการจัดกลุ่มคำในรูปแบบผันแปรต่างๆ เพื่อให้สามารถวิเคราะห์เป็นรายการเดียวได้ Lemmatization คล้ายกับการกำเนิด แต่นำบริบทมาสู่คำ ดังนั้นจึงเชื่อมโยงคำที่มีความหมายคล้ายกันเป็นคำเดียว

* 1. **ความก้าวหน้า 3 ทดสอบ Library NLTK** 
     1. รายละเอียดการทดลอง
        1. Wordnet with NLTK

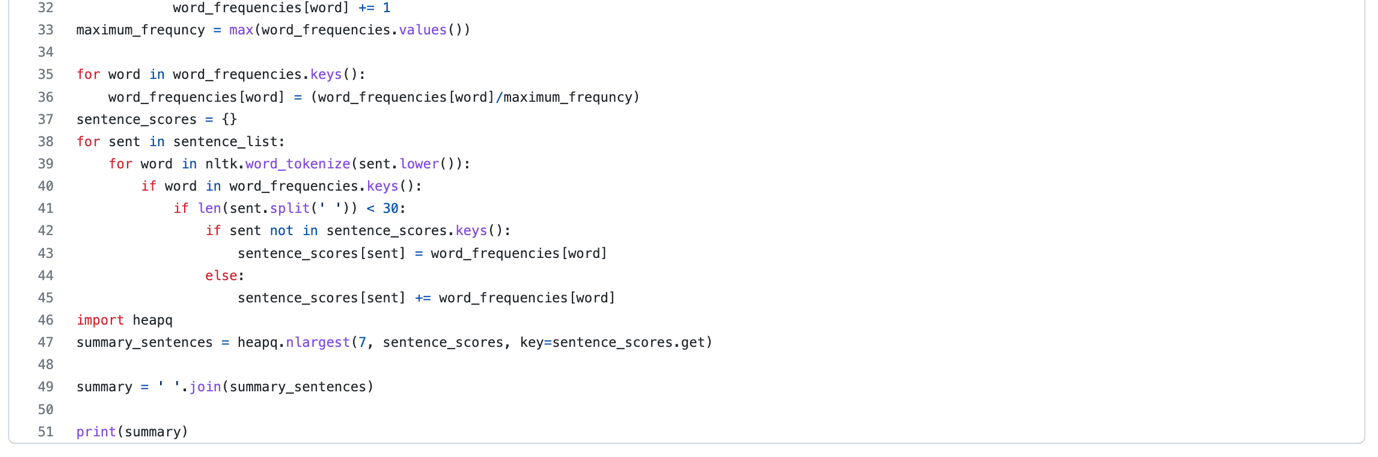


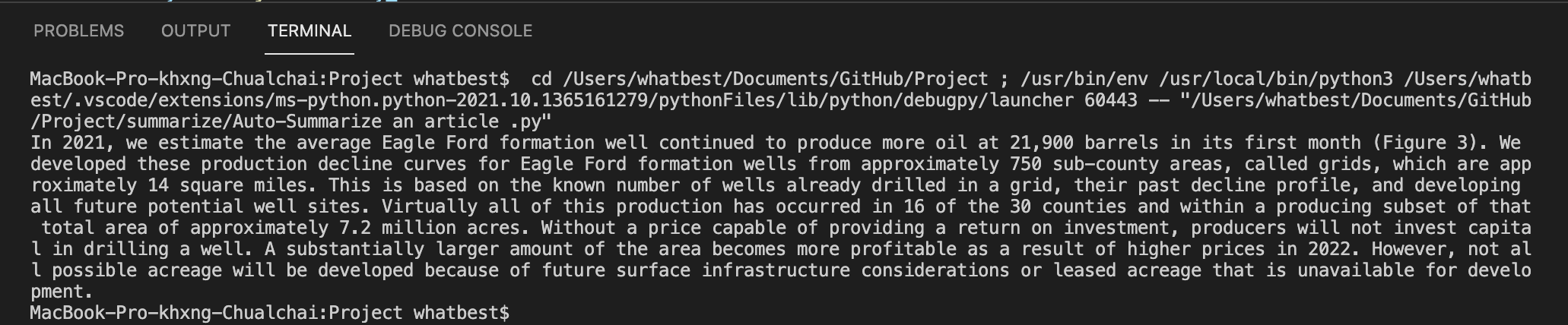
[**WordNet**](https://wordnet.princeton.edu/)เป็นฐานข้อมูลคำศัพท์สำหรับภาษาอังกฤษ ซึ่งสร้างโดย Princeton และเป็นส่วนหนึ่งของคลังข้อมูล NLTK

* + - 1. **Text Summarization with NLTK in Python**

รูปภาพประกอบด้วย ข้อความ

คำอธิบายที่สร้างโดยอัตโนมัติ





ในสคริปต์ด้านบน นำเข้าไลบรารีที่สำคัญที่จำเป็นสำหรับการดึงข้อมูลจากเว็บก่อน จากนั้นใช้urlopenฟังก์ชันจากurllib.requestยูทิลิตี้เพื่อขูดข้อมูล ต่อไป ต้องเรียกreadใช้ฟังก์ชันบนวัตถุที่ส่งคืนโดยurlopenฟังก์ชันเพื่ออ่านข้อมูล ในการแยกวิเคราะห์ข้อมูล ใช้BeautifulSoupอ็อบเจ็กต์และส่งผ่านออบเจ็กต์ข้อมูลที่คัดลอกมา เช่นarticleและlxmlตัวแยกวิเคราะห์

article\_textมีข้อความโดยไม่มีวงเล็บ อย่างไรก็ตาม จะไม่ลบอะไรไปจากบทความเนื่องจากเป็นบทความต้นฉบับ จะไม่ลบตัวเลข เครื่องหมายวรรคตอน และสัญลักษณ์พิเศษอื่นๆ ออกจากข้อความนี้ เนื่องจากเราจะใช้ข้อความนี้เพื่อสร้างบทสรุปและความถี่ของคำแบบถ่วงน้ำหนัก

มีสองอ็อบเจ็กต์article\_textซึ่งประกอบด้วยบทความต้นฉบับและบทความformatted\_article\_textที่จัดรูปแบบ จะใช้formatted\_article\_textเพื่อสร้างฮิสโทแกรมความถี่ถ่วงน้ำหนักสำหรับคำนั้นๆ และจะแทนที่ความถี่ถ่วงน้ำหนักเหล่านี้ด้วยคำในarticle\_textออบเจ็กต์

**Converting Text To Sentences** ได้ทำการประมวลผลข้อมูลล่วงหน้าแล้ว ต่อไป ต้องแปลงบทความให้เป็นประโยค จะใช้article\_textวัตถุสำหรับ tokenizing บทความเป็นประโยคเนื่องจากมีการหยุดเต็ม formatted\_article\_textไม่มีเครื่องหมายวรรคตอนใดๆ ดังนั้นจึงไม่สามารถแปลงเป็นประโยคโดยใช้จุดเต็มเป็นพารามิเตอร์ได้

**Converting Text To Sentences** ในการหาความถี่ของการเกิดขึ้นของแต่ละคำ ใช้formatted\_article\_textตัวแปร ใช้ตัวแปรนี้เพื่อค้นหาความถี่ของการเกิดเนื่องจากไม่มีเครื่องหมายวรรคตอน ตัวเลข หรืออักขระพิเศษอื่นๆ

ขั้นแรกจะเก็บคำหยุดภาษาอังกฤษทั้งหมดจากnltkไลบรารีลงในstopwordsตัวแปร ต่อไป จะวนรอบประโยคทั้งหมดแล้วตามด้วยคำที่เกี่ยวข้องเพื่อตรวจดูก่อนว่าเป็นคำหยุดหรือไม่ หากไม่เป็นเช่นนั้น จะดำเนินการตรวจสอบว่าคำนั้นมีอยู่ในword\_frequencyพจนานุกรมหรือไม่ เช่นword\_frequenciesหรือไม่ หากพบคำนี้เป็นครั้งแรก คำนั้นจะถูกเพิ่มลงในพจนานุกรมเป็นคีย์และตั้งค่าเป็น 1 มิฉะนั้น หากคำนั้นมีอยู่ในพจนานุกรมก่อนหน้านี้ ค่าของคำนั้นจะถูกอัพเดตเพียง 1

**Calculating Sentence Scores** ตอนนี้ได้คำนวณความถี่ถ่วงน้ำหนักสำหรับคำทั้งหมดแล้ว ตอนนี้เป็นเวลาที่จะคำนวณคะแนนสำหรับแต่ละประโยคโดยการเพิ่มความถี่ถ่วงน้ำหนักของคำที่เกิดขึ้นในประโยคนั้นโดยเฉพาะ

ขั้นแรกจะสร้างsentence\_scoresพจนานุกรม เปล่า กุญแจของพจนานุกรมนี้จะเป็นตัวประโยคเอง และค่าจะเป็นคะแนนที่สอดคล้องกันของประโยค ต่อไป จะวนรอบแต่ละประโยคใน the sentence\_listและแปลงประโยคเป็นคำ

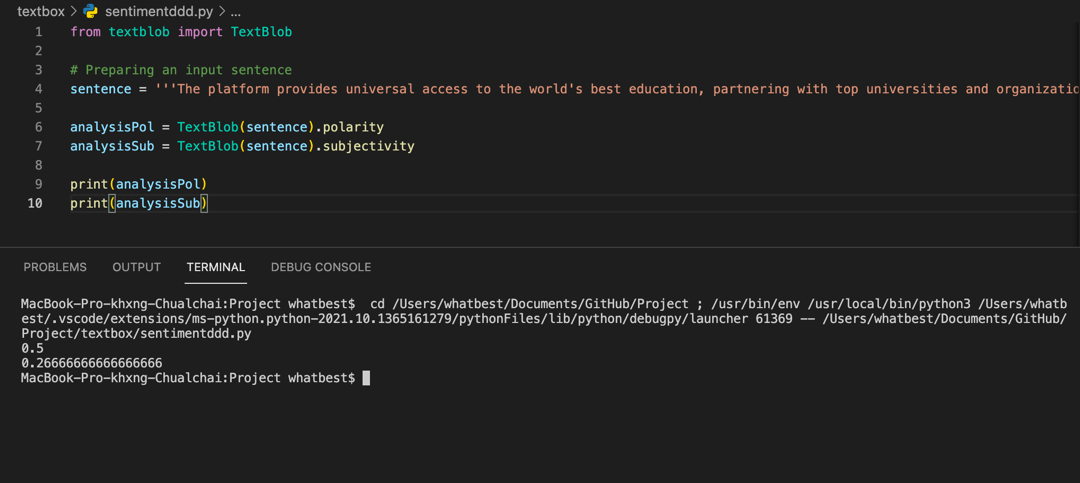
จากนั้นจะตรวจสอบว่าคำนั้นมีอยู่ในword\_frequenciesพจนานุกรมหรือไม่ การตรวจสอบนี้ดำเนินการเนื่องจากสร้างsentence\_listรายการจากarticle\_textวัตถุ ในทางกลับกัน ความถี่ของคำถูกคำนวณโดยใช้formatted\_article\_textออบเจกต์ ซึ่งไม่มีคำหยุด ตัวเลข ฯลฯ

ไม่ต้องการประโยคที่ยาวมากในการสรุป ดังนั้นจึงคำนวณคะแนนสำหรับประโยคที่มีคำน้อยกว่า 30 คำเท่านั้น (แม้ว่าจะปรับแต่งพารามิเตอร์นี้สำหรับกรณีการใช้งานของเองได้ก็ตาม) ต่อไปจะตรวจสอบว่าประโยคนั้นมีอยู่ในsentence\_scoresพจนานุกรมหรือไม่ หากไม่มีประโยคดังกล่าว จะเพิ่มลงในsentence\_scoresพจนานุกรมเป็นคีย์และกำหนดความถี่ถ่วงน้ำหนักของคำแรกในประโยคเป็นค่าของประโยค ในทางตรงกันข้าม หากประโยคนั้นมีอยู่ในพจนานุกรม เพียงแค่เพิ่มความถี่ถ่วงน้ำหนักของคำนั้นให้กับค่าที่มีอยู่

**Getting the Summary** ตอนนี้มีsentence\_scoresพจนานุกรมที่มีประโยคที่มีคะแนนตรงกัน เพื่อสรุปบทความสามารถใช้ประโยค N อันดับแรกที่มีคะแนนสูงสุด สคริปต์ต่อไปนี้ดึงประโยค 7 อันดับแรก

ใช้ ไลบรารี่ heapqและเรียกใช้nlargestฟังก์ชันเพื่อดึงประโยค 7 อันดับแรกที่มีคะแนนสูงสุด

* + - 1. **Sentiment Analasis**



อาต์พุตของ TextBlob สำหรับงานวิเคราะห์การวิเคราะห์จะภายในช่วง[-1.0, 1.0]ที่-1.0เป็นขั้วลบและ1.0เป็นบวก คะแนนนี้ยังสามารถเท่ากับ0ซึ่งหมายถึงการประเมินที่เป็นกลางของคำสั่ง เนื่องจากไม่มีคำใด ๆ จากชุดการฝึก

ในขณะที่งานการ ระบุ อัตนัย / ความเป็นวัตถุรายงานการลอยตัวภายในช่วง[0.0, 1.0]ที่0.0เป็นประโยคที่เป็นกลางและ1.0เป็นอัตนัยมาก

เมื่อนำเข้าแล้ว เราจะโหลดประโยคเพื่อวิเคราะห์และสร้างอินสแตนซ์ของTextBlobวัตถุ รวมทั้งกำหนดsentimentคุณสมบัติให้กับของเราเองanalysis:

คุณสมบัติsentimentเป็น a ของ แบบnamedtupleฟอร์มSentiment(polarity, subjectivity)

ผลลัพธ์ที่คาดหวังของการวิเคราะห์คือ:

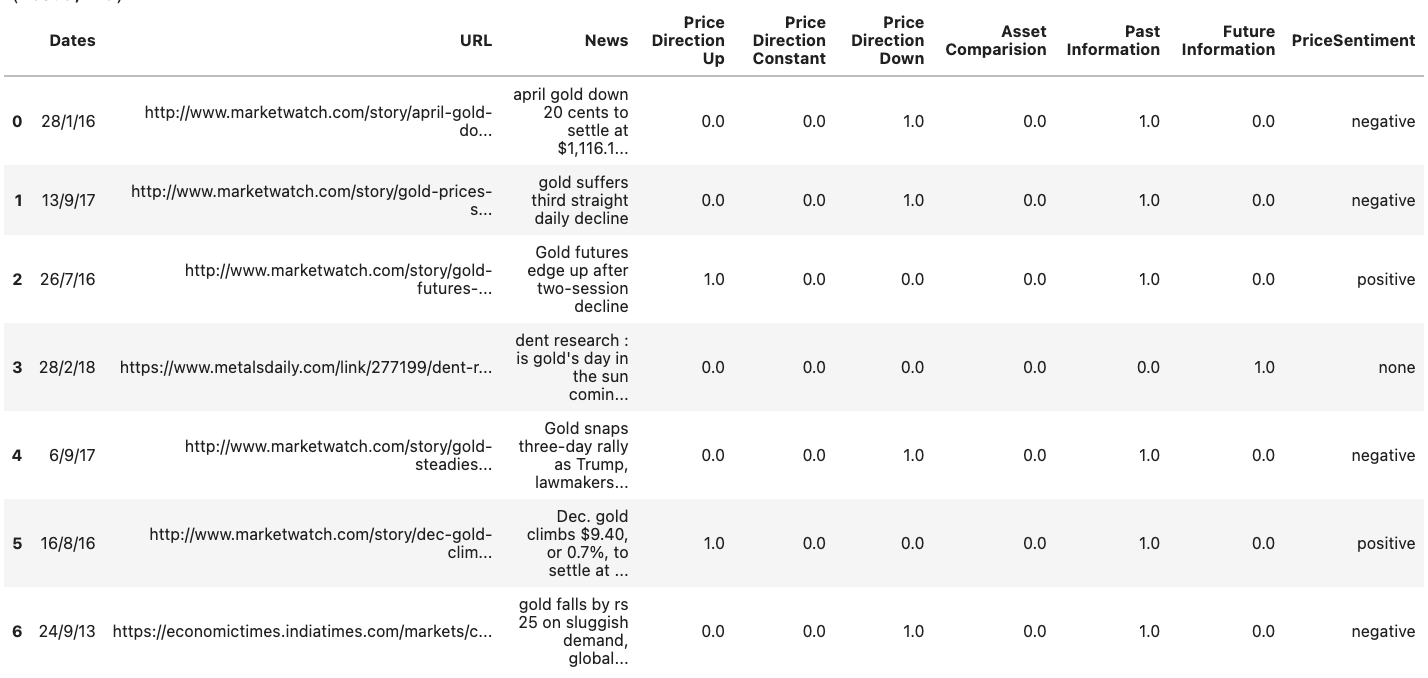
สิ่งที่ยอดเยี่ยมอย่างหนึ่งเกี่ยวกับ TextBlob คือช่วยให้ผู้ใช้สามารถเลือกอัลกอริธึมสำหรับการใช้งาน NLP ระดับสูงได้:

PatternAnalyzer- ตัวแยกประเภทเริ่มต้นที่สร้างขึ้นบนไลบรารีรูปแบบ

NaiveBayesAnalyzer- โมเดล NLTK ที่ได้รับการฝึกอบรมเกี่ยวกับคลังบทวิจารณ์ภาพยนตร์

* 1. **ความก้าวหน้า 4 สร้างเหมืองข้อมูล และทดสอบ**

ชุดข้อมูลนี้รวบรวมจากแหล่งข่าวต่างๆ และใส่คำอธิบายประกอบโดยนักบันทึกย่อที่เป็นมนุษย์สามคนซึ่งเป็นผู้เชี่ยวชาญในเรื่อง แต่ละพาดหัวข่าวได้รับการประเมินในมิติต่างๆ เช่น หากพาดหัวข่าวเป็นข่าวที่เกี่ยวข้องกับราคา ทิศทางของการเคลื่อนไหวของราคาที่พูดถึงคืออะไร ไม่ว่าพาดหัวข่าวจะพูดถึงอดีตหรืออนาคต รายการข่าวกำลังพูดถึงการเปรียบเทียบสินทรัพย์หรือไม่ เป็นต้น

****

ในช่วงไม่กี่ปีที่ผ่านมา มีการใช้วิธีการเรียนรู้ด้วยเครื่องเพื่อดึงข้อมูลจากกระแสข่าวในโดเมนการเงิน อย่างไรก็ตาม ข้อมูลนี้ส่วนใหญ่อยู่ในรูปแบบของความรู้สึกทางการเงินที่มีอยู่ในหัวข้อข่าว โดยเฉพาะราคาหุ้น ในงานปัจจุบันของเรา เราขอเสนอว่าสามารถดึงข้อมูลมิติอื่นๆ ที่หลากหลายจากหัวข้อข่าว ซึ่งจะเป็นที่สนใจของนักลงทุน ผู้กำหนดนโยบาย และผู้ปฏิบัติงานอื่นๆ เราเสนอกรอบการทำงานที่ดึงข้อมูล เช่น การเคลื่อนไหวในอดีตและทิศทางที่คาดหวังในด้านราคา การเปรียบเทียบสินทรัพย์ และข้อมูลทั่วไปอื่นๆ ที่ข่าวกล่าวถึง เราใช้กรอบการทำงานนี้กับสินค้าโภคภัณฑ์ "ทองคำ" และฝึกโมเดลการเรียนรู้ของเครื่องโดยใช้ชุดข้อมูลหัวข้อข่าวที่มีคำอธิบายประกอบโดยมนุษย์ 11,412 รายการ (เผยแพร่พร้อมกับการศึกษานี้) รวบรวมตั้งแต่ช่วง พ.ศ. 2543-2562 เราทดลองเพื่อตรวจสอบผลกระทบของกระแสข่าวที่มีต่อราคาทองคำ และสังเกตว่าข้อมูลที่ผลิตจากกรอบของเราส่งผลกระทบอย่างมีนัยสำคัญต่อราคาทองคำในอนาคต

* 1. **ความก้าวหน้า 5 ทดสอบ Model** Training and Test datasets.

**ขั้นตอนการทำงาน มี 7 ขั้นตอน**

Step 1 - Loading the required libraries and modules.

Step 2 - Loading the data and performing basic data checks.

Step 3 - Pre-processing the raw text and getting it ready for machine learning.

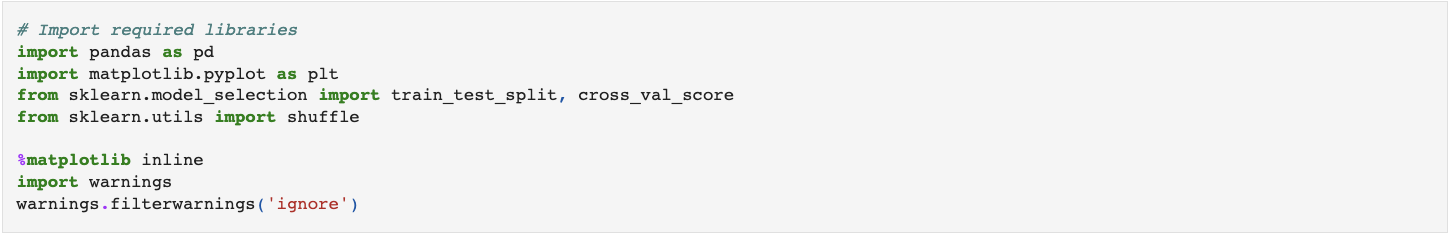
Step 4 - Creating the Training and Test datasets.

Step 5 - Converting text to word frequency vectors with TfidfVectorizer.

Step 6 - Create and fit the classifier.

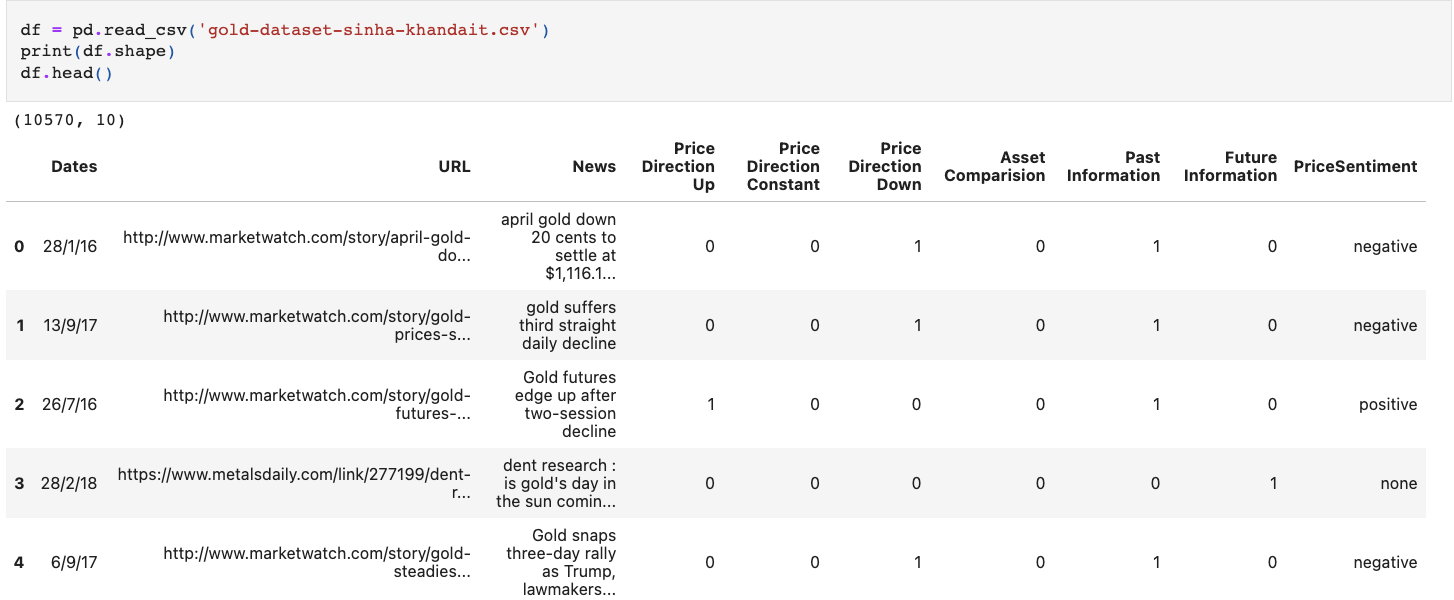
Step 7 - Computing the evaluation metrics.

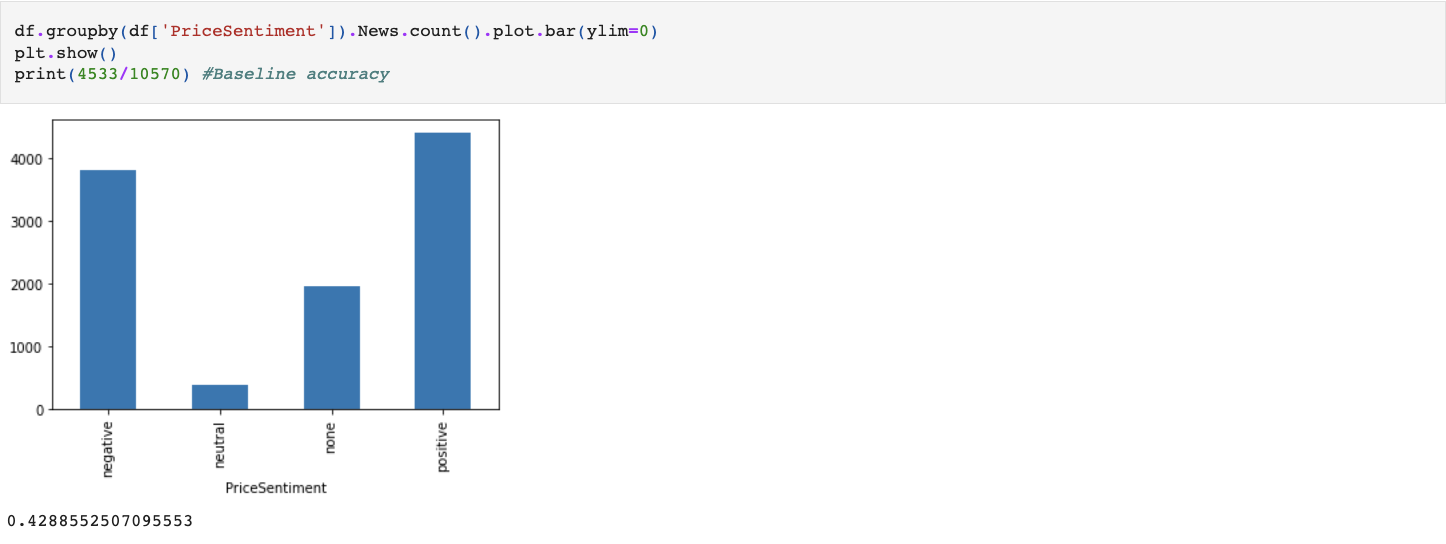
### **Step 1 - Loading the Required Libraries and Modules**

****

### **Step 2 - Loading the Data and Performing Basic Data Checks**

โค้ด บรรทัดแรกอ่านในข้อมูลเป็น data frame ของ pandas ในขณะที่บรรทัดที่สองพิมพ์รูปร่าง – 10,570 การสังเกตจาก 10 ตัวแปร บรรทัดที่สามพิมพ์ข้อสังเกตห้าข้อแรก

****

****

### Step 3 – Pre-processing the Raw Text and Getting It Ready for Machine Learning

ขั้นตอนก่อนการประมวลผลทั่วไปคือ:

การลบเครื่องหมายวรรคตอน - หลักการทั่วไปคือการลบทุกอย่างที่ไม่อยู่ในรูปแบบ x,y,z

การนำคำหยุดออก - คำเหล่านี้เป็นคำที่ไม่มีประโยชน์ เช่น 'the', 'is', 'at' สิ่งเหล่านี้ไม่มีประโยชน์เพราะความถี่ของคำหยุดนั้นอยู่ในคลังข้อมูลสูง แต่ก็ไม่ได้ช่วยในการสร้างความแตกต่างของคลาสเป้าหมาย การลบ Stopwords ยังช่วยลดขนาดข้อมูลอีกด้วย

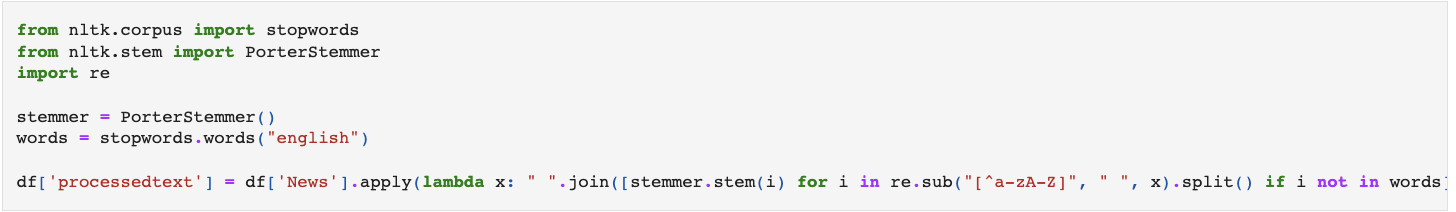
การแปลงเป็นตัวพิมพ์เล็ก - คำเช่น 'คลินิก' และ 'คลินิก' จำเป็นต้องถือเป็นคำเดียว ดังนั้นสิ่งเหล่านี้จะถูกแปลงเป็นตัวพิมพ์เล็ก

Stemming - เป้าหมายของ Stemming คือการลดจำนวนรูปแบบการผันคำที่ปรากฏในข้อความ ซึ่งทำให้คำต่างๆ เช่น "argue", "argued", "arguing", "argues" ถูกลดทอนเป็นคำว่า "argu" ซึ่งช่วยในการลดขนาดของพื้นที่คำศัพท์ มีหลายวิธีในการดำเนินการ Stemming วิธีที่นิยมเป็นวิธี “Porter Stemmer” โดย Martin Porter

สำหรับการทำตามขั้นตอนที่กล่าวไว้ข้างต้นให้เสร็จสิ้น เราจะต้องโหลดแพ็คเกจ nltk ซึ่งทำในโค้ดบรรทัดแรกด้านล่าง บรรทัดที่สองดาวน์โหลดรายการ 'คำหยุด' ในแพ็คเกจ nltk

****

โค้ด ที่สี่ถึงบรรทัดที่หกทำหน้าที่ประมวลผลข้อความล่วงหน้าที่กล่าวถึงข้างต้น

****

ตอนนี้เราจะมีชุดข้อมูลที่ประมวลผลล่วงหน้าซึ่งมีคอลัมน์ ' processedtext' ใหม่

**รูปภาพประกอบด้วย โต๊ะ

คำอธิบายที่สร้างโดยอัตโนมัติ**

### Step 4 - Creating the Training and Test Datasets

โค้ด บรรทัดแรกด้านล่างนำเข้าโมดูลสำหรับสร้างชุดข้อมูลการฝึกอบรมและทดสอบ บรรทัดที่สองสร้างอาร์เรย์ของตัวแปรเป้าหมาย เรียกว่า 'เป้าหมาย'

บรรทัดที่สามสร้างอาร์เรย์การฝึกอบรม (X\_train, y\_train) และชุดการทดสอบ (X-test, y\_test) โดยจะเก็บข้อมูลไว้ 30% สำหรับการทดสอบโมเดล อาร์กิวเมนต์ 'random\_state' ช่วยให้แน่ใจว่าผลลัพธ์สามารถทำซ้ำได้

บรรทัดที่สี่พิมพ์รูปร่างของชุดข้อมูลโดยรวม การฝึกอบรม และการทดสอบ ตามลำดับ

**รูปภาพประกอบด้วย ข้อความ

คำอธิบายที่สร้างโดยอัตโนมัติ**

### Step 5 - Converting Text to Word Frequency Vectors with TfidfVectorizer.

เราได้ประมวลผลข้อความแล้ว แต่เราต้องแปลงเป็นเวกเตอร์ความถี่คำเพื่อสร้างแบบจำลองการเรียนรู้ของเครื่อง มีหลายวิธีในการทำเช่นนี้ เช่น การใช้ CountVectorizer และ HashingVectorizer แต่ TfidfVectorizer เป็นวิธีที่ได้รับความนิยมมากที่สุด

TF-IDF เป็นตัวย่อที่ย่อมาจาก 'Term Frequency-Inverse Document Frequency' มันถูกใช้เป็นปัจจัยถ่วงน้ำหนักในแอปพลิเคชันการทำเหมืองข้อความ

Term Frequency (TF): สรุป Term Frequency ที่เป็นมาตรฐานภายในเอกสาร

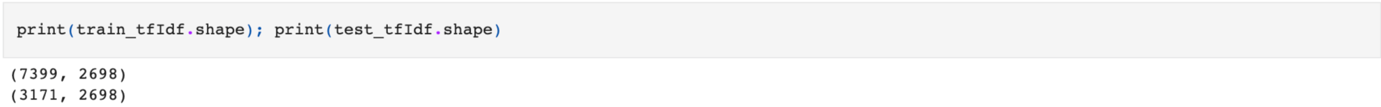
ความถี่เอกสารผกผัน (IDF): วิธีนี้จะช่วยลดน้ำหนักของคำที่ปรากฏบ่อยในเอกสารต่างๆ พูดง่ายๆ ก็คือ TF-IDF พยายามเน้นคำสำคัญที่มักพบในเอกสารแต่ไม่อยู่ในเอกสาร เราจะพยายามสร้างเวกเตอร์ TF-IDF สำหรับเอกสารของเรา

บรรทัดแรกของโค้ดด้านล่างนำเข้า TfidfVectorizer จากโมดูล 'sklearn.feature\_extraction.text' บรรทัดที่สองเริ่มต้นวัตถุ TfidfVectorizer ที่เรียกว่า 'vectorizer\_tfidf'

บรรทัดที่สามพอดีและแปลงข้อมูลการฝึก โค้ดบรรทัด ที่สี่จะเปลี่ยนข้อมูลการทดสอบ ในขณะที่บรรทัดที่ห้าจะพิมพ์คุณลักษณะ 10 รายการแรก

**รูปภาพประกอบด้วย ข้อความ

คำอธิบายที่สร้างโดยอัตโนมัติ**

****

### Step 6 - Create and Fit the Classifier.

ตอนนี้ เราจะสร้างแบบจำลองการจัดประเภทข้อความ อัลกอริทึมที่เราจะเลือกคือ Naive Bayes Classifier ซึ่งมักใช้สำหรับปัญหาการจัดประเภทข้อความ เนื่องจากขึ้นอยู่กับความน่าจะเป็น มันง่ายและมีประสิทธิภาพในการตอบคำถามเช่น "ให้คำเฉพาะในเอกสารอะไรเป็นโอกาส (ความน่าจะเป็น) ที่จะเป็นของชั้นเรียนนั้น"

เราเริ่มต้นด้วยการนำเข้าโมดูลที่จำเป็นซึ่งเสร็จสิ้นในโค้ดสองบรรทัดแรกด้านล่าง บรรทัดที่สามสร้างตัวแยกประเภท Multinomial Naive Bayes เรียกว่า 'nb\_classifier' รหัสบรรทัดที่สี่เหมาะกับตัวแยกประเภทในข้อมูลการฝึกอบรม

สุดท้าย โมเดลของเราได้รับการฝึกฝนและพร้อมที่จะสร้างการคาดการณ์เกี่ยวกับข้อมูลที่มองไม่เห็น การดำเนินการนี้ดำเนินการในโค้ดบรรทัดที่ห้าในขณะที่บรรทัดที่หกจะพิมพ์คลาสที่คาดการณ์ไว้สำหรับระเบียน 10 รายการแรกในข้อมูลการทดสอบ

****

### Step 7 - Computing the Evaluation Metrics

ตอนนี้เราพร้อมที่จะประเมินประสิทธิภาพของแบบจำลองของเรากับข้อมูลการทดสอบแล้ว การใช้ฟังก์ชัน 'metrics.accuracy\_score' เราคำนวณความถูกต้องในบรรทัดแรกของโค้ดด้านล่าง และพิมพ์ผลลัพธ์โดยใช้โค้ดบรรทัดที่สอง เราเห็นว่าความแม่นยำอยู่ที่ 73% ซึ่งเป็นคะแนนที่ดี

นอกจากนี้เรายังสามารถคำนวณความถูกต้องผ่านตัวชี้วัดความสับสน โค้ดบรรทัด ที่สามด้านล่างสร้างตัวชี้วัดความสับสน ซึ่งอาร์กิวเมนต์ 'labels' ใช้เพื่อระบุป้ายกำกับคลาสเป้าหมาย ('positive' , 'negative', ‘none’, ‘neutral’ ในกรณีของเรา) บรรทัดที่สี่พิมพ์ตัวชี้วัด

**รูปภาพประกอบด้วย ข้อความ

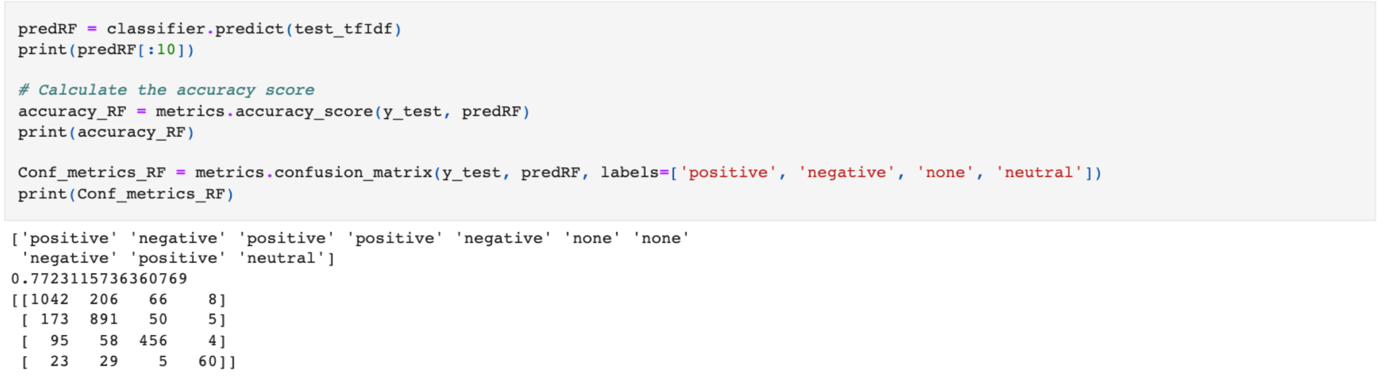
คำอธิบายที่สร้างโดยอัตโนมัติ**

## Building Random Forest Classifier

โค้ดสองบรรทัดแรกด้านล่างนำเข้าโมดูลที่จำเป็น บรรทัดที่สามสร้าง Random Forest Classifier ในขณะที่บรรทัดที่สี่เหมาะกับตัวแยกประเภทในข้อมูลการฝึกอบรม

****

เมื่อการฝึกโมเดลเสร็จสิ้น เราใช้โมเดลเพื่อสร้างการคาดการณ์เกี่ยวกับข้อมูลการทดสอบ ซึ่งจะทำในโค้ดบรรทัดแรกด้านล่าง บรรทัดที่สองพิมพ์คลาสที่คาดการณ์ไว้สำหรับ 10 ระเบียนแรกในข้อมูลการทดสอบ โค้ดบรรทัด ที่สามและสี่จะคำนวณและพิมพ์คะแนนความถูกต้องตามลำดับ เราเห็นว่าความแม่นยำเพิ่มขึ้นเหลือ 77.2%

****

* 1. **ความก้าวหน้า 6 ศึกษา Library Python NLP และทดสอบ**
  2. **ความก้าวหน้า 7 ศึกษา Library Python NLP และทดสอบ**

**บทที่ 5**

**สรุป**

**5.1. สรุปผลการดำเนินงาน**

1. สามารถสรุปบทความได้ ใช้ Library NLTK NLP

2. สามารถวิเคราะห์ บทความได้ระดับเบื้องต้น

**5.2. ปัญหาและอุปสรรค**

1. ยังใช้ภาษา Python ได้ไม่ชำนาญ

2. Library ที่จะใช้ค่อนข้างเยอะ จึงทำให้เริ่ม Project ช้า

3. มีปัญหาเรื่อง Version ที่จะติดตั้ง Library

**5.3. งานที่จะดำเนินการต่อไป**

1. ปรับปรุงแก้ไขการทำ sentiment Analysis

2. เทรนข้อมูลที่เกี่ยวกับ Forex

3. ทดสอบ Algorithm ระหว่าง SpaCy กับ NLTK

**บรรณานุกรม**

[1] S. Shalev-Shwartz, S. Ben-David, [Understanding Machine Learning: From Theory to Algorithms](https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/) (2014), Cambridge University Press , เข้าถึงล่าสุด 15 มกราคม 2565

[2] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. [Distributed Representations of Words and Phrases and their Compositionality](https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf) (2013), Advances in Neural Information Processing Systems 26 เข้าถึงล่าสุด 15 มกราคม 2565

[3] J. Pennington, R. Socher, and C. D. Manning, [GloVe: Global Vectors for Word Representation](https://www.aclweb.org/anthology/D14-1162) (2014), In EMNLP. เข้าถึงล่าสุด 16 มกราคม 2565

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. [Enriching word vectors with subword information](https://aclweb.org/anthology/Q17-1010) (2016), arXiv preprint เข้าถึงล่าสุด 17 มกราคม 2565

# [5] NLP Implementations : URL : <https://towardsdatascience.com/gentle-start-to-natural-language-processing-using-python-6e46c07addf3> เข้าถึงล่าสุด 18 มกราคม 2565

# [6] The theory you need to know before you start an NLP : URL : [https://towardsdatascience.com/the-theory-you-need-to-know-before-you-start-an-nlp-project-1890f5bbb793](https://towardsdatascience.com/the-theory-you-need-to-know-before-you-start-an-nlp-project-1890f5bbb793%20เข้าถึงล่าสุด%2012%20มีนาคม%202565) เข้าถึงล่าสุด 12 มีนาคม 2565

# [7] Us Department of labor : URL : <https://www.dol.gov/> เข้าถึงล่าสุด 12 มีนาคม 2565

# [8] Energy information Administration : URL : <https://www.eia.gov/> เข้าถึงล่าสุด 12 มีนาคม 2565

# [9] กองทุน SPDR : URL : <https://traderider.com/forex/spdr-%E0%B8%81%E0%B8%AD%E0%B8%87%E0%B8%97%E0%B8%B8%E0%B8%99%E0%B8%97%E0%B8%AD%E0%B8%87%E0%B8%84%E0%B8%B3%E0%B9%81%E0%B8%97%E0%B9%88%E0%B8%87> เข้าถึงล่าสุด 12 มีนาคม 2565

# [10] Federal Reserve : URL : <https://www.federalreserve.gov/> เข้าถึงล่าสุด 12 มีนาคม 2565

# [11] Bloomberg : URL : <https://www.bloomberg.com/asia> เข้าถึงล่าสุด 12 มีนาคม 2565

# [12] Twitter : URL : <https://twitter.com/> เข้าถึงล่าสุด 12 มีนาคม 2562

# [13] <https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP> ไปทำความเข้าใจแล้วเพิ่มในรายงาน

# [14] Sentiment Analysis of Commodity News (Gold) : URL : <https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-in-commodity-market-gold> เข้าถึงล่าสุด 1 กันยายน 2562