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

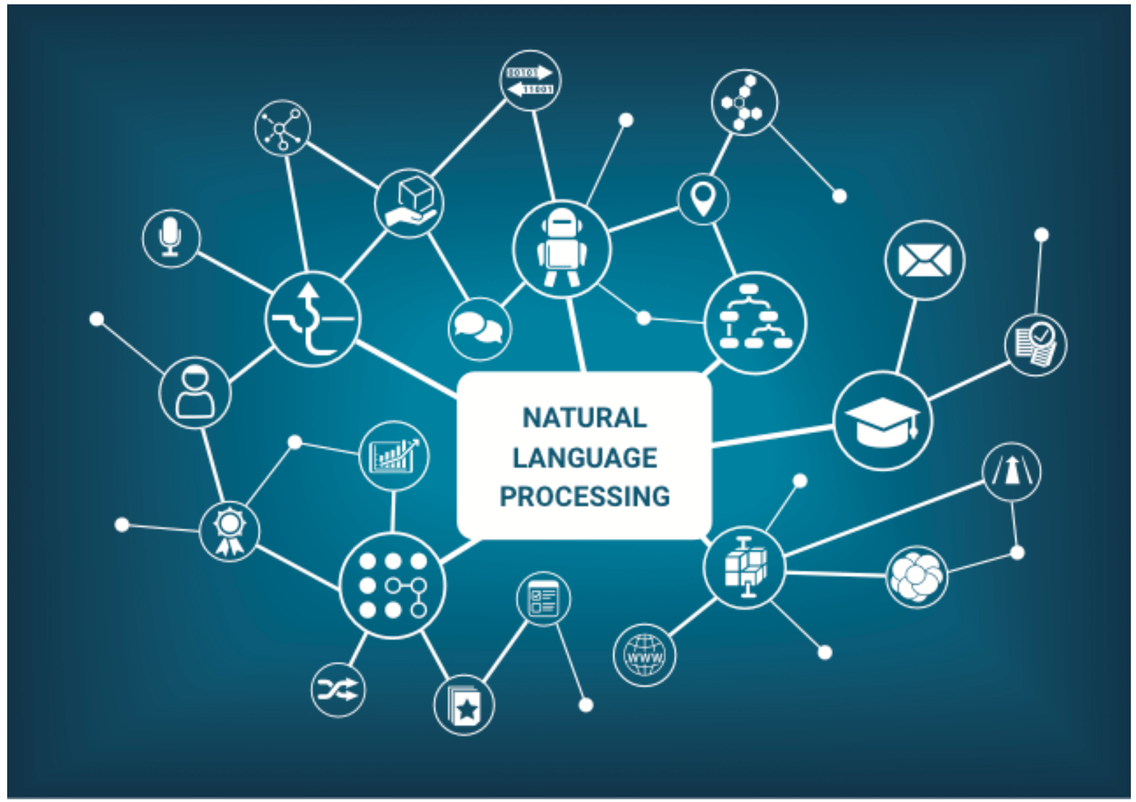
ผู้จัดทำโครงงาน
นายชวลชัย อภิชาติฐิติวรณ์ รหัสนักศึกษา 6210110646

อาจารย์ที่ปรึกษาโครงงาน
รศ.ดร.มนตรี กาญจนะเดชะ

OUTLINE

- บทคัดย่อ
- แผนการดำเนินงาน
- รายละเอียดการดำเนินงาน
- ความก้าวหน้าการดำเนินงาน
- สรุป

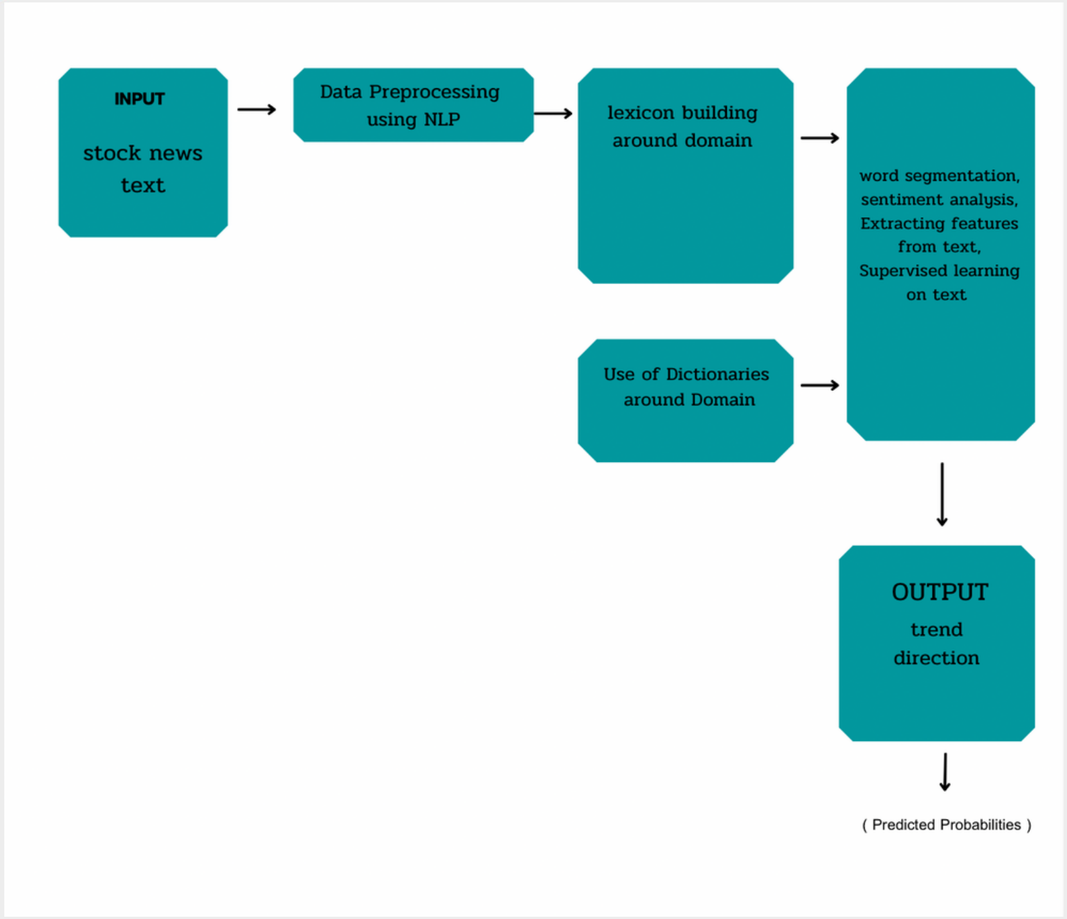
บทคัดย่อ



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	■	■																
(2) พัฒนาเขียนโปรแกรมด้วย Python		■	■	■	■	■												
(3) เทรนสมองกล							■	■	■	■	■	■	■	■				
(4) ทดสอบระบบ							■	■	■	■	■	■	■	■				
(5) ทดลองนำไปใช้จริง กับบัญชี Forex demo											■	■	■	■	■	■	■	■
(6) ปรับปรุงแก้ไขอุปกรณ์ให้สมบูรณ์											■	■	■	■	■			
(7) จัดทำเว็บไซต์															■	■	■	
(8) จัดทำรายงานให้สมบูรณ์															■	■	■	■

รายละเอียดการดำเนินงาน



ขั้นตอน INPUT

ข่าวสารtext

ขั้นตอน Process

ใช้ **algorithms**

ขั้นตอน Output

คาร์ดการณ์ล่วงหน้า

สรุปข้อความ

ขั้นตอน Input

ข่าวสาร	ข่าวประเภท	อ้างอิงค์
1. Census Bureau	กระทรวงพาณิชย์สหรัฐ	[7]
2. Us Department of labo	กระทรวงแรงงาน	[8]
3. Energy information Administration	ข้อมูลด้านพลังงาน	[9]
4. กองทุน SPDR	การซื้อขายทองคำของกองทุน SPDR	[10]
5. Federal Reserve	ธนาคารกลางสหรัฐ	[11]
6. Bloomberg	รายงานข่าวทั่วไปรอบโลก	[12]
7. Twitter	ข่าวทั่วไป	[13]

รูปที่ 1 แหล่งข่าวสาร

ขั้นตอน Process

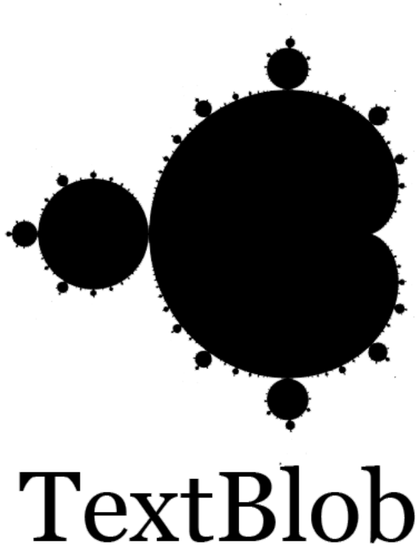
- sentiment analysis
  - tokenization
  - stopword
  - stemmer
- model naive baye

ขั้นตอน Output

- สามารถคาดการ์ณล่วงหน้าได้
- สามารถสรุปบทความได้

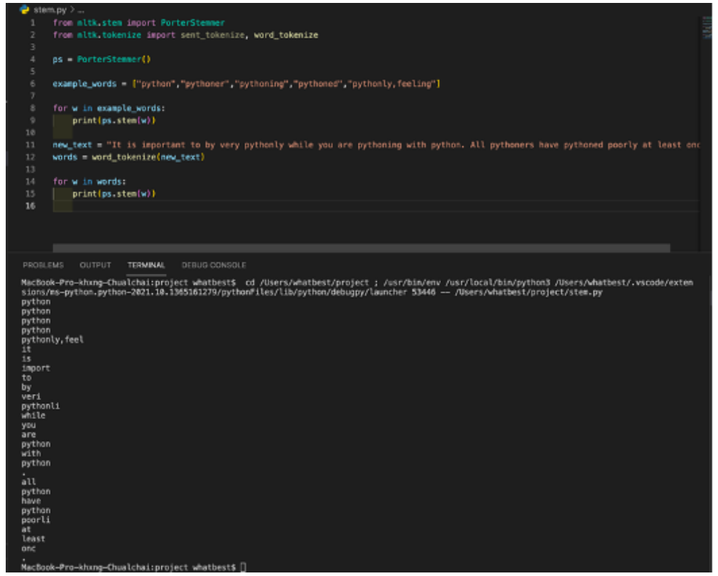
ความก้าวหน้าการดำเนินงาน

ความก้าวหน้า 1 ศึกษา Library Python NLP และทดสอบ

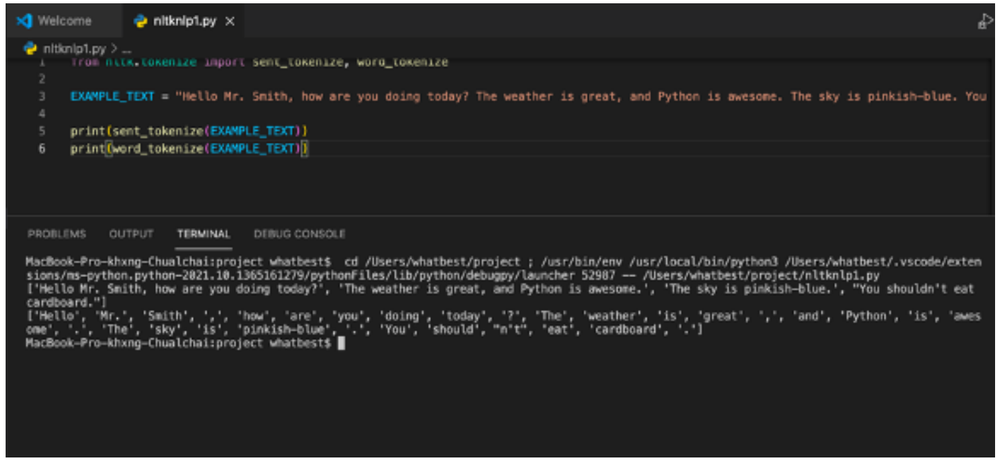


ความก้าวหน้าการดำเนินงาน

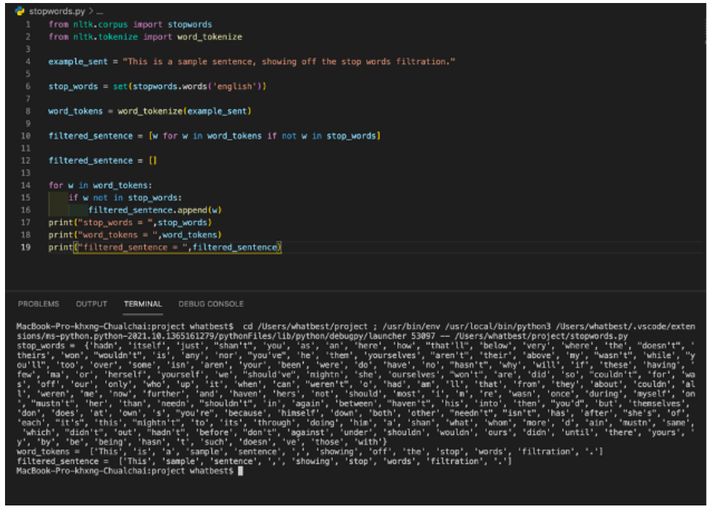
ความก้าวหน้า 2 ทดสอบ Library NLTK



Stemmer words with NLTK



Tokenizing Words and Sentences with NLTK



Stop words with NLTK

ความก้าวหน้าการดำเนินงาน

ความก้าวหน้า 3 Text Summarization with NLTK in Python

main ▾ Project / summarize / Auto-Summarize an article .py / <> Jump to ▾ Go to file ···

Chualchai Apichatitiworn commit — Latest commit 44eb6a6 28 minutes ago — History

0 contributors

51 lines (42 sloc) 1.67 KB — Raw Blame

```python
import bs4 as bs
import urllib.request
import re
import nltk

scraped_data = urllib.request.urlopen('https://www.eia.gov/petroleum/weekly/')
article = scraped_data.read()

parsed_article = bs.BeautifulSoup(article,'lxml')

paragraphs = parsed_article.find_all('p')

article_text = ""

for p in paragraphs:
    article_text += p.text
# Removing Square Brackets and Extra Spaces
article_text = re.sub(r'\[[0-9]*\]', ' ', article_text)
article_text = re.sub(r'\s+', ' ', article_text)
# Removing special characters and digits
formatted_article_text = re.sub('[^a-zA-Z]', ' ', article_text )
formatted_article_text = re.sub(r'\s+', ' ', formatted_article_text)
sentence_list = nltk.sent_tokenize(article_text)
stopwords = nltk.corpus.stopwords.words('english')

word_frequencies = {}
for word in nltk.word_tokenize(formatted_article_text):
    if word not in stopwords:
        if word not in word_frequencies.keys():
            word_frequencies[word] = 1
        else:
            word_frequencies[word] += 1
maximum_frequncy = max(word_frequencies.values())

for word in word_frequencies.keys():
    word_frequencies[word] = (word_frequencies[word]/maximum_frequncy)
sentence_scores = {}
for sent in sentence_list:
    for word in nltk.word_tokenize(sent.lower()):
        if word in word_frequencies.keys():
            if len(sent.split(' ')) < 30:
                if sent not in sentence_scores.keys():
                    sentence_scores[sent] = word_frequencies[word]
                else:
                    sentence_scores[sent] += word_frequencies[word]
import heapq
summary_sentences = heapq.nlargest(7, sentence_scores, key=sentence_scores.get)

summary = ' '.join(summary_sentences)

print(summary)
```

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

```
MacBook-Pro-khxng-Chualchai:Project whatbest$  cd /Users/whatbest/Documents/GitHub/Project ; /usr/bin/env /usr/local/bin/python3 /Users/whatbest/.vscode/extensions/ms-python.python-2021.10.1365161279/pythonFiles/lib/python/debugpy/launcher 60443 -- "/Users/whatbest/Documents/GitHub/Project/summarize/Auto-Summarize an article .py"
In 2021, we estimate the average Eagle Ford formation well continued to produce more oil at 21,900 barrels in its first month (Figure 3). We developed these production decline curves for Eagle Ford formation wells from approximately 750 sub-county areas, called grids, which are approximately 14 square miles. This is based on the known number of wells already drilled in a grid, their past decline profile, and developing all future potential well sites. Virtually all of this production has occurred in 16 of the 30 counties and within a producing subset of that total area of approximately 7.2 million acres. Without a price capable of providing a return on investment, producers will not invest capital in drilling a well. A substantially larger amount of the area becomes more profitable as a result of higher prices in 2022. However, not all possible acreage will be developed because of future surface infrastructure considerations or leased acreage that is unavailable for development.
MacBook-Pro-khxng-Chualchai:Project whatbest$
```

Tokennization

stopword

stemmer

ความก้าวหน้าการดำเนินงาน

ความก้าวหน้า 4 สร้างเหมืองข้อมูล

	Dates	URL	News	Price Direction Up	Price Direction Constant	Price Direction Down	Asset Comparision	Past Information	Future Information	PriceSentiment
0	28/1/16	http://www.marketwatch.com/story/april-gold-do...	april gold down 20 cents to settle at $1,116.1...	0.0	0.0	1.0	0.0	1.0	0.0	negative
1	13/9/17	http://www.marketwatch.com/story/gold-prices-s...	gold suffers third straight daily decline	0.0	0.0	1.0	0.0	1.0	0.0	negative
2	26/7/16	http://www.marketwatch.com/story/gold-futures-...	Gold futures edge up after two-session decline	1.0	0.0	0.0	0.0	1.0	0.0	positive
3	28/2/18	https://www.metalsdaily.com/link/277199/dent-r...	dent research : is gold's day in the sun comin...	0.0	0.0	0.0	0.0	0.0	1.0	none
4	6/9/17	http://www.marketwatch.com/story/gold-steadies...	Gold snaps three-day rally as Trump, lawmakers...	0.0	0.0	1.0	0.0	1.0	0.0	negative
5	16/8/16	http://www.marketwatch.com/story/dec-gold-clim...	Dec. gold climbs $9.40, or 0.7%, to settle at ...	1.0	0.0	0.0	0.0	1.0	0.0	positive
6	24/9/13	https://economictimes.indiatimes.com/markets/c...	gold falls by rs 25 on sluggish demand, global...	0.0	0.0	1.0	0.0	1.0	0.0	negative

ความก้าวหน้าการดำเนินงาน

ความก้าวหน้า 5 ทดสอบนำเหมืองข้อมูลมาใช้งาน Model Trainnin and Test dataset

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

```
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import re

stemmer = PorterStemmer()
words = stopwords.words("english")

df['processedtext'] = df['News'].apply(lambda x: " ".join([stemmer.stem(i) for i in re.sub("[^a-zA-Z]", " ", x).split() if i not in words
```

ความก้าวหน้าการดำเนินงาน

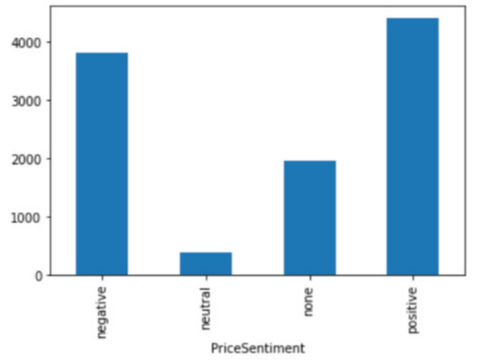
Step 2 - Loading the Data and Performing Basic Data Checks

```
df = pd.read_csv('gold-dataset-sinha-khandait.csv')
print(df.shape)
df.head()
```

```
(10570, 10)
```

	Dates	URL	News	Price Direction Up	Price Direction Constant	Price Direction Down	Asset Comparision	Past Information	Future Information	PriceSentiment
0	28/1/16	http://www.marketwatch.com/story/april-gold-do...	april gold down 20 cents to settle at $1,116.1...	0	0	1	0	1	0	negative
1	13/9/17	http://www.marketwatch.com/story/gold-prices-s...	gold suffers third straight daily decline	0	0	1	0	1	0	negative
2	26/7/16	http://www.marketwatch.com/story/gold-futures-...	Gold futures edge up after two-session decline	1	0	0	0	1	0	positive
3	28/2/18	https://www.metalsdaily.com/link/277199/dent-r...	dent research : is gold's day in the sun comin...	0	0	0	0	0	1	none
4	6/9/17	http://www.marketwatch.com/story/gold-steadies...	Gold snaps three-day rally as Trump, lawmakers...	0	0	1	0	1	0	negative

```
df.groupby(df['PriceSentiment']).News.count().plot.bar(ylim=0)
plt.show()
print(4533/10570) #Baseline accuracy
```




```
0.4288552507095553
```

ความก้าวหน้าการดำเนินงาน

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

```
import nltk
nltk.download('stopwords')
```

```
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import re

stemmer = PorterStemmer()
words = stopwords.words("english")

df['processedtext'] = df['News'].apply(lambda x: " ".join([stemmer.stem(i) for i in re.sub("[^a-zA-Z]", " ", x).split() if i not in words
```

```
print(df.shape)
df.head(10)
```

(10570, 11)

	Dates	URL	News	Price Direction Up	Price Direction Constant	Price Direction Down	Asset Comparision	Past Information	Future Information	PriceSentiment	processed
0	28/1/16	http://www.marketwatch.com/story/april-gold-do...	april gold down 20 cents to settle at $1,116.1...	0	0	1	0	1	0	negative	april gold set
1	13/9/17	http://www.marketwatch.com/story/gold-prices-s...	gold suffers third straight daily decline	0	0	1	0	1	0	negative	gold s third stra daili d
2	26/7/16	http://www.marketwatch.com/story/gold-futures-...	Gold futures edge up after two-session decline	1	0	0	0	1	0	positive	gold futur two ses d
3	28/2/18	https://www.metalsdaily.com/link/277199/dent-r...	dent research : is gold's day in the sun comin...	0	0	0	0	0	1	none	dent rese gold day come
4	6/9/17	http://www.marketwatch.com/story/gold-steadies...	Gold snaps three-day rally as Trump, lawmakers...	0	0	1	0	1	0	negative	gold three day trump law react
5	16/8/16	http://www.marketwatch.com/story/dec-gold-clim...	Dec. gold climbs $9.40, or 0.7%, to settle at ...	1	0	0	0	1	0	positive	dec gold c set
6	24/9/13	https://economictimes.indiatimes.com/markets/c...	gold falls by rs 25 on sluggish demand, global...	0	0	1	0	1	0	negative	gold f slug demand gl
7	23/9/16	http://www.marketwatch.com/story/gold-futures-...	Gold futures fall for the session, but gain fo...	1	0	1	0	1	0	positive	gold futu session
8	21/10/12	https://www.thehindubusinessline.com/opinion/c...	Gold struggles; silver slides, base metals falter	0	1	0	1	1	0	neutral	gold str silver base m f
9	16/3/18	http://www.marketwatch.com/story/april-gold-ho...	april gold holds slight gain, up $2.50, or 0.2...	1	0	0	0	1	0	positive	april gold slight ga

ความก้าวหน้าการดำเนินงาน

Step 4 - Creating the Training and Test Datasets

```
from sklearn.model_selection import train_test_split

target = df['PriceSentiment']

X_train, X_test, y_train, y_test = train_test_split(df['processedtext'], target, test_size=0.30, random_state=100)

print(df.shape); print(X_train.shape); print(X_test.shape)
```

```
(10570, 11)
(7399,)
(3171,)
```

ความก้าวหน้าการดำเนินงาน

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

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer_tfidf = TfidfVectorizer(stop_words='english', max_df=0.7)

train_tfIdf = vectorizer_tfidf.fit_transform(X_train.values.astype('U'))

test_tfIdf = vectorizer_tfidf.transform(X_test.values.astype('U'))

print(vectorizer_tfidf.get_feature_names()[:10])
```

```
['aayog', 'abat', 'abbrevi', 'abc', 'abn', 'acacia', 'acceler', 'access', 'account', 'accredit']
```

```
print(train_tfIdf.shape); print(test_tfIdf.shape)
```

```
(7399, 2698)
(3171, 2698)
```

ความก้าวหน้าการดำเนินงาน

Step 6 - Create and Fit the Classifier.

การประเมิน Naïve Bayes Model

```
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics

nb_classifier = MultinomialNB()

nb_classifier.fit(train_tfIdf, y_train)

pred2 = nb_classifier.predict(test_tfIdf)
print(pred2[:10])
```

```
['positive' 'negative' 'positive' 'positive' 'negative' 'none' 'none'
 'positive' 'negative' 'neutral']
```

ความก้าวหน้าการดำเนินงาน

Step 7 - Computing the Evaluation Metrics

```
# Calculate the accuracy score: score
accuracy_tfidf = metrics.accuracy_score(y_test, pred2)
print(accuracy_tfidf)

Conf_metrics_tfidf = metrics.confusion_matrix(y_test, pred2, labels=['positive', 'negative', 'none', 'neutral'])
print(Conf_metrics_tfidf)
```

```
0.7303689687795648
[[1142  151   29    0]
 [ 297  785   37    0]
 [ 172   64  377    0]
 [  62   40    3   12]]
```

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix

classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 100)

classifier.fit(train_tfIdf, y_train)
```

```
predRF = classifier.predict(test_tfIdf)
print(predRF[:10])

# Calculate the accuracy score
accuracy_RF = metrics.accuracy_score(y_test, predRF)
print(accuracy_RF)

Conf_metrics_RF = metrics.confusion_matrix(y_test, predRF, labels=['positive', 'negative', 'none', 'neutral'])
print(Conf_metrics_RF)
```

```
['positive' 'negative' 'positive' 'positive' 'negative' 'none' 'none'
 'negative' 'positive' 'neutral']
0.7723115736360769
[[1042  206   66    8]
 [ 173  891   50    5]
 [  95   58  456    4]
 [  23   29    5   60]]
```

สรุปผล

ระหว่างRandomForestกับnaïve baye ผลที่ได้

Random Forest มีประเซ็นความแม่นยำที่เยอะกว่า 77%

และ naïve baye 73 %

ความก้าวหน้าการดำเนินงาน

ความก้าวหน้า 6 ทดลองทำ Sentiment Analysis Naïve Bayes Classifier

Read into Python

Let's first read the required data from CSV file using Pandas library.

In [309…]
```
import pandas as pd
import csv
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.utils import shuffle
from sklearn.feature_extraction.text import CountVectorizer
import numpy as np                    #linear algebra
import pandas as pd                   # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt       #For Visualisation
import seaborn as sns                 #For better Visualisation
from bs4 import BeautifulSoup         #For Text Parsing

%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

In [310…]
```
data = pd.read_csv('main.csv')
data = data[['Dates','News','PriceSentiment']]
print(data.shape)
data.head(7)
```

```
(9999, 3)
```

Out[310…]

	Dates	News	PriceSentiment
0	28/1/16	april gold down 20 cents to settle at $1,116.1...	negative
1	13/9/17	gold suffers third straight daily decline	negative
2	26/7/16	Gold futures edge up after two-session decline	positive
3	28/2/18	dent research : is gold's day in the sun comin...	none
4	6/9/17	Gold snaps three-day rally as Trump, lawmakers...	negative
5	16/8/16	Dec. gold climbs $9.40, or 0.7%, to settle at ...	positive
6	24/9/13	gold falls by rs 25 on sluggish demand, global...	negative

Now, show the data how looks like...

In [311…]
```
data.isnull().sum()
```

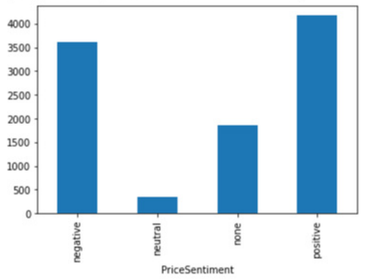
Out[311…]
```
Dates             0
News              0
PriceSentiment    0
dtype: int64
```

In [312…]
```
data=data.dropna()
data.isnull().sum()
```

Out[312…]
```
Dates             0
News              0
PriceSentiment    0
dtype: int64
```

In [313…]
```
Sentiment = data['PriceSentiment'].unique()
print(Sentiment)

data.groupby(data['PriceSentiment']).News.count().plot.bar(ylim=0)
plt.show()
```

```
['negative' 'positive' 'none' 'neutral']
```



ความก้าวหน้าการดำเนินงาน

In [314…
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/whatbest/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[314…
```
True
```

In [315…
```
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import re

stemmer = PorterStemmer()
words = stopwords.words("english")

data['processedtext'] = data['News'].apply(lambda x: " ".join([stemmer.stem(i) for i in re.sub("[^a-zA-Z]", " ", x).
```

In [316…
```
print(data.shape)
data.head(10)
```

```
(9999, 4)
```

Out[316…

	Dates	News	PriceSentiment	processedtext
0	28/1/16	april gold down 20 cents to settle at $1,116.1...	negative	april gold cent settl oz
1	13/9/17	gold suffers third straight daily decline	negative	gold suffer third straight daili declin
2	26/7/16	Gold futures edge up after two-session decline	positive	gold futur edg two session declin
3	28/2/18	dent research : is gold's day in the sun comin...	none	dent research gold day sun come soon
4	6/9/17	Gold snaps three-day rally as Trump, lawmakers...	negative	gold snap three day ralli trump lawmak reach d...
5	16/8/16	Dec. gold climbs $9.40, or 0.7%, to settle at ...	positive	dec gold climb settl oz
6	24/9/13	gold falls by rs 25 on sluggish demand, global...	negative	gold fall rs sluggish demand global cue
7	23/9/16	Gold futures fall for the session, but gain fo...	positive	gold futur fall session gain week
8	21/10/12	Gold struggles; silver slides, base metals falter	neutral	gold struggl silver slide base metal falter
9	16/3/18	april gold holds slight gain, up $2.50, or 0.2...	positive	april gold hold slight gain oz

Pre-process Data

We need to remove package name as it's not relevant. Then convert text to lowercase for CSV data. So, this is data pre-process stage.

In [317…
```
def preprocess_data(data):
    # Remove package name as it's not relevant
    #data = data.drop('News', axis=1)

    # Convert text to lowercase
    data['processedtext'] = data['processedtext'].str.strip().str.lower()
    return data
```

In [318…
```
data = preprocess_data(data)
```

Splitting Data

First, separate the columns into dependent and independent variables (or features and label). Then you split those variables into train and test set.

In [319…
```
df = data
# Split into training and testing data
x = data['processedtext']
y = data['PriceSentiment']
x, x_test, y, y_test = train_test_split(x,y, stratify=y, test_size=0.25, random_state=42)
```

Vectorize text reviews to numbers.

In [320…
```
# Vectorize text reviews to numbers
vec = CountVectorizer(stop_words='english')
x = vec.fit_transform(x).toarray()
x_test = vec.transform(x_test).toarray()
```

โดยใช้เทสในระบบ 25% และ เทรนในระบบ 75%

ความก้าวหน้าการดำเนินงาน

Model Generation

After splitting and vectorize text reviews to number, we will generate a random forest model on the training set and perform prediction on test set features.

In [321…
```
from sklearn.naive_bayes import MultinomialNB

model = MultinomialNB()
model.fit(x, y)
```

Out[321…
```
MultinomialNB()
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

ผลที่ได้คือความแม่นยำจาการเทสและเทรน 76%

```
['none'] [False]
Gold heading for worst week since November on rate hike worries
['negative'] [ True]
64.3859649122807
```

In [146…
```
model.predict(vec.transform(['Changes in non-farm payrolls increase.']))
```

Out[146…
```
array(['none'], dtype='<U8')
```

Average hourly earnings, m/m, remain unchanged. รายได้เฉลี่ยต่อชั่วโมง m/m ยังคงไม่เปลี่ยนแปลง The change in non-farm payrolls increased from the previous time. การเปลี่ยนแปลงในการจ้างงานนอกภาคเกษตรเพิ่มขึ้นจากครั้งก่อน lower unemployment rate อัตราการว่างงานลดลง

Evaluating Model

After model generation, check the accuracy using actual and predicted values.

In [322…
```
model.score(x_test, y_test)*100
```

Out[322…
```
75.92
```

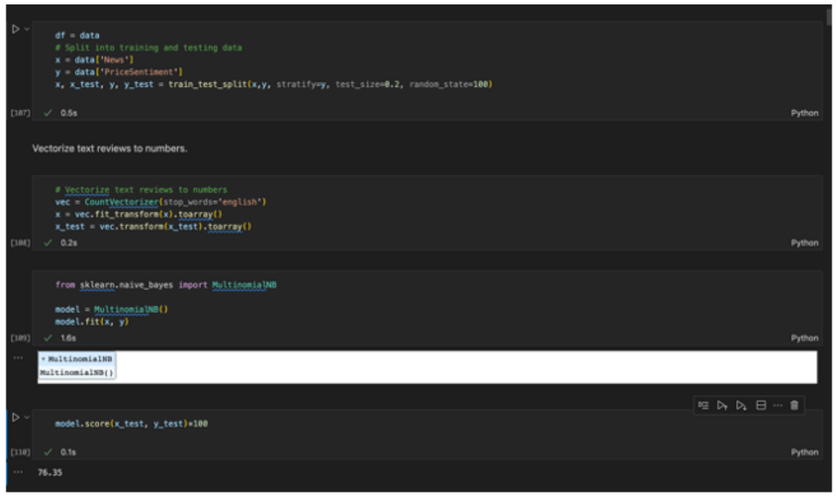
Then check prediction...

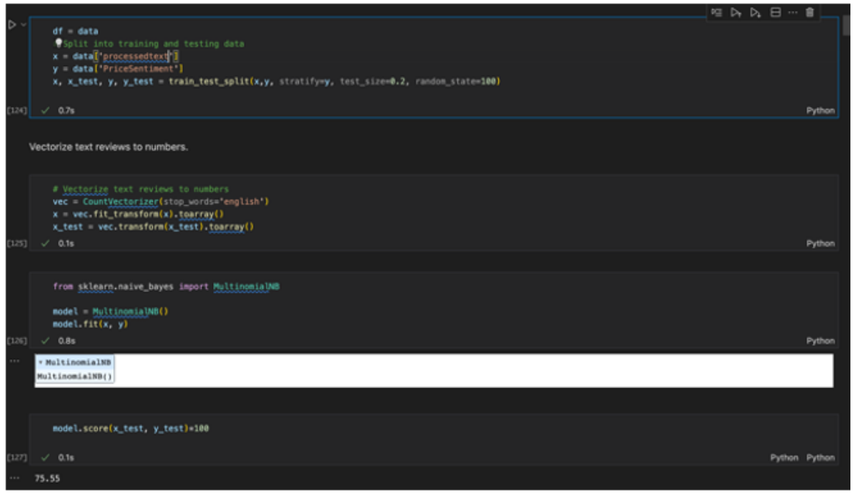
In [323…
```
from itertools import count
import pandas as pd
df = pd.read_csv('gold-dataset-sinha-khandait.csv', sep=',', header=None)
start = 10000
end = 10570
df = df[start - 1:end - 1]
correct = 0
for i in range(len(df)):
    print(df.values[i][2])
    print(model.predict(vec.transform([df.values[i][2]])),df.values[i][9] == model.predict(vec.transform([df.values[

    if df.values[i][9] == model.predict(vec.transform([df.values[i][2]])):
        correct += 1

print(correct / len(df) * 100 )
```

```
Kerala govt studying tax reduction on gold
['none'] [ True]
gold gets a festive lift, tops rs 31,000-mark
['positive'] [ True]
gold ends at new record high of $837.50 an ounce on nymex
['positive'] [ True]
Gold futures prices trade near the session's lows on Comex
['negative'] [ True]
gold settles at a more than 1-week high as dollar softens ahead of fed decision
['positive'] [ True]
Gold prices mark first climb in 7 sessions
['positive'] [ True]
Gold investment demand falls 34%, but WGC says it was still 'robust'
['negative'] [False]
gold futures end higher, but post a weekly decline
['positive'] [ True]
```

ความก้าวหน้าการดำเนินงาน

ความก้าวหน้า 7 การเปรียบเทียบ โดยใช้ Model Naïve baye





เป็นการเปรียบเทียบระหว่างใช้ NLP และ ไม่ใช้ NLP ของในระบบ

ความแม่นยำโดยไม่ใช้NLP 76.35%

ความแม่นยำโดยผ่านกระบวนการ NLP 75.55%

ความก้าวหน้าการดำเนินงาน

ความก้าวหน้า 7 การเปรียบเทียบ โดยใช้ Model Naïve baye

Then check prediction...

```
from itertools import count
import pandas as pd
df = pd.read_csv('gold-dataset-sinha-khandait.csv', sep=',', header=None)
start = 10000
end = 10570
df = df[start - 1:end - 1]
correct = 0
for i in range(len(df)):
    (df.values[i][2])
    (model.predict(vec.transform([df.values[i][2]])),df.values[i][9] == model.predict(vec.transform([df.values[i][2]])))

    if df.values[i][9] == model.predict(vec.transform([df.values[i][2]])):
        correct += 1

print(correct / len(df) * 100 )
```

[166] ✓ 0.5s

```
75.26315789473685
```

```
from itertools import count
import pandas as pd
df = pd.read_csv('gold-dataset-sinha-khandait.csv', sep=',', header=None)
start = 10000
end = 10570
df = df[start - 1:end - 1]
correct = 0
for i in range(len(df)):
    (df.values[i][2])
    (model.predict(vec.transform([df.values[i][2]])),df.values[i][9] == model.predict(vec.transform([df.values[i][2]])))

    if df.values[i][9] == model.predict(vec.transform([df.values[i][2]])):
        correct += 1

print(correct / len(df) * 100 )
```

[181] ✓ 0.5s

```
64.3859649122807
```

การทดสอบ โดยการนำข่าวสาร 570 record มาทดสอบจริง

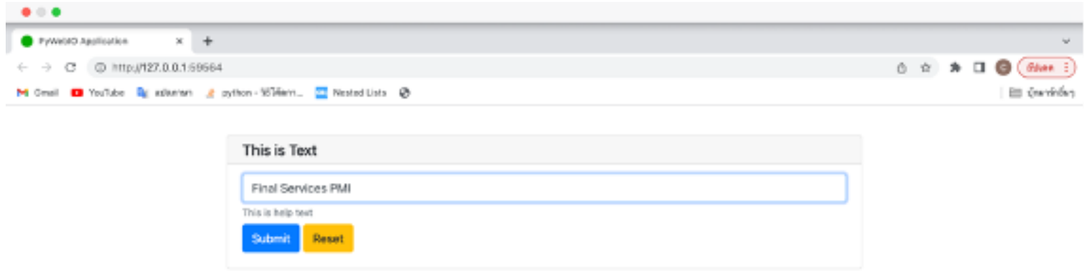
ความแม่นยำโดยผ่านกระบวนการ NLP 64%

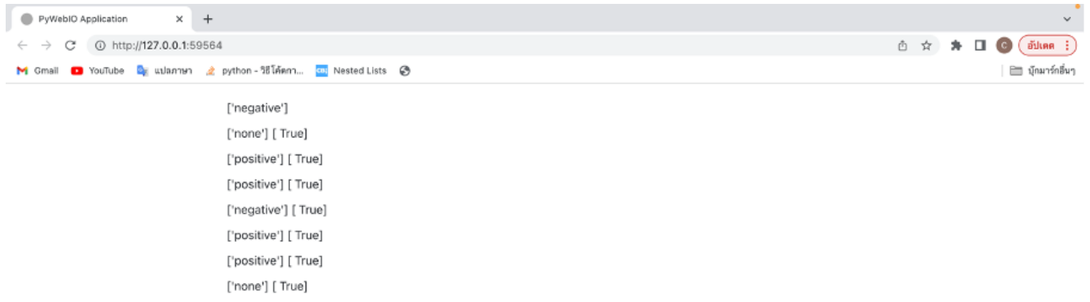
ความแม่นยำโดยไม่ใช้NLP 75.26%

ความก้าวหน้าการดำเนินงาน

ความก้าวหน้า 8 ทดลองทำ pywebio

```python
##pywebio##
from pywebio.input import input, FLOAT, TEXT
from pywebio.output import put_text
def main():
## Model Generation
    from sklearn.naive_bayes import MultinomialNB

    model = MultinomialNB()
    model.fit(x, y)
    from itertools import count
    import pandas as pd
    df = pd.read_csv('gold-dataset-sinha-khandait.csv', sep=',', header=None)
    start = 10000
    end = 10570
    df = df[start - 1:end - 1]
    correct = 0
    str = input('This is label', type=TEXT, placeholder='This is placeholder',
        help_text='This is help text', required=True)
    put_text(model.predict(vec.transform([str])))
    for i in range(len(df)):
        print(df.values[i][2])
        put_text(model.predict(vec.transform([df.values[i][2]])),df.values[i][9] == model.predict(vec.transform([df.values[i][2]])))

        if df.values[i][9] == model.predict(vec.transform([df.values[i][2]])):
            correct += 1

    print(correct / len(df) * 100 )


if __name__ == '__main__':
    main()
```





สรุป

สรุปผลการดำเนินงาน

1. สามารถสรุปบทความได้ ใช้ Library NLTK NLP
2. สามารถวิเคราะห์ บทความได้ระดับเบื้องต้น
3. จัดทำเหมืองข้อมูล Dataset
4. เปรียบเทียบอัลกอริทึมระหว่าง Random Forest Classifier กับ Naïve Bayes Classifier
5. ทำ Sentiment Analysis Naïve Bayes Classifier
6. ทำการเปรียบเทียบระหว่างข่าวสารดั้งเดิม และ ข่าวสารที่ผ่านระบบ NLP
7. มีการนำ pywebio เข้ามาในการทำ Web Application

ปัญหาและอุปสรรค

1. Library ที่จะใช้ค่อนข้างเยอะ จึงทำให้เริ่ม Project ช้า
2. ไม่สามารถนำ ตัวเลขที่ข่าวประกาศไปวิเคราห์ หรือ เก็บข้อมูลได้เนื่องจากตอนนี้การเทรนนิ่ง ต้องใช้ตัวอักษรภาษาอังกฤษเท่านั้น
3. ผลจากกาทำ Sentiment Analysis ผลที่ได้รับ ไม่ตรงกับที่เทสในระบบ
4. เนื่องจากข่าวสารที่ใช้ในการทดสอบ มีตัวเลขและบางประโยคสำคัญทำให้ NLP ตัดบางส่วนที่สำคัญไป

งานที่จะดำเนินการต่อไป

1. จัดทำ Web Applicationให้สมบูรณ์
2. ศึกษา Framework Django

เอกสารอ้างอิง

[1] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms (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 (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 (2014), In EMNLP. เข้าถึงล่าสุด 16 มกราคม 2565

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information (2016), arXiv preprint เข้าถึงล่าสุด 17 มกราคม 2565

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

เอกสารอ้างอิง

[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 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

คำถามและข้อเสนอแนะ


PRINCE OF
SONGKLA
UNIVERSITY