โครงงานทางวิศวกรรม

# Senior Project Proposal

เรื่อง

การวิเคราะห์ความเปลี่ยนแปลงในตลาดหุ้นโดยใช้เว็บบอร์ด

# Stock Market Sentiment Analysis using Webboard


โดย

นาย ชัชชนก อาศุเวทย์

อาจารย์ที่ปรึกษาโครงงาน

ผศ. ดร. ณัฐวุฒิ หนูไพโรจน์


รายงานนี้เป็นส่วนหนึ่งของการศึกษาวิชาโครงการวิศวกรรมคอมพิวเตอร์ หลักสูตรปริญญาวิศวกรรมศาสตร์บัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์จุฬาลงกรณ์มหาวิทยาลัย ปีการศึกษา 2560

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

โครงการวิศวกรรมคอมพิวเตอร์ ประจำปีการศึกษา 2560

**ชื่อโครงการ** (ภาษาไทย) การวิเคราะห์ความเปลี่ยนแปลงในตลาดหุ้นโดยใช้เว็บบอร์ด

(ภาษาอังกฤษ) Stock Market Sentiment Analysis using Webboard

| **ชื่อ-นามสกุล** | **เลขประจำตัว** | **ลายมือชื่อ** |
|---|---|---|
| นาย ชัชชนก อาศุเวทย์ | 573102202 | ..... |

| **อาจารย์ที่ปรึกษาโครงงาน** | **ลายมือชื่อ** |
|---|---|
| ผศ.ดร.ณัฐวุฒิ หนูไพโรจน์ | ..... |

# บทคัดย่อ

ทุกวันนี้หุ้นการเป็นสิ่งที่เข้าถึงได้ง่ายเนื่องจากมีเว็บไซต์ที่อำนวยความสะดวกในการเล่นหุ้นมากมายและมีหนังสือสอนการเล่นหุ้นตีพิมพ์ออกมามากอีกทั้งยังสามารถสืบค้นข้อมูลเกี่ยวกับหุ้นได้จากหลายเว็บไซต์ทำให้โซเชียลเน็ตเวิคได้เป็นที่ๆสำหรับแลกเปลี่ยนข้อมูลใหญ่ยกตัวอย่างเช่นเว็บไซต์พันทิพถือเป็นเวปที่มีการแลกเปลี่ยนกันของข้อมูลขนาดใหญ่ โดยได้แบ่งออกเป็นหลายๆห้องเช่นในห้องสินธรจะเป็นห้องที่พูดคุยกันเรื่องของตลาดหลักทรัพย์ ซึ่งถ้าหากสามารถวิเคราะห์ถึงคอมเม้นที่เอ่ยถึงชื่อหุ้นและนำมาวิเคราะห์ถึงความสัมพันธ์ระหว่างการพูดถึงและราคาหุ้น ในแต่ละวันว่าจะมีความสอดคล้องกันอย่างไรและจะนำข้อมูลที่ได้มาให้นักเศรษฐศาสตร์ทำการวิเคราะห์ต่อไป

# กิตติกรรมประกาศ

โครงงานการวิเคราะห์ความเปลี่ยนแปลงในตลาดหุ้นโดยใช้เว็บบอร์ด สามารถสำเร็จลุล่วงได้ ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจาก ผศ.ดร.ณัฐวุฒิ หนูไพโรจน์ อาจารย์ที่ปรึกษา ที่ได้กรุณาให้คำแนะนำ และตรวจสอบ แก้ไข ข้อบกพร่องทุกขั้นตอนของการจัดทำโครงงาน

ขอขอบคุณ ดร. วรประภา นาควัชนะ อาจารย์ คณะเศรษฐศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ค่อยให้ข้อมูลเกี่ยวกับตลาดหลักทรัพย์และค่อยแนะนำแนวทางในการทำโครงงาน

ขอขอบคุณภาควิชาวิศวกรรมศาสตร์คอมพิวเตอร์และคณาจารย์คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ที่มอบความรู้ทั้งทางด้านวิชา รวมไปถึงประสบการณ์ทางวิชาชีพตลอดระยะเวลาศึกษาทั้งสี่ปีที่ผ่านมา

ขอขอบพระคุณ บิดา มารดา และเพื่อน ตลอดจนผู้ที่เกี่ยวข้องทุกท่านที่ไม่ได้กล่าวนามไว้ณ ที่นี้ที่ได้ให้กำลังใจและมีส่วนช่วยเหลือให้โครงงานฉบับนี้สำเร็จลุล่วงได้ด้วยดี
ท้ายที่สุด ผู้จัดทำโครงงานหวังว่าโครงงานฉบับนี้จะเป็นประโยชน์กับผู้สนใจไม่มากก็น้อย

ชัชชนก อาศุเวทย์

**สารบัญ**

**สารบัญรูป**

# 1.บทนำ

## 1.1ที่มาและความสำคัญของปัญหา

ในปัจจุบัน โซเชียลเน็ตเวิคได้เป็นที่ๆสำหรับแลกเปลี่ยนข้อมูลใหญ่ยกตัวอย่างเช่นเว็บไซต์พันทิพถือเป็นเวปที่มีการแลกเปลี่ยนกันของข้อมูลขนาดใหญ่ โดยได้แบ่งออกเป็นหลายๆห้องเช่นในห้องสินธรจะเป็นห้องที่พูดคุยกันเรื่องของตลาดหลักทรัพย์ ซึ่งถ้าหากสามารถวิเคราะห์ถึงคอมเม้นที่เอ่ยถึงชื่อหุ้นและนำมาวิเคราะห์ถึงความสัมพันธ์ระหว่างการพูดถึงและราคาหุ้น ในแต่ละวันว่าจะมีความสอดคล้องกันอย่างไร
และนำข้อมูลที่ได้มาให้นักเศรษฐศาสตร์ทำการวิเคราะห์ต่อไป

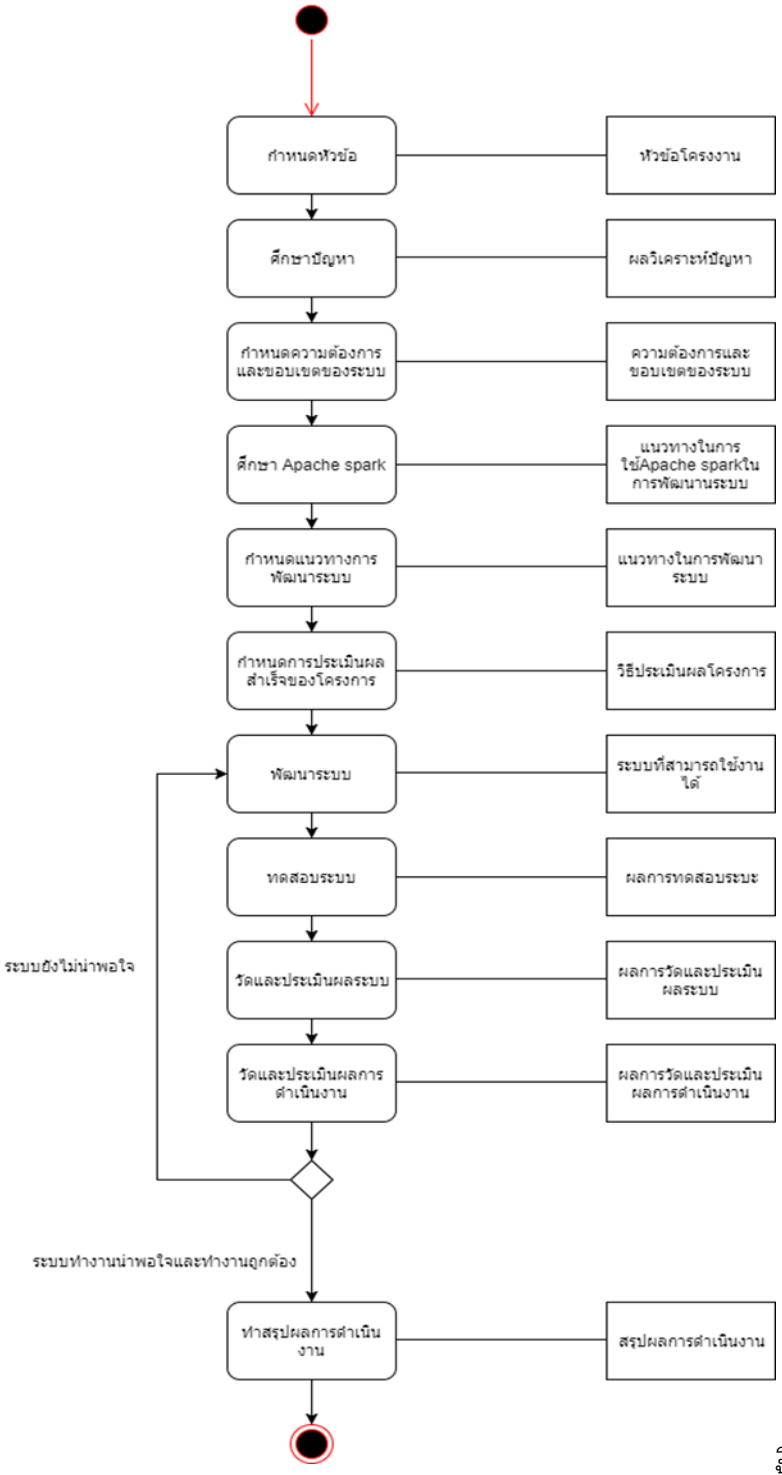
## 1.2วัตถุประสงค์

ในโครงงานนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบถึงจำนวนครั้งที่กล่าวถึงชื่อหุ้นจากคอมเม้นในห้องสินทรของเว็บไซต์pantipและนำมาเปรียบเทียบกับราคาขายจริงของหุ้นตัวนั้นโดยนำเสนอออกมาเป็นกราฟเพื่อให้นักเศรษฐศาสตร์ทำการวิเคราะห์ต่อไป

## 1.3ขอบเขตของโครงงาน

- ไฟล์ที่ใช้สำหรับนำเข้าเป็นไฟล์ข้อมูลของกระทู้ในพันทิพตั้งแต่วันที่ถึงเป็นไฟล์สกุล .json
- ไฟล์ข้อมูลชื่อหุ้น,ราคาขาย,วันที่ขาย ฯลฯ จากsetsmartเป็นไฟล์สกุล .csv

# 1.4ขั้นตอนการดำเนินงาน




รูปที่1.4ขั้นตอนการดำเนินงาน

# 1.5ประโยชน์ที่คาดว่าจะได้รับ

- เก็บเป็นข้อมูลทางสถิติเพื่อให้นักเศรษฐศาสตร์นำไปวิเคราะห์ต่อในเรื่องความสัมพันธ์ระหว่างความถี่ที่ปรากฏของชื่อหุ้นที่กล่าวถึงและการเปลี่ยนแปลงที่เกิดขึ้นจริงของตัวหุ้น

# 2.ความรู้, ทฤษฎีและเทคโนโลยีที่เกี่ยวข้อง

## 2.1 Apache spark

เป็นเทคโนโลยีในการประมวลข้อมูลขนาดใหญ่ โดยสามารถจะประมวลผลข้อมูลทั้งที่อยู่ใน HDFS หรือแหล่งอื่นๆ อาทิเช่น Cloud Storage, NoSQL, RDBMS ทั้งนี้ Spark สามารถทำงานแบบ Standalone หรือจะทำงานบน Hadoop Cluster ผ่าน YARN ซึ่ง Spark มีบริการประมวลผลแบบต่างๆ ดังนี้

- **Spark core** ก็คือระบบประมวลผลโดยผ่าน API ซึ่งให้ผู้ใช้บริการสามารถเลือกใช้ภาษา Java, Scala, Python หรือ R
- **Spark streaming** สำหรับการประมวลผลแบบ Realtime Streaming
- **Spark SQL** สำหรับการประมวลผลที่ใช้ภาษาคล้ายกับ SQL
- **MLlib** สำหรับการประมวลที่เป็นแบบ Machine Learning

## 2.2 Pantip

เป็นเว็บไซต์ไทยที่ให้บริการเว็บบอร์ดของไทยที่มีชื่อเสียง มีห้องสนทนาหลายเรื่อง ที่สมาชิกสามารถตั้งกระทู้เพื่อให้สมาชิกคนอื่นเข้ามาตอบกระทู้ได้ ปัจจุบันพันทิปมีห้องทั้งหมด 37 ห้อง

## 2.3JSON(JavaScript Object Notation)

คือรูปแบบของข้อมูลที่สำหรับแลกเปลี่ยนข้อมูลที่มีขนาดเล็กซึ่งคนสามารถทำความเข้าใจได้ง่ายและสามารถถูกสร้างและอ่านโดยเครื่องได้ง่าย มักถูกกำหนดภายใต้ภาษา JavaScript JSON เป็นรูปแบบข้อมูลตัวอักษรที่มีความเป็นอิสระอย่างสมบูรณ์ แต่จะมีหลักการการเขียนที่

คุ้นเคยกับนักเขียนโปรแกรมภาษาต่างๆได้ ไม่ว่าจะเป็น ภาษา C,C++C#, Java , Javascript , Perl , Python และอื่นๆ

## 2.4jupyter notebook

Web-base ในการเขียนcodeภาษาpythonสำหรับโครงการนี้

## 2.5pandas,bokeh

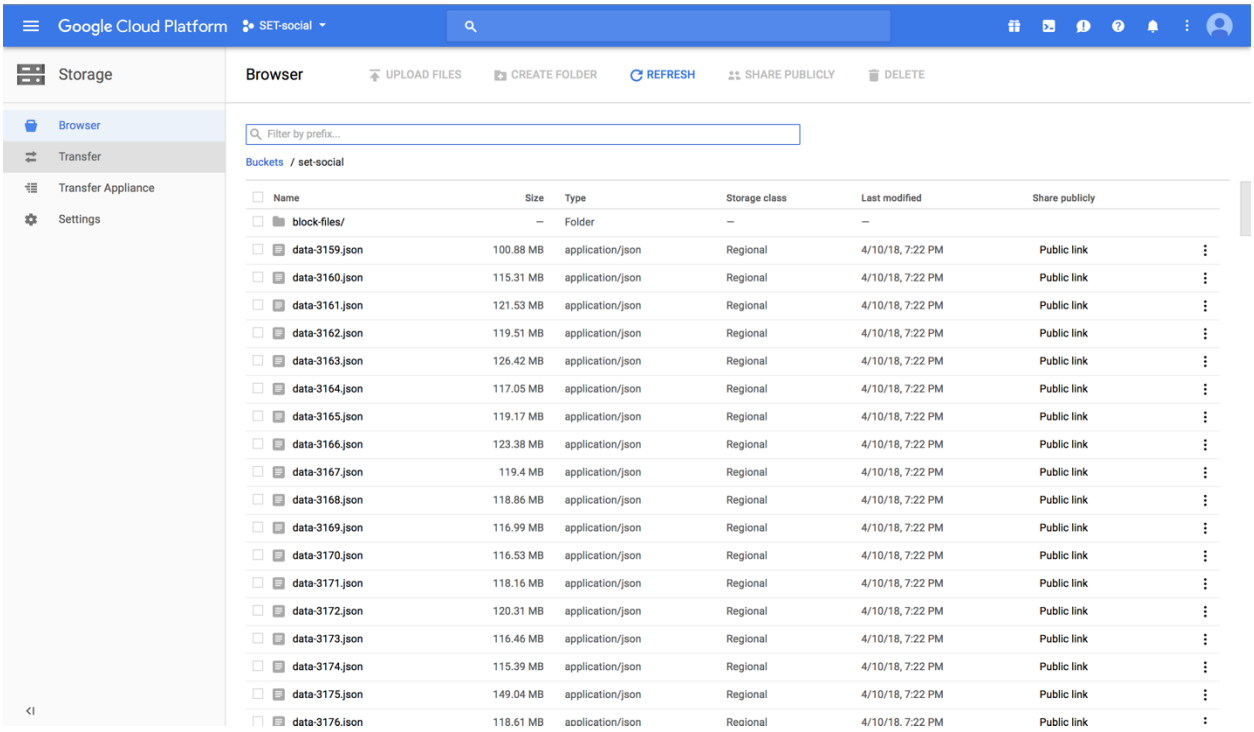
libraryที่ทำงานบนpython สำหรับทำโครงงานนี้โดยจะใช้pandasในการจัดการกับdataframeและใช้Bokehในการทำ data visualization

# บทที่3.การออกแบบและการพัฒนา

## 3.1การกำหนดdata set

Dataที่เป็นinputสำหรับโครงการนี้จะมาจาก3แหล่งด้วยกันคือ

dataของกระทู้pantipระหว่างวันที่ถึงวันที่ โดย ผศ.ดร.ณัฐวุฒิ หนูไพโรจน์ได้ดึงมาให้เก็บเป็นนามสกุล.json ซึ่งจะเก็บไว้ในbucketของgooglecloudเป็นจำนวน120ไฟล์ โดยไฟล์นึงจะมี10000กระทู้ ข้อมูลจากpantipนี้เป็นข้อมูลปี2014




รูปที่3.1.1input_datasetจากgoogle cloud

อีกdataหนึ่งคือข้อมูลของsetsmartเป็นไฟล์นามสกุล.csvซึ่งจะประกอบไปด้วย['D_TRADE',
'I_SECURITY', 'I_MARKET', 'I_INDUSTRY', 'I_SECTOR','I_SUBSECTOR', 'Z_PRIOR
', 'P_YIELD_PRIOR', 'Z_OPEN', 'P_YIELD_OPEN','Z_HIGH', 'P_YIELD_HIGH',
'Z_LOW', 'P_YIELD_LOW', 'Z_CLOSE','P_YIELD_CLOSE', 'P_YIELD_CHG', 'Z_ACR_I
NTEREST', 'Z_LAST_BID','P_YIELD_LAST_BID', 'Z_LAST_OFFER', 'P_YIELD_LAST_O
FFER', 'Q_TRANS','Q_VOLUME', 'M_VALUE', 'Q_TRANS_TR', 'Q_VOLUME_TR', 'M_VA
LUE_TR','Q_TRANS_ODD', 'Q_VOLUME_ODD', 'M_VALUE_ODD', 'M_MKT_CAP', 'D_AS_O
F',
'R_PE', 'R_PB', 'M_BOOK_VALUE', 'P_DVD_YIELD', 'Z_PAR','Q_SHARE_LISTED',
'Q_TOTAL_VOLUME', 'M_TOTAL_VALUE', 'R_TURNOVER','N_STATUS', 'N_BENEFIT',
'R_BETA', 'R_ROI', 'R_ADJUST_FACTOR','D_PRIOR', 'R_PRIOR_ADJUST_FACTOR',
'R_EPS', 'D_EARNING', 'P_CHANGE','Q_AVG_VOLUME', 'M_AVG_VALUE', 'R_IV',
'R_DELTA']

โดยเป็นข้อมูลของปี2014เช่นกันโดยแต่ละparameterมีความหมายดังนี้

## 2.30 DAILY_SECURITY_TRADING

**Description** ข้อมูลการซื้อขายและสถิติของหลักทรัพย์ สรุปเป็นรายวัน

| Field Name | Description | Type | Length | Key/Index | Null | Possible Value |
|---|---|---|---|---|---|---|
| D_TRADE | วันที่ทำการซื้อขาย | DATE | | 1 | N | dd/mm/yyyy |
| I_SECURITY | รหัสหลักทรัพย์ | INTEGER | | 2 | N | |
| ~~I_BOARD~~ | ~~รหัสประเภทกระดาน~~ | ~~CHAR~~ | | | ~~Y~~ | ~~M/F~~ |
| I_MARKET | รหัสตลาด | CHAR | | | Y | |
| I_INDUSTRY | รหัสกลุ่มอุตสาหกรรม | SMALLINT | | | Y | |
| I_SECTOR | รหัสหมวดอุตสาหกรรม | SMALLINT | | | Y | |
| I_SUBSECTOR | รหัสหมวดอุตสาหกรรมย่อย | SMALLINT | | | Y | |
| Z_PRIOR | ราคา ณ สิ้นวันก่อน | DECIMAL | 8, 2 | | Y | |
| P_YIELD_PRIOR | อัตราผลตอบแทน ณ ราคาปิดวันก่อนหน้า | FLOAT | | | Y | 3.6 |
| Z_OPEN | ราคาเปิด | DECIMAL | 8, 2 | | Y | |
| P_YIELD_OPEN | อัตราผลตอบแทน ณ ราคาเปิดของหลักทรัพย์ | FLOAT | | | Y | 3.6 |
| Z_HIGH | ราคาสูงสุด | DECIMAL | 8, 2 | | Y | |
| P_YIELD_HIGH | อัตราผลตอบแทน ณ ราคาสูงสุดของหลักทรัพย์ | FLOAT | | | Y | 3.6 |
| Z_LOW | ราคาต่ำสุด | DECIMAL | 8, 2 | | Y | |
| P_YIELD_LOW | อัตราผลตอบแทน ณ ราคาต่ำสุดของหลักทรัพย์ | FLOAT | | | Y | 3.6 |
| Z_CLOSE | ราคาปิด | DECIMAL | 8, 2 | | Y | |
| P_YIELD_CLOSE | อัตราผลตอบแทน ณ ราคาปิดของหลักทรัพย์ | FLOAT | | | Y | 3.6 |
| P_YIELD_CHG | อัตราผลตอบแทน ณ ราคาเสนอขายครั้งสุดท้าย | FLOAT | | | Y | 3.6 |
| Z_ACR_INTEREST | ดอกเบี้ยค้างรับ ต่อหน่วย ( บาท ) | DOUBLE | | | Y | |
| Z_LAST_BID | ราคาเสนอซื้อครั้งสุดท้าย | DECIMAL | 8, 2 | | Y | |

| | | | | | |
|---|---|---|---|---|---|
| P_YIELD_LAST_BID | การเปลี่ยนแปลงของอัตราผลตอบแทน | FLOAT | | Y | |
| Z_LAST_OFFER | ราคาเสนอขายครั้งสุดท้าย | DECIMAL | 8, 2 | Y | |
| P_YIELD_LAST_OFFER | อัตราผลตอบแทน ณ ราคาเสนอซื้อครั้งสุดท้าย | FLOAT | | Y | 3.6 |
| Q_TRANS | จำนวนรายการซื้อขายแบบ AOM | BIGINT | | Y | |
| Q_VOLUME | จำนวนหุ้นซื้อขายแบบ AOM | BIGINT | | Y | |
| M_VALUE | มูลค่าการซื้อขายแบบ AOM | DOUBLE | | Y | |
| Q_TRANS_TR | จำนวนรายการซื้อขายแบบ TRADE REPORT | BIGINT | | Y | |
| Q_VOLUME_ TR | จำนวนหุ้นซื้อขายแบบ TRADE REPORT | BIGINT | | Y | |
| M_VALUE_ TR | มูลค่าการซื้อขายแบบ TRADE REPORT | DOUBLE | | Y | |
| Q_TRANS_ODD | จำนวนรายการซื้อขายแบบ Odd Lot | BIGINT | | Y | |
| Q_VOLUME_OD D | จำนวนหุ้นซื้อขายแบบ Odd Lot | BIGINT | | Y | |
| M_VALUE_ODD | มูลค่าการซื้อขายแบบ Odd Lot | DOUBLE | | Y | |
| M_MKT_CAP | Market Capitalization | DOUBLE | | Y | |
| D_AS_OF | วันที่ใช้คำนวณ EPS | DATE | | Y | |
| R_PE | P/E Ratio | DOUBLE | | Y | |
| R_PB | P/BV Ratio | DOUBLE | | Y | |
| M_BOOK_VALUE | Book Value | DOUBLE | | | |
| P_DVD_YIELD | Dividend yield | REAL | | Y | |
| Z_PAR | ราคาพาร์ | DECIMAL | 16,5 | Y | |
| Q_SHARE_LISTE D | จำนวนหุ้นที่จดทะเบียนกับตลาดหลักทรัพย์ | BIGINT | | Y | |
| Q_TOTAL_VOLU ME | ปริมาณการซื้อขายรวมทุกวิธีการซื้อขายของหุ้น local และ foreign | BIGINT | | Y | |

| | | | | | |
|---|---|---|---|---|---|
| M_TOTAL_VALUE | มูลค่าการซื้อขายรวมทุกวิธีการซื้อขายของหุ้น local และ foreign | DOUBLE | | Y | |
| R_TURNOVER | อัตราการหมุนเวียนซื้อขาย | REAL | | Y | |
| N_STATUS | ชื่อเครื่องหมาย | CHAR | 2 | Y | NP, NR, SP, ST, CM |
| N_BENEFIT | ชื่อย่อเครื่องหมายผลประโยชน์ | CHAR | 2 | Y | |
| R_BETA | ข้อมูล Beta ของหลักทรัพย์ | DECIMAL | 8, 2 | Y | *Data currently not available* |
| R_ROI | ข้อมูล Return on Investment ของหลักทรัพย์ | DECIMAL | 8, 2 | Y | *Data currently not available* |
| R_ADJUST_FACTOR | อัตราส่วนการปรับปรุงราคา | DOUBLE | | Y | |
| D_PRIOR | วันที่เกิดราคาก่อนหน้า | DATE | | Y | |
| R_PRIOR_ADJUST_FACTOR | อัตราส่วนการปรับปรุงราคาของราคาก่อนหน้า | DOUBLE | | Y | |
| R_EPS | กำไรสุทธิต่อหุ้น | DECIMAL | 12,5 | Y | |
| D_EARNING | วันที่ตามมูลค่ากำไร/ขาดทุน สุทธิ (ปี ค.ศ.) | DATETIME | | Y | |
| R_IV | Implied Volatility (for DW) | DECIMAL | 8,2 | | *Data currently not available* |
| R_DELTA | Delta (for DW) | DECIMAL | | | *Data currently not available* |

รูปที่3.1.2 รูปข้อมูลdaily_security_trading

สังเกตุว่าข้อมูลจากsetsmartจะไม่มีชื่อหุ้นแต่จะใช้i_securityแทนในการระบุถึงตัวหุ้นจึงทำให้มีอีก

inputหนึ่งคือsecurity.csvซึ่งจะประกอบไปด้วย 'I_SECURITY', 'I_COMPANY', 'I_MARKET', 'I_INDUSTRY', 'I_SECTOR','I_SUBSECTOR', 'I_SEC_TYPE', 'N_SECURITY', 'N_SECURITY_T','N_SECURITY_E', 'I_SECURITY_LOCAL', 'I_ISIN', 'I_NATIVE']

โดยแต่ละparameterมีความหมายดังนี้

### 2.97 SECURITY

**Description** ข้อมูลหลักทรัพย์

| Field Name | Description | Type | Length | Key/Index | Null | Possible Value |
|---|---|---|---|---|---|---|
| I_security | รหัสหลักทรัพย์ | Integer | | 1 | N | |
| I_company | รหัสบริษัท | Smallint | | FK | N | |
| I_market | รหัสตลาด | Char | 1 | FK | N | A, S, B,O |
| I_industry | รหัสกลุ่มอุตสาหกรรม | Smallint | | FK | Y | |
| I_sector | รหัสหมวดอุตสาหกรรม | Smallint | | FK | Y | |
| I_subsector | รหัสหมวดอุตสาหกรรมย่อย | Smallint | | FK | Y | |
| I_sec_type | รหัสประเภทของหลักทรัพย์ | Char | 1 | | Y | |
| N_security | ชื่อย่อหลักทรัพย์ | Char | 20 | | Y | |
| N_security_t | ชื่อขยายหลักทรัพย์ (ไทย) | Varchar | 60 | | Y | |
| N_security_e | ชื่อขยายหลักทรัพย์ (อังกฤษ) | Varchar | 60 | | Y | |
| I_security_local | รหัสหลักทรัพย์ Local ของหุ้นนั้น | Integer | | | Y | |
| I_isin | รหัส ISIN Code | Char | 14 | | Y | |
| I_native | ภูมิลำเนา | Char | 1 | | Y | L, F, U, R |

รูปที่3.1.3 รูปข้อมูลหลักทรัพย์

## 3.2การเตรียมการข้อมูล

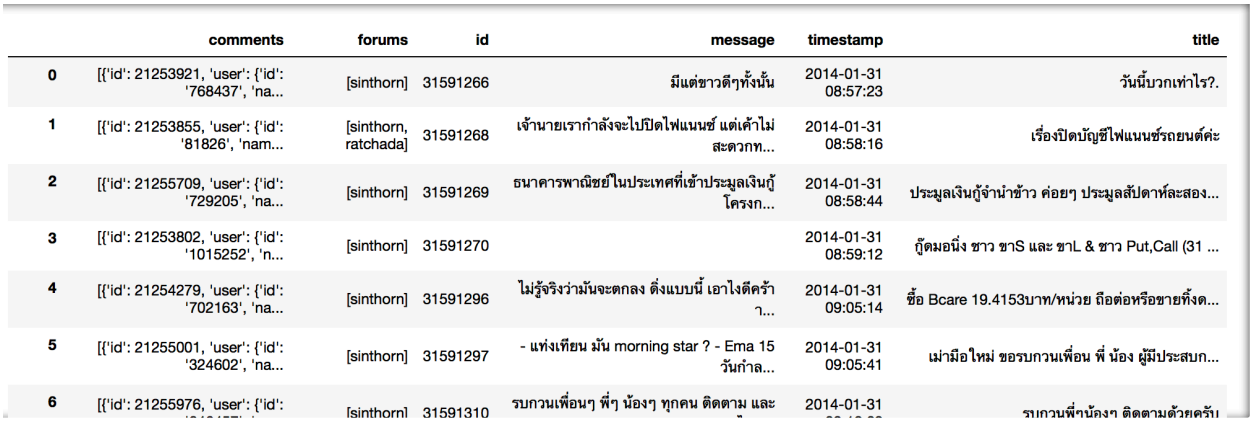
### 3.2.1ใช้sparkเพื่อดูschemaของข้อมูลจากpantip รูปพันทิพ

```
root
 |-- comments: array (nullable = true)
 |     |-- element: struct (containsNull = true)
 |     |     |-- commentNumber: string (nullable = true)
 |     |     |-- id: long (nullable = true)
 |     |     |-- message: string (nullable = true)
 |     |     |-- replies: array (nullable = true)
 |     |     |     |-- element: struct (containsNull = true)
 |     |     |     |     |-- id: long (nullable = true)
 |     |     |     |     |-- message: string (nullable = true)
 |     |     |     |     |-- replyNumber: string (nullable = true)
 |     |     |     |     |-- timestamp: long (nullable = true)
 |     |     |     |     |-- timestampISO: string (nullable = true)
 |     |     |     |     |-- user: struct (nullable = true)
 |     |     |     |     |     |-- id: string (nullable = true)
 |     |     |     |     |     |-- ip: string (nullable = true)
 |     |     |     |     |     |-- name: string (nullable = true)
 |     |     |-- timestamp: long (nullable = true)
 |     |     |-- timestampISO: string (nullable = true)
 |     |     |-- user: struct (nullable = true)
 |     |     |     |-- id: string (nullable = true)
 |     |     |     |-- ip: string (nullable = true)
 |     |     |     |-- name: string (nullable = true)
```

รูปที่3.2.1 schemaของข้อมูลจากpantip

จะเห็นว่ากระทู้pantipจะเก็บข้อมูลเป็นลำดับชั้นที่ภายในกระทู้นึงจะมีหลาย commentและภายในcommentนึงก็จะมีหลายreplyและถ้าหากหากต้องการที่จะนับจำนวนครั้งที่กล่าวถึงชื่อหุ้นแต่ละตัวนั้นจะจากนั้นจะนับจากtitleของกระทู้commentและ reply

3.2.2จากนั้นทำการอ่านข้อมูล.jsonมาที่ละไฟล์แล้วจะทำการfilterข้อมูลกระทู้pantipทั้งหมดและเลือกมาเฉพาะกระทู้ที่อยู่ในห้องsinthornเท่านั้นจากนั้นใช้pandasจัดให้เป็นdataframeดังรูป



| | comments | forums | id | message | timestamp | title |
|---|---|---|---|---|---|---|
| 0 | [{'id': 21253921, 'user': {'id': '768437', 'na... | [sinthorn] | 31591266 | มีแต่ขาวดีๆทั้งนั้น | 2014-01-31 08:57:23 | วันนี้บวกเท่าไร?. |
| 1 | [{'id': 21253855, 'user': {'id': '81826', 'nam... | [sinthorn, ratchada] | 31591268 | เจ้านายเรากำลังจะไปปิดไฟแนนซ์ แต่เค้าไม่สะดวกท... | 2014-01-31 08:58:16 | เรื่องปิดบัญชีไฟแนนซ์รถยนต์ค่ะ |
| 2 | [{'id': 21255709, 'user': {'id': '729205', 'na... | [sinthorn] | 31591269 | ธนาคารพาณิชย์ในประเทศที่เข้าประมูลเงินกู้โครงก... | 2014-01-31 08:58:44 | ประมูลเงินกู้จำนำข้าว ค่อยๆ ประมูลสัปดาห์ละสอง... |
| 3 | [{'id': 21253802, 'user': {'id': '1015252', 'n... | [sinthorn] | 31591270 | | 2014-01-31 08:59:12 | กู๊ดมอนิ่ง ชาว ชาS และ ชาL & ชาว Put,Call (31 ... |
| 4 | [{'id': 21254279, 'user': {'id': '702163', 'na... | [sinthorn] | 31591296 | ไม่รู้จริงว่ามันจะตกลง ดิ่งแบบนี้ เอาไงดีคร้าา... | 2014-01-31 09:05:14 | ซื้อ Bcare 19.4153บาท/หน่วย ถือต่อหรือขายทิ้งด... |
| 5 | [{'id': 21255001, 'user': {'id': '324602', 'na... | [sinthorn] | 31591297 | - แท่งเทียน มัน morning star ? - Ema 15 วันกำล... | 2014-01-31 09:05:41 | เม่ามือใหม่ ขอรบกวนเพื่อน พี่ น้อง ผู้มีประสบก... |
| 6 | [{'id': 21255976, 'user': {'id': | [sinthorn] | 31591310 | รบกวนเพื่อนๆ พี่ๆ น้องๆ ทุกคน ติดตาม และ | 2014-01-31 | รบกวนพี่ๆน้องๆ ติดตามด้วยครับ |

รูปที่3.2.2 dataframe pantipเฉพาะห้องsinthorn

จากรูปนั้นในแต่ละrowก็คือ1กระทู้(messageคือcommentของเจ้าของกระทู้เนื่องจากpantipไม่สามารถตั้งแค่ชื่อกระทู้ได้จำเป็นต้องเขียนcommentด้วย)

3.2.3ทำการfilterอีกรอบโดยเอามาแค่commentและreplyที่ตอบกระทู้ภายในวันเดียวกันเท่านั้น(เหตุผลทางเศรษฐศาสตร์เพราะว่าหากมีบางreplyที่มาตอบกระทู้ที่เก่ามากๆจะทำให้ข้อมูลผิดเพี้ยน)พร้อมทั้งระบุวันที่ด้วยว่าแต่ละกระทู้,comment,replyนั้นเป็นของวันไหน

| | ID | Timestamp | flag | text |
|---|---|---|---|---|
| 0 | 31591266 | 2014-01-31 08:57:23 | title | วันนี้บวกเท่าไร?. |
| 1 | 31591266 | 2014-01-31 08:57:23 | message | มีแต่ขาวดีๆทั้งนั้น |
| 2 | 21253921 | 2014-01-31 09:02:41 | comment | ต่างชาติขายทุกวัน เราหันมาเล่นตัวเล็กดีกว่า ให... |
| 3 | 21253924 | 2014-01-31 09:02:46 | comment | +ซัก5จุด AOT ก็คง+ไป20บาทแล้วมั้งขนาดลงทุกวันม... |
| 4 | 5481404 | 2014-01-31 09:04:22 | reply | AOT น่าจะพี่กองกับปอปลาก |
| 5 | 5481456 | 2014-01-31 09:10:46 | reply | เกรง AOT จะเปลี่ยนอาชีพไปเป็นชาวสวน |
| 6 | 21254006 | 2014-01-31 09:05:18 | comment | เม่ายังไม่มั่นใจ <br />\n   &nb... |
| 7 | 21254119 | 2014-01-31 09:08:39 | comment | บวกเช้า ลบ 4 โมงเย็น เม่า+ปอป ขายกลัววันที่ 2 |
| 8 | 21254834 | 2014-01-31 09:27:11 | comment | ระวังปอปลากไปกินตับ<br />\n<img class="img-in-... |
| 9 | 21254870 | 2014-01-31 09:27:55 | comment | ตลาดเอเชียปิดกันเยอะ แถมยังกังวลทั้งปัจจัยภายใ... |

รูปที่3.2.3dataframeผ่านการfilterแล้ว

3.2.4ทำการรับชื่อหุ้นที่สนใจแล้วนำไปค้นในdataframeจากข้อที่แล้วและจัดเก็บเป็นอีกdataframeนึง โดยแสดงเป็นวันที่ และข้อความที่มีชื่อหุ้นปรากฏ

รูปที่3.2.4.1dataframeที่filterเอามาเฉพาะหุ้นที่สนใจ

| | date | text | type |
|---|---|---|---|
| 0 | 2014-01-31 13:44:15 | เทคนิคการวาง put เพื่อลดความเสียหาย: มีหุ้น PT... | title |
| 1 | 2014-01-31 13:44:15 | ผมมีหุ้น PTT ในพอร์ตมูลค่า 686,000 บาท ทุน 274... | message |
| 2 | 2014-01-31 13:59:17 | ไม่ควรไปเสี่ยงเพิ่มนะครับ เพราะมันมีอายุจำกัด<... | comment |
| 3 | 2014-01-31 14:20:56 | อย่าเล่น dw เด็ดขาดนะครับ<br />\n<br />\ndw จะ... | comment |
| 4 | 2014-01-31 14:54:40 | จากที่อ่านผมคิดว่า <br />\n<br />\nคุณเป็นประเ... | comment |
| 5 | 2014-01-31 16:04:24 | PTT มีข่าวไรปะคับแชรง | title |
| 6 | 2014-02-01 10:04:22 | เลขคณิตคิดสนุก ดูจาก set.or.th นั่นแหละฟังเขาเ... | comment |
| 7 | 2014-02-01 15:02:41 | วันนี้เป็นวันหวยออก นึกถึงสุภาษิตที่ว่า"คนจนเล... | message |
| 8 | 2014-02-01 21:33:13 | หุ้น PTT Goodddddd | title |
| 9 | 2014-02-01 21:33:13 | อาทิตย์หน้ารีบสะสมกันเลยเพราะรถไฟไกล้จะออกแล้ว... | message |
| 24 | 2014-02-03 21:04:47 | ซื้อหุ้น PTT ครับ ปันผลเดือนหน้า | comment |
| 25 | 2014-02-03 20:46:10 | ขอถามผู้รู้ในห้องนี้นะครับ วันนี้ ตั้งซื้อ ตอน... | message |
| 26 | 2014-02-03 22:29:27 | ซื้อ PTT ครับ รอวิ่งยาว | comment |
| 27 | 2014-02-03 23:45:16 | กลยุทธ์ส่วนตัว:<br />\n<br />\nS เพื่อ Closed ... | comment |
| 28 | 2014-02-03 23:52:48 | ตอนนี้มีอยู่ 5 ตัว GYT PTT TTW EA ... | message |
| 29 | 2014-02-04 21:04:36 | ptt pttgc ค่ะ | comment |
| ... | ... | ... | ... |
| 2701 | 2014-10-29 09:51:51 | <b>โบรกฯชู STPI หุ้นทรงสวย อัพไซด์รอบนี้ไม่ต่ำ... | comment |
| 2702 | 2014-10-29 12:57:00 | สรุปก็คือ  เปนราคา จริงๆแท้ๆตามโลกคว... | comment |
| 2703 | 2014-10-29 15:36:08 | กองทุน ลากมาขาย ด้วยครับ<br />\nวันนี้ ขาย kba... | reply |
| 2704 | 2014-10-29 21:04:36 | หลังจากที่ Pttgc ได้กลับตัวจากจุดต่ำสุดที่ 56.... | message |

ทำการจัดกลุ่มตามวันเพื่อดูว่าในแต่ละวันนับจำนวนครั้งที่พูดถึงชื่อหุ้นได้กี่ครั้ง

| | count | date |
|---|---|---|
| **0** | 6 | 2014-1-31 |
| **1** | 4 | 2014-2-1 |
| **2** | 7 | 2014-2-2 |
| **3** | 12 | 2014-2-3 |
| **4** | 1 | 2014-2-4 |
| **5** | 6 | 2014-2-5 |
| **6** | 8 | 2014-2-6 |
| **7** | 41 | 2014-2-7 |
| **8** | 12 | 2014-2-8 |
| **9** | 5 | 2014-2-9 |
| **10** | 18 | 2014-2-10 |
| **11** | 9 | 2014-2-11 |
| **12** | 6 | 2014-2-12 |
| **13** | 5 | 2014-2-13 |
| **14** | 1 | 2014-2-15 |
| **15** | 2 | 2014-2-16 |
| **16** | 13 | 2014-2-17 |
| **17** | 10 | 2014-2-18 |
| **18** | 11 | 2014-2-19 |
| **19** | 6 | 2014-2-20 |
| **20** | 10 | 2014-2-21 |

รูปที่3.2.4.2dataframeที่จัดกลุ่มตามวันและจำนวนครั้งที่หุ้นปรากฏ

3.2.5ส่วนข้อมูลจากsetsmartจะfilterจากparameter,ชื่อหุ้นที่สนใจและใช้pandasจัดให้เป็นdataframeเช่นกัน(จากรูปด้านล่างสนใจM_TOTAL_VALUE)

| | Date | M_TOTAL_VALUE |
|---|---|---|
| 0 | 3/3/2014 00:00 | 1.795527e+09 |
| 1 | 4/3/2014 00:00 | 9.343916e+08 |
| 2 | 2/1/2014 00:00 | 9.693125e+08 |
| 3 | 3/1/2014 00:00 | 1.239200e+09 |
| 4 | 5/6/2014 00:00 | 1.524353e+09 |
| 5 | 6/6/2014 00:00 | 8.038056e+08 |
| 6 | 6/1/2014 00:00 | 1.639226e+09 |
| 7 | 5/3/2014 00:00 | 1.506905e+09 |
| 8 | 9/1/2014 00:00 | 2.058637e+09 |
| 9 | 6/3/2014 00:00 | 1.124174e+09 |
| 10 | 8/1/2014 00:00 | 1.886561e+09 |
| 11 | 7/3/2014 00:00 | 8.891363e+08 |
| 12 | 10/3/2014 00:00 | 8.397985e+08 |
| 13 | 11/3/2014 00:00 | 1.558645e+09 |
| 14 | 10/6/2014 00:00 | 7.250693e+08 |
| 15 | 7/1/2014 00:00 | 2.117638e+09 |
| 16 | 9/6/2014 00:00 | 5.658084e+08 |
| 17 | 14/1/2014 00:00 | 1.065173e+09 |
| 18 | 16/1/2014 00:00 | 1.493811e+09 |
| 19 | 12/6/2014 00:00 | 6.714811e+08 |
| 20 | 13/1/2014 00:00 | 1.071964e+09 |
| 21 | 14/3/2014 00:00 | 1.286967e+09 |
| 22 | 11/6/2014 00:00 | 6.435414e+08 |

รูปที่3.2.5dataframeจากsetsmart

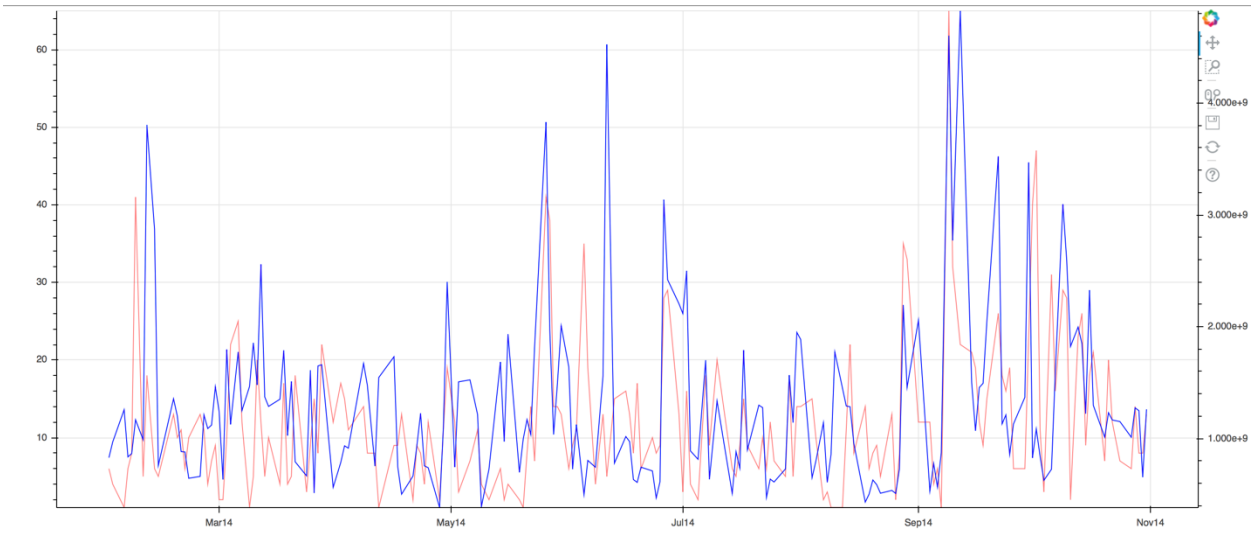
3.2.6ทำการรวมdataframeจากpantipและsetsmartเข้าด้วยกันเพื่อดูว่าในแต่ละวันนั้นมีการพูดถึงชื่อหุ้นนี้กี่ครั้งละราคาในตลาดจริงเป็นเท่าไหร่

| | count | date | M_TOTAL_VALUE |
|---|---|---|---|
| **0** | 6 | 2014-1-31 | 8.284117e+08 |
| **1** | 4 | 2014-2-1 | 9.693125e+08 |
| **2** | 1 | 2014-2-4 | 1.257528e+09 |
| **3** | 6 | 2014-2-5 | 8.354520e+08 |
| **4** | 8 | 2014-2-6 | 8.654535e+08 |
| **5** | 41 | 2014-2-7 | 1.167509e+09 |
| **6** | 5 | 2014-2-9 | 9.890055e+08 |
| **7** | 18 | 2014-2-10 | 3.804009e+09 |
| **8** | 6 | 2014-2-12 | 2.873469e+09 |
| **9** | 5 | 2014-2-13 | 7.584551e+08 |
| **10** | 13 | 2014-2-17 | 1.355539e+09 |
| **11** | 10 | 2014-2-18 | 1.200082e+09 |
| **12** | 11 | 2014-2-19 | 8.827327e+08 |
| **13** | 6 | 2014-2-20 | 8.796141e+08 |
| **14** | 10 | 2014-2-21 | 6.440775e+08 |
| **15** | 13 | 2014-2-24 | 6.609436e+08 |
| **16** | 11 | 2014-2-25 | 1.211102e+09 |
| **17** | 4 | 2014-2-26 | 1.088923e+09 |
| **18** | 7 | 2014-2-27 | 1.119978e+09 |
| **19** | 9 | 2014-2-28 | 1.462775e+09 |
| **20** | 2 | 2014-3-1 | 1.239200e+09 |

รูปที่3.2.6dataframeที่รวมจากpantipและsetsmart

## 3.3การแสดงผลลัพธ์

จะแสดงผลโดยใช้bokehซึ่งจะแสดงผลออกมาเป็นไฟล์htmlโดยแสดงออกมาในรูปแบบกราฟที่มีแกนxคือวันที่แกนyมีสองแกนคือจำนวนcomment,replyที่นับได้และparameterจากsetsmartที่สนใจ

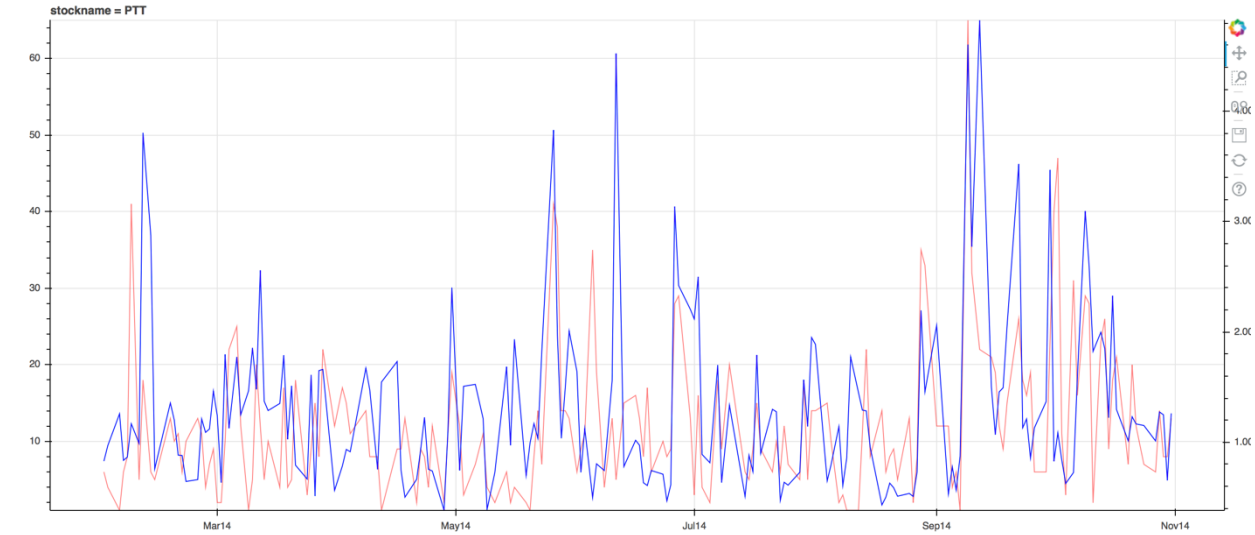



รูปที่3.3รูปกราฟผลลัพธ์

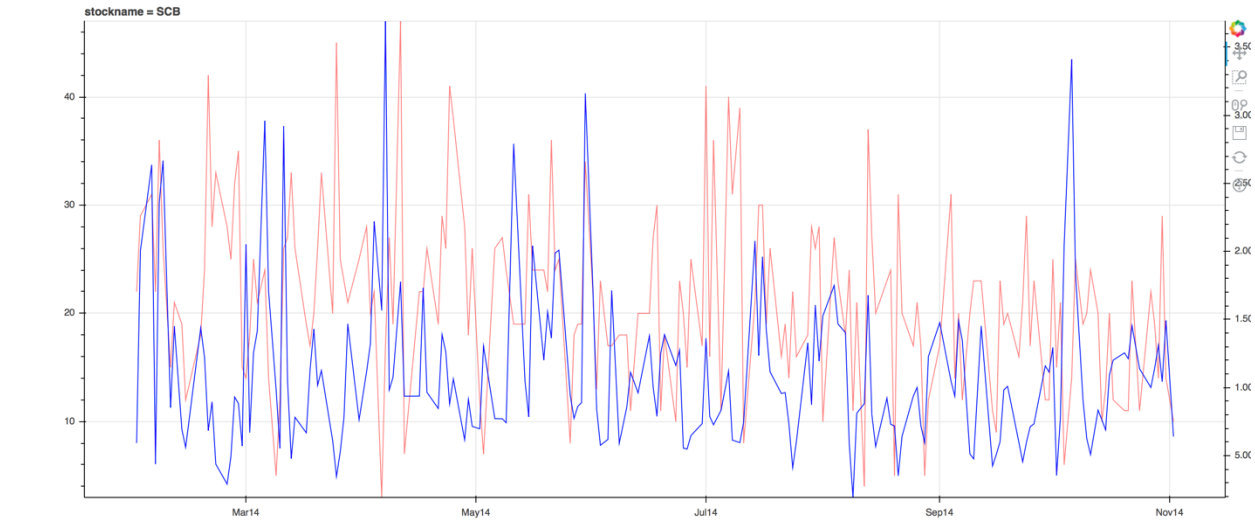
# 4.การทดสอบระบบและผลการทดสอบระบบ

ได้ลองทดสอบระบบตั้งแต่แรกจนถึงขั้นตอนสุดท้ายไม่ว่าจะเป็นการรับข้อมูลที่จะเป็นไฟล์.jsonเข้ามาเพิ่มการเปลี่ยนชื่อหุ้นที่สนใจ,เปลี่ยนparameterจากsetsmartจนถึงการ

นำแสดงผลออกมาในรูปแบบกราฟ




รูปที่4.1รูปแสดงผลการทดลองโดยใช้ชื่อหุ้นPTT




รูปที่4.2รูปแสดงผลการทดลองโดยใช้ชื่อหุ้นSCB

# 5.ปัญหาและอุปสรรคที่พบ

5.1ด้านperformanceการทำงานยังทำได้ช้าคิดว่ายังสามารถแก้ไขโค้ดให้ทำงานได้รวดเร็วขึ้นมากกว่านี้ได้อีก หรือทำเป็นclusterแล้วรันด้วยspark

5.2ยังไม่สามารถเรียกข้อมูลนามสกุล.jsonจากgoogle cloudมาเป็นinputได้โดยตรงจำเป็นจะต้องทำการดาวโหลดมาไว้บนเครื่องตัวเองก่อน

5.3การขอข้อมูลของsetsmartจะต้องทำเอกสารกับทางคณะบัญชีก่อนไม่สามารถหยิบข้อมูลออกมาใช้ได้เอง

5.4ข้อมูลของsetsmartที่ได้รับมามีparameterที่ไม่เข้าใจหลายตัวจึงจำเป็นต้องที่จะสอบถามจากอาจารย์ทางคณะเศรษฐศาสตร์

# 6.ข้อสรุปและแนวทางในการพัฒนางาน

6.1ข้อสรุป โครงการ การวิเคราะห์ความเปลี่ยนแปลงในตลาดหุ้นโดยใช้เว็บบอร์ดได้ถูกพัฒนาขึ้น ซึ่งมีเป้าหมายในการนับความถี่จำนวนcomment,replyที่กล่าวถึงชื่อหุ้นจากเว็บไซต์pantipนำมาเปรียบเทียบกับราคาในตลาดหลักทรัพย์จริงซึ่งดูได้จากฐานข้อมูลของsetsmart โดยผลลัพธ์ของโครงการที่ได้คือ กราฟที่แสดงถึงจำนวนความถี่ที่ปรากฏของชื่อหุ้นและparameterที่สนใจจากsetsmart เพื่อให้นักเศรษฐศาสตร์นำข้อมูลที่ได้นี้ไปทำการวิเคราะห์ต่อ

6.2แนวทางในการพัฒนางาน

6.2.1ข้อมูลที่ใช้ในโครงการเป็นข้อมูลเก่าคือปี2014สามารถที่พัฒนาได้มาขึ้นได้หากเรามีinputของข้อมูลปีใหม่ๆ

6.2.2 สามารถรับข้อมูลเข้าได้นอกเหนือจากpantip

6.2.3ทำให้เช็คว่าการพูดถึงชื่อหุ้นนั้นเป็นการพูดถึงเชิงบวก,เป็นกลางหรือพูดถึงในแง่ลบ

# 7.เอกสารอ้างอิง

[1] Apache-spark https://spark.apache.org/docs/latest/quick-start.html

**[2] bokeh**
**https://bokeh.pydata.org/en/latest/docs/user_guide/data.html**

**[3] SSE-PSIMS Data Model Version 4.12 Effective Date: January 3, 2018**