

**המחלקה להנדסת תוכנה וחשמל**

1. **שם הפרויקט: מערכת משולבת למסחר בשוק ההון**

Project Name: Integrated Trading System

דוח ביניים

|  |  |
| --- | --- |
| שם הסטודנט: | אבירם זוזות, חגי ז'ביליק, יריב מזרחי |
| מספר תעודת זהות: | 304965890, 036828069, 313576654 |
| שם המנחה: | עמית שטקל |
| חתימת המנחה: | אישור במייל בעמוד הבא |
| תאריך ההגשה: | 31.01.2021 |

1. **אישור מנחה:**
2. **תוכן עיניינים**

[**1.**](#_30j0zll) 1

[**2.**](#_1fob9te) 2

[**3.**](#_3znysh7) 3

[**4.**](#_2et92p0) 5

[מטרות הפרוייקט : 5](#_tyjcwt)

[מטרות: 5](#_3dy6vkm)

[מטרות הפרוייקט : יעדים: 5](#_1t3h5sf)

[2.](#_4d34og8) 6

[ריכוז שינויים: 6](#_2s8eyo1)

[עיקרי מסקנות: 10](#_17dp8vu)

[**5.**](#_3rdcrjn) 11

[**6.**](#_26in1rg) 14

[**7.**](#_lnxbz9) 15

[**8.**](#_35nkun2) 17

[**9.**](#_1ksv4uv) 18

[**10.**](#_44sinio) 20

[תיאור הארכיטקטורה: 20](#_2jxsxqh)

[**11.**](#_z337ya) 23

[**12.**](#_3j2qqm3) 26

[**13.**](#_1y810tw) 27

[**14.**](#_4i7ojhp) 28

[**15.**](#_2xcytpi) 30

[**16.**](#_1ci93xb) 32

[קוד כרייה: 32](#_3whwml4)

[קוד מידול: 39](#_2bn6wsx)

[**נספח ב1 - (Software Requirements Document (SRD, מסמך דרישות למודולי תכנה 43**](#_qsh70q)

[הקדמה: 43](#_3as4poj)

[דיאגרמת מחלקה: 44](#_1pxezwc)

[זחלן: 44](#_49x2ik5)

1. **תקציר:**

מסחר בשוק ההון הינו מסחר בעל נפח מסחר גבוהה היכול להגיע ל- 237 מיליארד דולר בשנה ואף יותר, במהלך סקירת הספרות נפגשנו עם מקורות מידע רבים אשר מציגים תנודתיות במסחר בסכומים דומים, בהתאם למדינה ולבורסה שבה נסחרים, כך שבבחירתנו בארה"ב בחרנו בבורסה התואמת מסחר סביב ל-237 מליארד דולר נכון ל-2021. מסחר מורכב ע"י סוחרים החל מהאדם הממוצע עד לסוחרים מתקדמים, חברות ביטוח, בנקים ותאגידים גדולים. סוחרים ללא עזרה של מערכת אוטומטית מתקשים לעקוב אחרי מספר רב של מניות, תזמון מכירה וקניה, ניתוק רגשות ופחדים מהתנהגות השוק. עם זאת המסחר האוטומטי מתבסס על אלגוריתמים שונים, מגוונים הנותנים מענה לבעיות הנ"ל ונותנים חיזוי המשפרים את ההצלחה בהשקעה במספר מניות רב.

מטרת המערכת היא לייצר מערכת אוטומטית המבוססת על למידת מכונה, אשר תבצע שימוש וסחר בניירות ערך בשוק ההון. זאת תוך ניתוח מודלים פיננסיים, ניהול סיכונים וניתוח הודעות מתפרצות רלוונטיות ממקורות מידע ברשת ומדיה חברתית, כאשר שאיפת המערכת לייצר תשואה חיובית בשוק ההון, תוך מתן דגש על אוטומציה וחיסכון בזמן.

היעדים הם:

איסוף נתונים - בקשת מידע וריכוזו, מאתרי מדיה כגון yahoo finance gurofucus ומשוק המניות האמריקאי לאורך 20 השנים האחרונות.

* 1. עיבוד מידע ראשוני – המרת המידע מטקסטים למאפיינים נומריים המשקפים את הערך הסנטימינטלי של המידע.
  2. לימוד מודל באמצעות למידת מכונה - ייצור מודל על בסיס המידע שנאסף תוך שימוש ברשת נוירונים ומערכת לזמן קצר/ארוך, לטובת חיזוי ערך הסגירה עבור מניה ביום הבא.
  3. דירוג מניות - הערכת ניירות ערך ליום המסחר הבא לפי פרמטרים סטטסטיים (נתוני מסחר בורסה) ודינמיים (ניתוח טקסטים במדיה חברתית).
  4. ניהול מסחר - פיתוח תהליך פקודות מסחר יומיות בבורסה.
  5. ניהול סיכוני תיק - בניית מודל מתאים ליישום תהליך ניהול סיכונים.
  6. הצגה - בניית ממשק משתמש להגדרת והפעלת המערכת.

צריך לעדכן את היעדים... יש צורך למקד אותם כדי שתוכלו לעמוד בהם.

תכנון סביבת העבודה שנבחרה היא .NET כאשר הIDE שנבחר הוא Visual Studio. בחרנו בסביבה זו בעיקר בגלל נוחות, ניסיון עבר, תמיכה במספר שפות וקהילה רחבה. שפות הפיתוח שנבחרה היא Python וC#, כאשר Python היא השפה הנפוצה ביותר להבאת נתונים, בניית מודלים. בנוסף קיים המון מידע וניסיון צבור ברחבי הרשת כאשר העלות היא חינם.

במערכת המתוכננת בצד הClient נבחרה פלטפורמת Angular של Google הפופולרית בעולם הWeb עם סביבת העבודה Visual Studio שהוזכרה למעלה. בצד הServer השימוש יהיה בסביבת Net Framework של Microsoft, ובעזרת API תהיה משיכת המידע מהבורסה ורשתות חברתיות.

תכנון כריית המידע תבוצע ע"י שפת Python, ובעזרת ספריות כמו selenium, nltk, keras, scikit--learn

**מטרות הפרוייקט :**

**מטרות:**  
יצירת מערכת אוטומטית ￼המבוססת על למידת מכונה, אשר תבצע שימוש וסחר בניירות ערך בשוק ההון. זאת תוך ניתוח מודלים פיננסיים, ניהול סיכונים וניתוח הודעות מתפרצות רלוונטיות ממקורות מידע ברשת ומדיה חברתית, כאשר שאיפת המערכת לייצר תשואה חיובית בשוק ההון, תוך מתן דגש על אוטומציה וחיסכון בזמן.

**מטרות הפרוייקט :  
יעדים:**

* 1. איסוף נתונים - בקשת מידע וריכוזו, מאתרי מדיה כגון yahoo finance gurofucus ומשוק המניות האמריקאי לאורך 20 השנים האחרונות.
  2. עיבוד מידע ראשוני – המרת המידע מטקסטים למאפיינים נומריים המשקפים את הערך הסנטימינטלי של המידע.
  3. לימוד מודל באמצעות למידת מכונה - ייצור מודל על בסיס המידע שנאסף תוך שימוש ברשת נוירונים ומערכת לזמן קצר/ארוך, לטובת חיזוי ערך הסגירה עבור מניה ביום הבא.
  4. דירוג מניות - הערכת ניירות ערך ליום המסחר הבא לפי פרמטרים סטטסטיים (נתוני מסחר בורסה) ודינמיים (ניתוח טקסטים במדיה חברתית).
  5. ניהול מסחר - פיתוח תהליך פקודות מסחר יומיות בבורסה.
  6. ניהול סיכוני תיק - בניית מודל מתאים ליישום תהליך ניהול סיכונים.
  7. הצגה - בניית ממשק משתמש להגדרת והפעלת המערכת.

1. **מדדים**:
   1. איסוף נתונים - איסוף נתונים טקסטואליים מ-2 מקורות מידע לפחות, ועיבוד של 20 טקסטים כלכליים חדשים(או ידיעות חדשותיות) לפחות בכל יום.
   2. עיבוד מידע ראשוני – בדיקת המודל אל מול label אנושי של המידע והצלחה בחיזוי ברמה של יותר מ60%
   3. לימוד מודל באמצעות למידת מכונה או למידה עמוקה- צפי לאחוז הצלחה גבוה בשלב ה-validation.
   4. דירוג מניות - ניהול רשימה של 10 המניות המומלצת למשתמש ברמה יומית.
   5. ניהול מסחר - בחירה מתוך רשימה של לפחות 100 ניירות( מניות) והשקעה בו זמנית ב- 10 מניות לפחות.
   6. ניהול סיכוני תיק - השגת יחס רווח לסיכון (מדד שארפ) של לפחות 0.5.

**ריכוז שינויים מופיע בפרק 7 בפורמט מחלקת חשמל ופרק 14 בתוכנה.**

**ריכוז שינויים:**

במסגרת כריית המידע בוצעה חקירה לגבי חוקיות כריית מידע וההבדלים בין כריית מידע עם זחלן ושימוש בAPI, בספטמבר 2019 חברת hiQ יצאה זכאית במשפט בהרכב של 3 שופטים אל מול חברת Linkedin שתבעה אותם על הפרת אמנת CFAA ובעצם יצרה תקדים חוקי לכריית מידע.   
בכל אתר קיים קובץ robots.txt המנחה את בוטים החיפוש איזה חלקים באתר מותר להציג בתוצאות חיפוש, אבל אין אזכור משפטי לגבי כריית מידע, וגם אפילו אם בוט חיפוש עובר על החוקים בקובץ. עם זאת מאוד נהוג להתייחס לקובץ כאל מנחה במה מותר להציג מהאתר למרות שאין לו שום חוזק משפטי. אפילו google הכריזה שהיא יוצאת נגד הפקודה noindex המבטלת את האופציה להציג את האתר בתוצאות החיפוש שלה.

הוחלט על שימוש עיקרי בספריות Python, בעיקר בSelenium ובScrapy & beautifulSoup כדי להשלים פקודות שלא יתממשו בSelenium. שאר הכלים שנבחנו עולים כסף ו/או לא נותנים פתרון מיידי, אלא שחינמים מוגבלים בכמות מידע שנכרה / זמן פעולת הרובוט והתוצאות למשתמש.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| פיצרים מיוחדים | זמן | תשלום | אוטומטי | שם מוצר |
| טיפול ב  Proxy  Browser  IP block  CAPTCHA | מיידי | כן | חצי | [Scraper API](https://www.scraperapi.com/) |
| Csv to mail | 1 –2 שבועות | כן | לא | [ScrapeSimple](https://www.scrapesimple.com/) |
| בניית זחלן לפי מאפיינים וללא קידוד (ממשק גרפי)  תמיכה בענן  Cvs , excel, API  IP rotation | לא מיידי, משתנה | לא אבל מוגבל  כן- ללא הגבלה | כן | [Octoparse](https://www.octoparse.com/) |
| ללא קוד  IP rotation  Login  Csv / json | לא מיידי, משתנה | לא - מוגבל  כן- ללא הגבלה | כן | [ParseHub](https://www.parsehub.com/) |
| **מחלקה פתוחה בpython** | **מיידי** | **לא** | **כן** | [**Scrapy**](https://scrapy.org/) |
| [**BeautifulSoup**](https://www.crummy.com/software/BeautifulSoup/bs4/doc/) |
| [**Selenium**](https://www.selenium.dev/) |
| Fine tune  Api  Crawl history | מיידי | כן | כן | [Diffbot](https://www.diffbot.com/) |
| מחלקה פתוחה בNodeJS | מיידי | לא | כן | [Cheerio](http://cheerio.js.org/) |
| [Puppeteer](https://github.com/puppeteer/puppeteer) |
| Scrape text, files, images and PDF content from web  TSV, CSV,XML, XLSX or JSON | לא מיידי, משתנה | כן | כן | [Mozenda](https://www.mozenda.com/) |
| ללא צורך בקוד | לא מיידי, משתנה | כן | כן | [Import.io](https://www.import.io/) |
| Stock monitor  Product analytics  Data reports | לא מיידי, משתנה | כן | כן | [Scrapinghub](https://www.scrapinghub.com/) |
| Proxy  Multiply crawlers  No code knowledge | לא מיידי, משתנה | כן | כן | [Dexi.io/CloudScrape](https://www.dexi.io/) |

הכרייה מתבססת על העובדה שהרוב המוחלט של אתרים נבנה ע"י html עם חוקים וקווים מנחים ברורים. לאחר חקירה של כל אתר מידע ניתן להבין את האופן בו הוא בנה את התאר והאופן בו הוא מציג את המידע.   
כל אתר מקבל תשומת לב והבנה איך ניתן למשוך בצורה הכי מהירה ואפקטיבית את המידע. אפשרות של API, קריאת html או איתור ערך ערך בצורה רלוונטית בקצב איטי.   
כריית המדיע צריכה להיות בצורה איטית ומדמה פעולות אל אדם ממוצע בהתנהגות אל מול האתר. רוב האתרים כיום מגנים על כמות הבקשות לשרת שלהם על ידי הגבלה בAPI שהם מספקים, חיפוש בקשות על ידי רובוט (כמות בקשות, כתובת המכשיר המבצע את הדרישה למידע, כמות הלחיצות באתר, לחיצות על מידע שקיים ויזואלית ולא מוחבא בקוד (מלוכדת רובוט בקוד), מהירות הפעולות באתר (נסיון למשוך מידע לפני שהוא נטען), ועוד..)   
בעת בניית הזחלן יש ללמוד את הגנות האתר על ידי ניסוי וטעיה ולנסות לצמצם את כמות הבקשות, לחיצות מהאתר כדי להקטין את הסיכון לזיהוי כזחלן רובוטי.

מכיוון שהמערכת היא לצורך לימוד ובדיקת יכולות, הוחלט לוודא שאכן הכרייה מתבצעת לפי חוקי בוט החיפוש(robots.txt) ובהגדרה אין משיכה של מידע בתשלום, אלא רק מה שמוצג באופן פומבי למשתמש הרגיל.

כרגע האתרים שנבחרו לכרייה הם אתרים שהמידע שאנחנו צריכים לגשת אליו הוא חינמי

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **stockwteets** | **Yahoo finanace** | **openInsider** | **gurofucus** |  |
| **אין** | **יש** | יש | **אין** | **API** |
| **אין** | **יש** | יש | **אין** | **תשלום על שירותי אתר** |
| **אין הגבלה על השירותים המבוקשים** | **אין הגבלה על השירותים המבוקשים** | יש הגבלה | **אין הגבלה על השירותים המבוקשים** | **Robots.txt** |

כרגע המערכת עוקבת אחרי ידיעות של משתמשים כלליים שרוצים להביע דעה על המניה אבל הם לא מומחים או אנשי מפתח במדיה היכולים להשפיע בצורה קיצוני על מחיר מניה. לאחר שלב הסיווג מידע, ובניית מודל חיזוי נוכל להרכיב רשימה של אנשים שהשפיעו / צדקו בהערכה שלהם על המניה והמערכת תדע לחפש בנוסף (כגורם קלט נוסף) ידיעות ספציפיות שלהם וגם נוכל להכניס לרשימה אנשי מפתח שאנחנו הגדרנו כרלוונטיים.

בהיבטי משימות הנוגעות לטקסט (ניתוח טקסט, הבנה סמנטית, חילוץ ישויות ואימון מודלים) נבחנו מספר ספריות PYTHON, המסוגלות למלא את המשימות המבוקשות, ביניהן NLTK, SPACY, ALLEN NLP, SPARKNLP, STANFORD. חלקן לא עמדו בכל המשימות הנדרשות ונפסלו וחלקן היו בשימוש מצומצם לעומת האחרות. לבסוף הוחלט להתקדם עם ספריית NLTK, היודעת לעמוד בכלל המשימות (חילוץ ישויות, קשרים, פירוק משפטים, ניתוח תחבירי ואימון מודלים) והיא ספריית בעלת השימוש הרחב ביותר כיום בתחום הNLP. בשלב ראשוני, על מנת לבצע ניתוח טקסט מבלי לחכות לכריית המידע, נפנה אל תוצרים מוכנים הנמצאים כיום ברשת וכך נוכל להפיק תובנות וניסיון כאשר המידע שאנו מעוניינים בו יהיה מוכן. סיכון בהיבט הזה הוא הבנה סמנטית, שכל עולם הNLP מנסה לפענח סמנטיקה בצורה גבוהה(בעיית הציניות למשל) מכיוון שזאת בעיה גלובלית שעדיין אל נפתרה נוכל להסתכל על כמות רבה של דיווחים לכל מניה ועל כמה אנשים שנתייג כבעלי

|  |  |  |
| --- | --- | --- |
| **יכולות** | **תאור** | **תשתית** |
| חילוץ ישויות, חילוץ קשרים, פירוק למשטים, ניתוח תחבירי | תשתית לביצוע סט רחב של פעולות NLP. בשימוש נרחב בתעשייה. | NLTK |
| חילוץ ישויות, חילוץ קשרים, פירוק למשטים, ניתוח תחבירי | תשתית לאימון מודלים למשימות NLP | SPACY |
| חילות ישויות, חילוץ קשרים | מבוססת על SPACY. מוסיפה יכולות הסבר למשתמש | ALLENNLP |
| חילוץ ישויות, חילוץ קשרים, פירוק למשטים, ניתוח תחבירי | תשתית לאימון מודלים למשימות NLP מבוססת ע MXNET בשימוש אמזון | GLUON NLP |
| חילוץ ישויות, חילוץ קשרים, פירוק למשטים, ניתוח תחבירי | תשתית לאימון מודלים למשימות NLP | SPARK NLP |
| ניתוח תחבירי | מנתח תחבירי בעברית | YAP |

אמינות גבוהה ולא נסתמך רק על דעה אחת.

בתחום המודל: נשתמש בשפת Python בהתאם לחלקים הקודמים בפרויקט.

מבחינת בחירת הספריות והמודל למימוש נבחנו מספר אפשרויות המסוכמות בטבלה הבאה:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **שם הספרייה** | **תחום התמחות** | **פופולריות** | **שימוש עיקרי** | **עדכני** | **דירוג התאמה** |
| TensorFlow | רשתות נוירונים | 5 | גרפים | 5 | 4 |
| **Scikit-learn** | **למידת מכונה, למידה עמוקה** | **4** | **נתונים** | **4** | **5** |
| PyTorch | למידת מכונה ורשתות נוירונים | 3 | גרפים | 5 | 2 |
| Keras | למידת מכונה | 4 | הדגמה | 4 | 1 |
| Shogun | למידת מכונה | 1 | אקדמי | 1 | 3 |
| Torch | למידת מכונה | 3 | חברות ענק | 2 | 3 |

השתמשנו בספריית Scikit-learn מכיוון והיא ממוקדת לתחום למידת המכונה, מכילה פונקציות רבות שעוסקות בניתוח נתונים, היא יחסית פופולרית ועדכנית.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **מודל** | **דירוג** | **מסווג** | **דירוג** | אמצעי הערכה | **דירוג** |
| ANN | 5 | Filtering | 3 | MAPE | 4 |
| NATIVE BAYES | 2 | Fuzzy based system | 5 | RMSE | 3 |
| HMM | 2 | Kmeans | 1 | ACCURACY | 5 |
| SVM | 4 | Optimization | 4 | NMSE | 1 |
| RNN | 3 | - | - | HIT RATE | 1 |
| DSS | 3 | - | - | PRECISION | 2 |

בחרנו להשתמש ב- RNN כמודל מכיוון ואנו רוצים לנסות משהו שונה מחומרים שנחשפנו אליהם כגון SVM ושיטות מסורתיות מצד אחד, אך לא להפריז ולנסות HMM שאולי יהיה לא פרקטי ליישום במסגרת הזמנים, מכאן ש- RNN הוא מודל מורכב אך ישים במסגרת הזמן.

מבחינת סיווג, בחרנו באופטימיזציה מכיוון וזו שינה המניבה תוצאות טובות בהערכת הסקר ספרות וזו שיטה שמתאימה יותר לסביבת העבודה ולמודל שנבחר, במקרה שלנו האופטימיזר הוא מסוג אדם, אשר מבטא שיטת חישוב לדעיכת גרדיאנט ועוד מומנט כאשר כל אחד מהם מוכפל במשקולת גמא המבטאת אחוז, כך שיש איזון בין התכנסות ומהירות שינוי. מבחינת אמצעי ההערכה, בחרנו RMSE מתוך שיקולי התאמה לסביבת העבודה ולמודל, ננסה לבחון שיטה נוספת להערכה.

**עיקרי מסקנות:**

בתחום הכרייה: יש לעקוב באופן רציף אחרי האתרים (אופן תצוגת המידע, הנגישות אליו והאחסון), פלט הכרייה (לראות שהוא לא השתנה) והספריות (שלא שינו שירותים ואופן הפעלה) כדי שנוכל לקבל את הקלט הנדרש. בנוסף יש לבדוק שהאתר לא מזהה אותנו כרובוט והגישה אל המידע בעזרת הזחלן תופסק.

בתחום המידול: כדי שמודל יהיה מוצלח, הוא צריך לנבע משהו שאחרים לא ניבעו, מכאן שברוב המקרים צריך לחדש מודלים כדי להמשיך להקדים את המשק בחיזוי ערך הסגירה ליום הבא.

מודלים אשר השימוש בהם היה נהוג בעבר בתחום כגון SVM מציגים יחס רווח לסיכון מתחת 60% לפי סקירת הספרות, לכן אנו נבחן מודלים עדכניים כדוגמת RNN ו- LSTM משולב עם RNN.

1. **סקירת ספרות וסקר שוק**
   1. בוצעה חקירה מעמיקה של הספריות איתם מתבצעת העבודה, עיקר החקירה הייתה קריאת מאמרים (לא אקדמיים) של אנשי מקצוע הביצעו פעולות זהות ושיתפו את הממצאים שלהם דרך האתר ([**www.medium.com**](http://www.medium.com) , [**www.towardsdatascience.com**](http://www.towardsdatascience.com)), וחקירה שוטפת של הספריות המקצועיות ופעולות שניתן לבצע איתם (מפורט ברשימת מקורות)

* 1. בחנו מקור מידע נוסף המראה בחינה של מודלים מסורתיים למטרה דומה לפרויקט זה, חיזוי יומי של ערך מניה. לפי מקור זה, המודלים המסורתיים כגון SVM לא הצליחו לחזות יחס רווח לסיכון גבוה, הם התקשו לעלות מעל 60% סיכויי הצלחה.  
     מקור הכתבה : [קישור](https://www.sciencedirect.com/science/article/pii/S1877050916311619)

Aparna Nayak, M. M. Manohara Pai, Radhika M. Pai,

Prediction Models for Indian Stock Market ,

Procedia Computer Science ,

Volume 89 ,

2016 ,

Pages 441-449 ,

ISSN 1877-0509 ,

(<https://www.sciencedirect.com/science/article/pii/S1877050916311619> )

* 1. האתר IKNOWFIRST, אשר מתיימר לספק תוצאות חיזוי בהתבסס על מידע עבר תוך שימוש במודלי AI. זוהי חברת פינטק אשר מספקת מערכת למידה אשר מתבססת על אלגוריתם חיזוי.  
     החברה מספקת חיזוי יומי להשקעות, מודל החברה מבוסס על מידע לאורך שנים ככניסה ומידול באמצעות DSS – Decision Support System כמערכת ייעול.  
     החברה מספקת את התחזיות לטווח קצר בינוני וארוך, בעיקר לחברות גדולות ולא לצרכן פרטי, מקריאת מאמר זה הבנו שניתן לבנות מערך קבלת החלטות על ידי מספר אלגוריתמים אשר יחד מהווים מערכת אחת מסועפת, ותוצאות אלו מניבות הכנסות בעולם העסקי של ימינו, מכאן שיש ערך כלכלי גם לעבודה בתחום זה ולא רק לתוצר המוגמר, לעבודה הבנו שניתן לשלב מספר מודלים אך יש עוד מחקר עומק שניתן לבצע על שילוב מודלים מתקדמים.  
     קישור: [www.iknowfirst.com](http://www.iknowfirst.com)
  2. סקירת ספרות על מודלים מסווגים ואמצעי הערכה  
     סקרנו מה הקריטריונים המקובלים להערכת המודלים ואילו מודלים פופולריים בתחום: [קישור למאמר](https://www.sciencedirect.com/science/article/abs/pii/S157401371930084X?via%3Dihub)

Dattatray P. Gandhmal, K. Kumar,

Systematic analysis and review of stock market prediction techniques,

Computer Science Review,

Volume 34,

2019,

100190,

ISSN 1574-0137,

https://doi.org/10.1016/j.cosrev.2019.08.001.

(<https://www.sciencedirect.com/science/article/pii/S157401371930084X>)

* 1. במסגרת סקירת הספרות נבחנו מספר סקירות על מודלים והשוואתם האחד לשני כגון:

Comparison of Financial Models for Stock Price Prediction

בקישור: [קישור למאמר](https://www.researchgate.net/publication/343657965_Comparison_of_Financial_Models_for_Stock_Price_Prediction)

Islam, Mohammad Rafiqul & Nguyen, Nguyet. (2020). Comparison of Financial Models for Stock Price Prediction. Journal of Risk and Financial Management. 13. 10.3390/jrfm13080181.

<https://www.researchgate.net/publication/343657965_Comparison_of_Financial_Models_for_Stock_Price_Prediction/citation/download>

* 1. <https://www.visualcapitalist.com/the-worlds-10-largest-stock-markets/>
  2. <https://data.worldbank.org/indicator/CM.MKT.TRAD.CD?locations=US> - נפח מסחר בארה”ב, ניתן לקבל גם נפח מסחר עולמי ועוד אפשרויות.

קראנו וסיכמנו את המידע לקבלת הערכה כללית משווי השוק ומה שווי בורסאות שונות בעולם, המסקנה היא שישנו כסף רב והתחום פופולרי מאד, מכאן שלמרות שהחסמים להיות מוביל שוק בתחום פינטק גבוהים, עדיין גם ללא הובלה כלל ואפילו הצלחה מועטה ניתן להרוויח המון.

* 1. אתר גורופוקוס הוקם על ידי ד"ר לפיזיקה מתוך הבנה שבאמצעות הבנה במסחר בשוק ההון ניתן להרוויח ולמנף במהירות, האתר מכיל ידע רב אודות המלצות של גדולי דעה כגון וורן באפט ואחרים. מקור מידע זה ענה לנו על שאלות כגון מהי ההערכה של אנשי מקצוע מהן הסטטיסטיקות הנפוצות ומה המגמות עליהן מצביעים אנשי דעה בתחום. האתר הזה טוב יותר ממקורות שונים שבחנו בכך שהוא מיועד לכל הרמות, כולל מתחילים ומכיל מידע מקיף ועמוק בתחום, משווי שוק, מגמות עולמיות, הערכות לטווחי זמן שונים ועוד, בהמשך נבחן אפשרות לקליטת מידע מאתר זה מכיוון והוא מפרסם עסקאות בעלי הון, עסקאות אלו בעלות משמעות רבה עבורנו בפרויקט זה מכיוון וזו אינדיקציה טובה האם בעלי המניה מאמין שהיא תעלה או שהגיע הזמן לממש וכמה. הכוונה בלממש הנה כמובן למכור במטרה לגרוף רווחים.

<https://www.gurufocus.com/new_index/?r=1410f230ed925fe5d599ddd5585ee0bf&utm_source=adwords&utm_medium=2&utm_campaign=homepage&gclid=CjwKCAjwjuqDBhAGEiwAdX2cjwFzNRAOeCLNev2dNs6HhwhEixDvi1fpdDPUmEjZl9T3SIWO5tCFDRoC7g4QAvD_BwE>

* 1. מאמר המדגים ערכית מערכת מבוססת על ניתוח אות וחישוב ממוצעים שונים, השאלה ששאלנו במאמר זה, היא האם ניתן ועד כמה ניתן לחזות את מחיר המניה באמצעות ניתוח אות, לסיכום נותחו מספר דוגמאות והוצגה מערכת לקבלת החלטות, שילוב של דברים אשר נלמדו בלימודינו, כאשר המסקנה בשורה התחתונה היא שהמחיר תלוי בעיקר בהיסטוריה והתנודתיות של המניה.  
     ננסה לבחון זאת בעבודתנו.  
       
     https://scholar.google.com/scholar?hl=iw&as\_sdt=0%2C5&q=Digital+Signal+Processing+for+Predicting+Stock+Prices&btnG=[‏](https://www.researchgate.net/profile/Ledisi_Kabari5/publication/340809604_Digital_Signal_Processing_for_Predicting_Stock_Prices_in_Financial_market/links/5ea86f7892851cb26760bd7d/Digital-Signal-Processing-for-Predicting-Stock-Prices-in-Financial-market.pdf)

DOUGLAS, T. Minafa; KABARI, L. Giok. Digital Signal Processing for Predicting Stock Prices in Financial market.‏

* 1. במאמר זה שאלנו את השאלה, מהם המודלים בתחום עיבוד אות אשר משמשים לחיזוי מחירי מניות ומה מידת דיוקם והקריטריונים לאבחנה זו, כגון rmse ומערכת lstm ועוד, להבנתנו מערכות אלו לא מניבות תוצאות מעל 80% ולכן יש עוד הרבה מקום לשיפור, בין אם באמצעות סוג המערכת או באמצעות שילוב מודלים קיימים.

An improved deep learning model for predicting stock market price time series  
<https://www.sciencedirect.com/science/article/abs/pii/S1051200420300865>

1. **מסמך דרישות ראשוני של המערכת:**

פירוט דרישות ותכנון מימוש של אלגוריתמים:

1. כריית מידע:
   * זמינות אתרים (ללא שינוי המידע- האופן בו התוכן נשמר ומוצג באתר)
   * הצלחה בכריית המידע ומשיכת המידע
   * אי חסימת הזחלן (גילוי כרובוט לא חוקי)
   * שלמות מידע והפרדה ברורה של המאפיינים(שדות) שנכרו (טקסט, מחיר, משתמש, תאריך...)
2. עיבוד מידע:
   * קבלת המידע לפי מה שנקבע
   * הצלחת חיזוי של אפיון טקסט
   * הצלחת המרת כלל המאפיינים לערכים נומריים
3. יצירת מודל:
   * קבלת כלל המאפיינים לפי דרישה
   * יצירת מודל עם מאפיינים ממדיה חברתית
4. מערכת:
   * יכולת שמירת נתוני משתמש (הרשמה)
   * יכולת בדיקה והשוואה של נתונים (התחברות וזיהוי)
   * הגדרת ושינוי מאפייני מסחר
   * משיכת מידע והצגת הנתונים מהמודלים.
5. אבטחה:
   * יכניסה למערכת (הצגה ושינוי של הנתונים) רק לאחר זיהוי אל מול המערכת.
   * חיסיון פרטים אישיים.
6. **חלופות פונקציונאליות למימוש:**

החלופות לסביבות עבודה שבחנו הן שרת אינטרנטי(מרוחק) ושרת מקומי(פיתוח בעזרת Visual studio ו.NET FRAMEWORK).

שרת אינטרנטי - סביבת עבודה ייעודית למטרת פיתוח אלגוריתמים, חסרון מאוד בולט הוא הקושי לשלב אותה עם מערכות נוספות , ביצועים נמוכים ומגבלות אבטחה.

- Visual studio.net

סביבת עבודה נוחה ומוכרת לפיתוח המתקשרת עם הרבה שפות, כאשר החסרון של הוא בביצועים פחות טובים עקב דרך המימוש בהעלאת הקוד לשרת.

על מנת להחליט מהי סביבת העבודה הנכונה עבור הפרויקט, בחרנו מס' פרמטרים, דירגנו כל סביבה וכך בחרנו בסביבת העבודה הנכונה ביותר.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | ציונים | | ציונים משוקללים | |
|  | משקל | Visual studio.net | שרת אינטרנטי | חלופה 1 | חלופה 2 |
| פשטות | 20% | 3 | 2 | 0.6 | 0.4 |
| עלות | 10% | 1 | 1 | 0.1 | 0.1 |
| ביצועים | 15% | 3 | 2 | 0.45 | 0.3 |
| פונקציונאליות | 30% | 3 | 3 | 0.9 | 0.9 |
| זמינות | 15% | 3 | 2 | 0.45 | 0.3 |
| ניסיון עבר של חברי הצוות | 10% | 3 | 1 | 0.3 | 0.1 |
|  | 100% |  |  | **2.8** | **2.1** |

בבחינת כלל החלופות, הוחלט כי ממשק VISUAL STUDIO ועבודה מול שרת מקומי יניב לנו את מירב היתרונות הרצויים מבחינת הצוות על מנת לספר את המע' בצורה הטובה ביותר, תוך מתן דגש על גמישות, זמינות, פונקציונאליות וכו'.

בהמשך לסביבת העבודה, בחנו גם מס' חלופות לטכנולוגיית הפיתוח(שפת הפיתוח) בה נרצה להשתמש, כאשר לנגד עינינו עמדו מס' פרמטרים הנחוצים לנו לצורך מימוש המע'.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | ציונים | | | | ציונים משוקללים | | | |
|  | משקל | Python | C# | C++ | Matlab | Python | C# | C++ | Matlab |
| פשטות | 20% | 10 | 8 | 5 | 10 | 2.0 | 1.6 | 1.0 | 2.0 |
| עלות | 10% | 10 | 3 | 4 | 0 | 1.0 | 0.3 | 0.4 | 0 |
| ביצועים | 20% | 6 | 5 | 8 | 9 | 1.2 | 1.0 | 1.6 | 1.8 |
| פונקציונאליות | 30% | 9 | 8 | 5 | 9 | 2.7 | 2.4 | 1.5 | 2.7 |
| ניסיון עבר של חברי הצוות | 20% | 9 | 8 | 2 | 3 | 1.8 | 1.6 | 0.4 | 0.6 |
|  | 100% |  |  |  |  | **8.7** | **6.9** | **4.9** | **7.1** |

בבחינת החלופות, הוחלט לבחור בPython כשפה העיקרית שתשמש אותנו לצורך בניית האלגוריתם, כאשר לבניית צד השרת של המע' נשתמש בC# ולצד המשתמש בפלטפורמת Angular.  
החלופות שבדקנו היו לצורך בחירת שפת התכנות בה נממש את האלגוריתם שלנו, כאשר עיקר הקריטריונים היו פשטות, פונקציונאליות וביצועים, אשר הינם קריטיים לצורך הצלחת המערכת.

1. **חלופות מערכתיות:**
2. חלופה א'
   1. מבחינת איסוף מידע: שימוש ב-API בתשלום כגון (אתר)
   2. מבחינת עיבוד מידע: נבחר את כל המאפיינים ונשתמש ב- exponential smooth כדי לעדן את התוצאות במטרה להגדיל את סיכויי החיזוי.

באופציה הזו אנו בוחרים את המאפיינים, לכן יש צורך להביא מספיק מאפיינים, לעשות בדיקה סטטסיטית שלהם ולהעריך את החשיבות של כל אחד מהם לחיזוי. אפשר לעשות את זה גם באופן נאיבי על ידי חישוב המתאם (קורלציה) בין המאפיין לתנועת המניות.

* 1. מבחינת מודל: שימוש במודל כגון SVM, ובחירת מאפיינים באופן

באופן ידני.

* + 1. שינוי משקולות באמצעות עדכון מומנט

1. חלופה ב'
   1. מבחינת איסוף מידע: שימוש ב-API חינמי כגון יאהו
   2. מבחינת עיבוד מידע: נבחר חלק מהמאפיינים וננרמלם בלבד.
   3. מבחינת מודל: שימוש במסווג מסוג RANDOM FOREST ושימוש ב- KNN (***k*-Nearest Neighbors algorithm)**
      1. שינוי משקולות באמצעות דעיכת גרדיאנט
2. חלופה ג'
   1. מבחינת איסוף מידע: שימוש בזחלן בתשלום
   2. מבחינת עיבוד מידע: נבחר רק את מחיר הסגירה וננרמל אותו.
   3. מבחינת מודל: נבחר רשת נוירונים עמוקה, כך שהמערכת תבחר את המשקולות בעצמה.
      1. נוכל לבחור בעדכון משקולות בצורה הבאה:
         1. דעיכת גרדיאנט - מתכנס לאט
         2. מומנט - משתנה מהר
         3. שילוב - משתנה מהר ושואף להתכנסות.
         4. שילוב והכפלה בפרמטר המבטא באחוזים כמה כל רכיב ישפיע - מתכנס מהר
3. **חלופות טכנולוגיות:**
4. כריית מידע מאתר gurofucus יכולה להתבצע בצורה "איטית וכבדה" יותר מאשר היא מוצגת בקוד:
   * 1. במקום לתת לדרייבר להכניס את הפרמטרים לחיפוש הפעולות על המניה, מי עשה אותם, איזה תאריכים וכמות תוצאות שמוגבלת ל100- זוהה pattern בUrl של כתובת האתר, כאשר מאפייני החיפוש מוזנים והטבלה מוצגת כל הפרמטרים של תוצאות החיפוש מוצגים בURL.  
        לאחר מחקר קל (ניסוי וטעיה) מסתבר שניתן לדלג על שלב הכנסת הנתונים לחיפוש ולהכניס את הפרמטרים ישירות לכתובת ולמשוך את הטבלה, ואפילו לבטל את הגבלת התוצאות מ100 ל5000.  
        לצורך זה הוכנה מתודה היודעת לפי תאריך מוכנס לייצר כתובות url מותאמות אישית לצורך הצגת הטבלה. דבר זה מאפשר וחוסך את זמן התצגת הטבלה. שיטה זאת נקראת Url manipulation
     2. במקום למשוך את הטבלה בצורה יסודית, מעבר על השורה הראשונה בטבלה, ולהגידר אותה כעמודות, ולאחר מכן לעברו שורה שורה ולהעתיק ערך ערך ולהכניס אותו לטבלה- לאחר חקירה של קוד הHtml, התגלה שהטבלה שמורה באופן פשוט ביותר כך שמספר פקודות פשוטות (2-3) מספיקות למשוך את טבלה ב1-2 שניות בודדות. בצורה הנאיבית והאיטית יכולה להגיע לאזור ה200-300 שורות קוד כדי להביא את אותה טבלה.
5. עיבוד מידע:  
   ניתן לממש מס' חלופות:  
   1. בחירה באחת מהספריות השונות שבחנו, למשל SPACY אשר נותנת יכולות דומות.  
   2. שימוש במנוע NLP כמו BERT או BIG BIRD, אשר נותנים סט יכולות ולמידה רחב ועשיר יותר של הבנת טקסט, חילוץ קשרים ועצמים. שימוש
6. מבחינת מודל:

א. ניתן לבחון את מערכת הלמידה באמצעות SVM – SUPPORT VECTOR MODEL, הכניסה תהיה נתונים סטטיים והערכת חדשות מתפרצות, כאשר ננסה להעריך האם המניה תעלה או תרד ביום למחרת, אך הערכה זו גסה יותר וטומנת בחובה מספר תת אפשרויות, כאשר אפשרויות אלו נבדקו בסקירת הספרות והציגו תוצאות לא טובות - מתחת ל-60% הצלחה, זו שיטה שהייתה מקובלת בעבר.

ב. ניתן לשנות את ה - optimizer, שיטת חישוב המשקולות, כעת אנו משתמשים באפשרות מתוך הספרייה שנקראת אדם המעריכה ירידה באופן של גרדיאנט ומומנט ועל שניהם יש פרמטר להבעת עוצמה של כל רכיב.  
ניתן להשתמש רק בירידה לפי גרדיאנט או במומנט.  
העדפנו את "אדם" כי להערכתנו הוא מתייחס במידה פרופורציונאלית ללמידה מתוך המידע, ומצליח להתמודד עם דוגמאות שהן חריגות, ומצד שני מתכנס מספיק מהר כדי שנוכל לקבל תוצאה בזמן סביר.

ניתן לנסות ולהשתמש בשילוב של מודלים מסורתיים ועדכניים, למיצוי מאפיינים.

המודל שנבחר ללמידת המערכת, מתבצע באמצעות מערכת לזמן קצר וארוך, ניתן לבחור חלונות זמן ארוכים שונים וכך לנסות למצוא התאמה מיטבית.

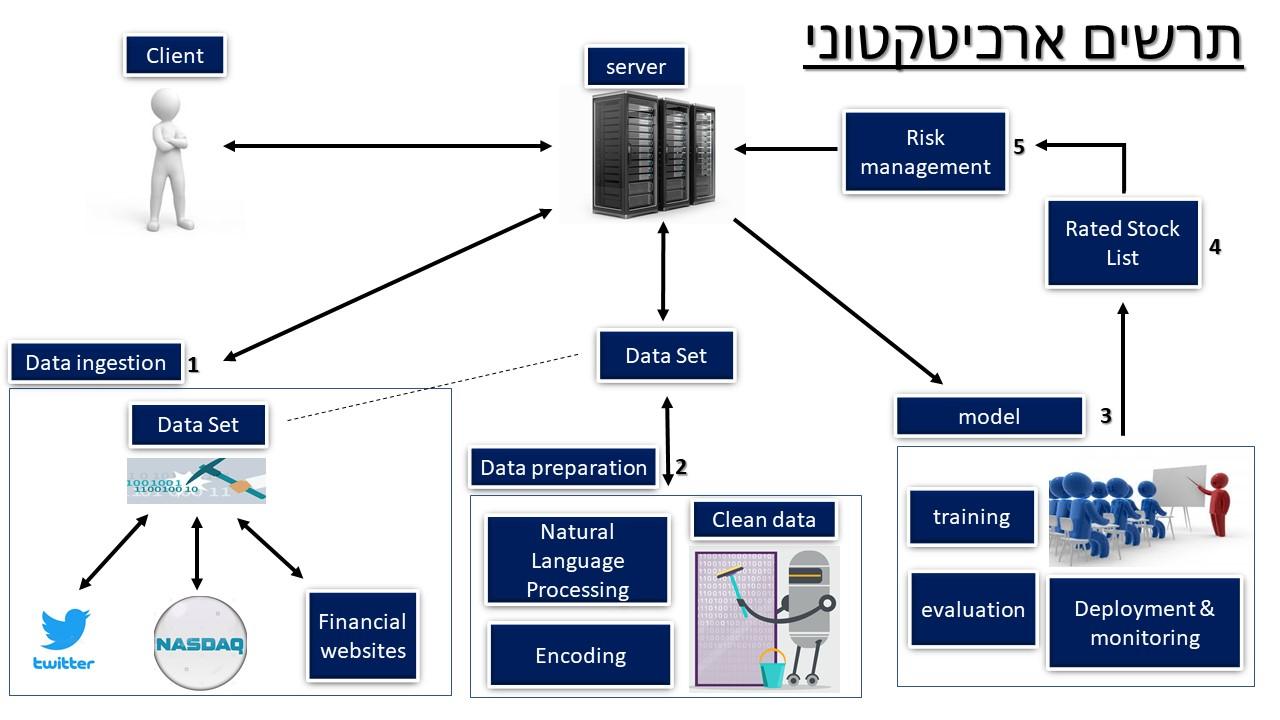
כעת נבחר חלון של חודשיים - 60 ימים עם מחיר סגירה יומי.

מבחינת שכבת הנוירונים נבחרו שתי שכבות האחת בגודל 50 נוירונים והשנייה 1, ניתן לנסות גודל שונה במקום ה-50 ו/או להוסיף שכבות. בכפוף לכוח המחשוב ובמטרה לקבל יחס רווח לסיכון גבוה יותר.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **שם הספרייה** | **תחום התמחות** | **פופולריות** | **שימוש עיקרי** | **עדכני** | **דירוג התאמה** |
| TensorFlow | רשתות נוירונים | 5 | גרפים | 5 | 4 |
| **Scikit-learn** | **למידת מכונה, למידה עמוקה** | **4** | **נתונים** | **4** | **5** |
| PyTorch | למידת מכונה ורשתות נוירונים | 3 | גרפים | 5 | 2 |
| Keras | למידת מכונה | 4 | הדגמה | 4 | 1 |
| Shogun | למידת מכונה | 1 | אקדמי | 1 | 3 |
| Torch | למידת מכונה | 3 | חברות ענק | 2 | 3 |

1. **ארכיטקטורה:**

**תיאור הארכיטקטורה:**  
השלב הראשוני הוא משיכת המידע משוק ההון האמריקאי, פעולות בבורסה של בעלי העיניין וניתוח מדיה חברתית.  
לאחר משיכת המידע המערכת תבצע עיבוד למידע, ניקוי שלו, עיבוד הטקסטים המרת כלל המאפיינים השונים בקלט לערכים, מודל לניבוי מחיר המניה באמצעות למידת מכונה, דירוג המניות.   
ולאור הגדרות המשתמש המערכת תבחר על איזה מניות לבצע פעולות מסחר כאשר בסוף התהליך נוכל להשוות את התשואה של התיק אל מול פעולות המסחר באלגוריתם ללא ניתוח מדיה חברתית.



**Use case**

1. תרחיש 1
   * שם התרחיש: הרשמה למע'
   * שחקנים: לקוח(לקוחות פרטיים המעוניינים להשקיע), מערכת
   * מטרה: לאפשר ללקוח להירשם למערכת
   * מסלול עיקרי-
     + - לקוח נכנס למערכת ומבקש להרשם.
       - לקוח מזין את פרטיו האישיים.
       - לקוח בוחר הגדרות המותאמות אישית לצרכיו(נתונים אישיים, כמות עסקאות יומית, אחוז סיכון, אחוז כסף).
       - המערכת שומרת את פרטי התחברות המשתמש והגדרותיו בבסיס הנתונים.
   * מסלול חלופי 1-
     + משתמש מזין פרטים לא נכונים (תקינות קלט)
     + המערכת מודיעה למשתמש ומבקשת קלט חוזר
   * מסלול חלופי 2-
     + משתמש לא מגדיר הגדרות אישיות
     + המערכת משתמש בהגדרות שנקבעו מראש.

1. תרחיש 2:
   * שם התרחיש: שינוי פרטי לקוח
   * שחקנים: לקוח(לקוחות פרטיים המעוניינים להשקיע), מערכת
   * מטרה: לאשר לקוח לשנות פרטים והגדרות
   * מסלול עיקרי:
     + לקוח מבצע התחברות למערכת.
     + הלקוח דורש לשנות את פרטי האישיים.
     + הלקוח בוחר הגדרות המותאמות אישית לצרכיו(נתונים אישיים, כמות עסקאות יומית, אחוז סיכון, אחוז כסף).
     + המערכת שומרת את פרטיו האישיים של הלקוח בבסיס הנתונים.
   * מסלול חלופי 1:
     + הלקוח משנה פרטים אישיים בקלט לא תקין
     + המערכת מדווחת על קלט לא תקין ומבקשת להכניס את הנתונים בשנית
   * מסלול חלופי 2:
     + הלקוח משנה הגדרות מסחר לא חוקיות
     + המערכת מדווחת על קלט לא תקין ומבקשת להכניס את הנתונים בשנית
2. תרחיש 3:
   * שם התרחיש: התחברות למערכת
   * שחקנים: לקוח(לקוחות פרטיים המעוניינים להשקיע), מערכת
   * מטרה: לאפשר התחברות למערכת
   * מסלול עיקרי-
     + לקוח מבצע התחברות למערכת.
     + מערכת בוחנת את פרטי ההתחברות שהזין הלקוח ומחפשת התאמה.
     + המערכת מעבירה את הלקוח אל דף הבית המותאם אישית לצרכיו.
   * מסלול חלופי-
     + אין התאמה במערכת על נתוני ההתחברות
     + המערכת מבקשת להכניס את הנתונים שנית כי אין התאמה
3. תרחיש 4:
   * שם התרחיש: מסחר בבורסה
   * שחקנים: לקוח(לקוחות פרטיים המעוניינים להשקיע), בורסה
   * מטרה: לאפשר למתשמש לקבל תמונת מצב
   * מסלול עיקרי:
     + לקוח מבצע התחברות למערכת.
     + המערכת סורקת את נתוני המשתמש ומחפשת הזדמנויות עסקיות עבורו בעזרת נתונים מהבורסה וממדיות חברתיות.
     + המערכת מבצעת פקודות קנייה/מכירה באופן אוטומטי בהתאם להגדרות המשתמש.
     + המערכת מעדכנת במסך הבית את המשתמש על פעולות שהיא מבצעת, כמות הרווח/הפסד בכל רגע נתון.
   * מסלול חלופי:
     + הלקוח לא מצליח להתחבר
     + המערכת מודיעה על אי התאמה ומבקשת להתחבר בשנית
   * מסלול חלופי 2:
     + המערכת סורקת נתונים (כריית מידע) ולא מקבלת את המידע המבוקש
     + המערכת מתריאה איזה מקור מידע לא תקין ולא ממשיכה במסחר עד לפתרון התקלה.
   * מסלול חלופי 3:
     + המערכת לא הצליחה לבצע את הפקודות קניה/ מכירה.
     + המערכת מעדכנת את הפעולה שהיא לא הצליחה לבצע.

תיאור האלגוריתמים בסעיף 11 בתכן.

1. **תכן הנדסי מפורט:**

* כריית מידע: כתיבת זחלן בשפת Python, עם ספריית Selenium הנותנת רמה מספקת של כלים לכריית מידע מהאתרים שנבחנו עד כה. קיימת ספרייה מובילה מקבילה של Beautifull Soup היודעת לבצע את אותן פעולות בגישה שונה.
  + הכרייה באופן כללי מבוצעת ע"י העלאת מדמה דפדפן אוטומטי מסוג Chrome, הטוען את כל המידע דרכו וניתן לבצע פעולות באופן אוטומטי, גישה קלה ונוחה לSourceCode המורכב מHtml.
  + רציונל הכרייה מסתמך על שפה המרכזית בעולם Html, שפת תגיות לתצוגה ועיצוב דפי אינטרנט ותוכן לתצוגה בדפדפן. השפה מאופיינת בכתיבה עם מבנה אחיד עם קווים מנחיה לכתיבת קוד ברורים ומסודרים אשר לא משתנים בין אתר לאתר. תצוגת המידע והאופן בו הוא מאוכסן שונים והאתגר הוא להבין את הצורה שבה נכתב האתר והיכולת למשוך את המידע באופן המהיר ביותר.
  + כלל מנחה מרכזי בכריית המידע הוא- אם המידע נגיש לי ויזואלית, אני יכול למשוך אותו. לדוגמא משיכת טבלה:
    - אם האתר העלה טבלה לאתר בצורה "פשוטה" נוכל למשוך למשוך אותה בעזרת שורות בודדות (כמו באתר gurufocus)
    - באתר אחר אותה טבלה מונפקת על ידי java script וכדי לכרות אותה צריך להריץ את הscript עם ערכים מסויימים, לקבל תצוגה של הטבלה המוצגת ולאחר מכן ללמוד את המבנה שבו המידע מוצג ולהתאים לו זחלן שיודע למשוך את המידע ערך ערך.

לכן כל אתר צריך להיבחן תחת זכוכית מגדלת כדי להבין את אופן המשיכה היעיל ביותר

מעבר לשימוש בספרייה המרכזית, בוצע שימוש באמצעים נוספים:

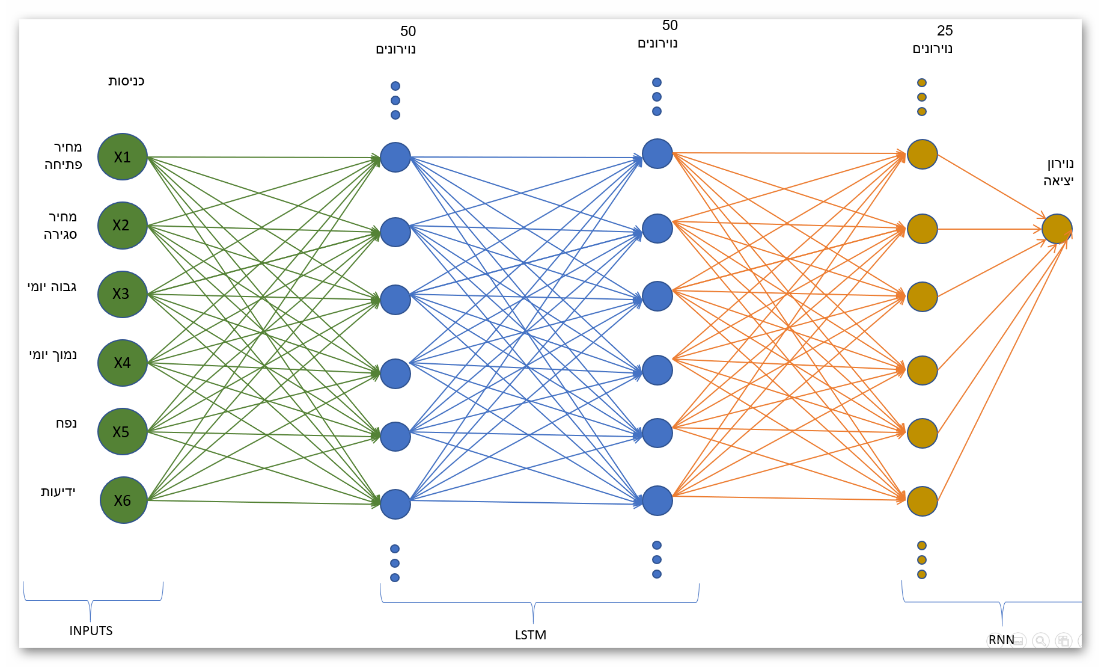
* URL manipolation (URL rewriting)- באתר gurufocus, עם הכנסת הפרמטרים לחיפוש עם קבלת תוצאות החיפוש , ניתן לראות את הפרמטרים שאותם הגדרנו לחיפוש בצורה מאוד ברורה ומובנת (עם ניסוי וטעיה) בשורת הUrl, ולאחר בדיקות קצרות ניתן היה גם להבין שניתן לקבל תוצאות מעבר למה שמוגבל בחיפוש (למשל לקבל 5000 תוצאות במקום 100). לכן בוצע התאמה של קוד היודע לייצר את הכתובות המבוקשות כמו שהמתשמש הזין באתר אבל בצורה אוטומטית. מחשבה זאת חוסכת מאיתנו את הצורך בהפעלת הdrive (והכנסה ידנית אוטומטית של הפרמטרים לחיפוש) ותורמת להקטנת את הסיכון בחשיפה של האתר למערכת אוטומטית ולא משתמש אנושי.
* באתר gadgetguru, זוהה שהמידע המבוקש מוצג בhtml בטבלה מסודרת ופשוטה. עם הזיהוי הנכון ניתן היה למשוך את הטבלה עם פקודת read\_html של ספריית Pandas ( המיועדת לניתוח ועיבוד נתונים, מבני נתונים נומריים)
* משיכת מידע על מניות מאתר Yahoo, בוצע על ידי פקודת Datareader של ספריית Pandas היודע למשוך מידע מאתרים נבחרים ישירות לתוך מבנה הנתונים.

אופן פעולת הכרייה:

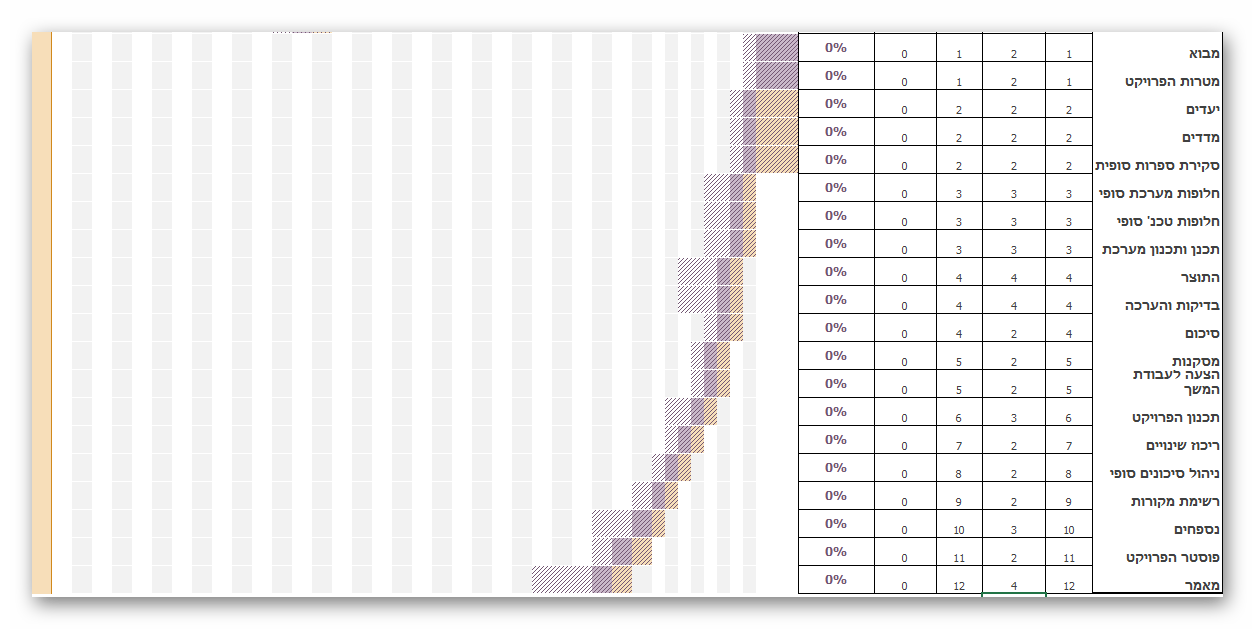
* + Yahoo finance-
    - הגדרת זמן אחרון למשיכת מידע כיום הפעלת הזחלן. (תאריך)
    - משיכת מידע של כל מניה לפי תאריכים עם מתודה Datareader של pandas הפועלת מול yahoo API.
    - טיפול במידע נוסף לנוחויות טיפול בנתונים (שם המאגר מידע ושם המניה)
    - שמירת המידע בdata frame של pandas
    - איחוד כל מאגרי המידע לקובץ אחד.
  + OpenInsider-
    - הכנת מאגר url על ידי שימוש במתודה make\_urls
      * Make\_urls:
        + קבלת פרמטר של תאריך לכרייה מתחילת המידע של האתר (החודש הנוכחי/ כל המידע שקיים/ כל המידע עד לחודש נוכחי (לא כלול) - ניתן לשינוי.
        + בניית כתובת url המורכבת מהתאריכים המבוקשים וכמות תצפיות של 5000 (במקום 100)
        + החזרת רשימה של כתובות
    - הרצת איטירטור על כל הכתובות
      * משיכת הטבלה המבוקשת מכל אתר בעזרת פקודת read\_html של pandas
    - איחוד כל הטבלאות לטבלה וקובץ אחד.
  + Stocktweets-
    - מעבר על רשימת המניות שרוצים לכרות, כל מניה:
      * העלאת דרייב של chrome, המדמה אתר עם הכתובת של המניה באתר.
      * גלגול לתחתית העמוד כדי לטעון ידיעות חדשות (כמות הטעינות מחושב לפי כמות הידיעות שהמערכת הגדירה למשוך)- כל גלגול נותן בערך 22 ידיעות נוספות)
      * שמירת כל הידיעות ברשימה.
      * ניקוי כל ידיעה ושמירה של הנתונים הבאים:
        + שם מניה.
        + שם משתמש- חיפוש בידיעה את שם המשתמש.
        + זמן הידיעה-שימוש במתודת עזר מיוחדת שממירה את המידע לגבי זמן הידיעה ("now/1-59m ago/XX:XX AM/PM/year-month-day XX:XX AM/Pm) לתאריך מסוג year-month-day)
        + טקסט הידיעה- שימוש במתודת עזר היודעת לנקות את הרעשי רגע בידיעה ולהוציא את הטקסט עצמו.
      * שמירת המידע html source code בקובץ לוקלי- אופציה לבחינת שיפור עתידי.
      * החזרת הידיעות וסידורם בdata frame וקובץ
* עיבוד מידע:
  + שימוש בספריית NLTK המספקת זיהוי יישויות, פירוק משפט לגורמים ומתן ציון סנטימנטלי למשפט(חיובי/שלילי/נייטרלי), במדד רציף הנע בין 1- ל 1.
  + בשלב הראשוני נאספו כ150 ידיעות, אשר להן ניתן ציון סנטימנטלי בצורה ידנית.
  + המודל נבנה כך שהוא מסתמך על מפיץ הידיעה, המנייה עליה נכתב הידיעה, תאריך הידיעה, מתן פידבק האם החשבון הינו חשבון מאומת ברשת חברתית כלשהי(בשלב זה הבדיקה נעשתה מול Twitter).
  + המודל בחן את הידיעות ונתן להם ציון המורכב מ4 פרמטרים: חיובי, נייטרלי, שלילי וסכם של כל השלושה.
  + בשלב האימון, בוצע אימון ב2 שלבים, תחילה המודל היה צריך לחזות את הציון הסנטימנטלי הידני שניתן ע"י, ולאחר מכן את ציון הסכם שניתן ע"י הספריה.
  + בבחינת של מס' מודלים ליניאריים ומודלים נוספים, בבחינה ראשונית החיזוי היה בסבירות של מעל ל70%, כאשר השאיפה היא להגיע ל90%.

מידול מידע:

* + מערכת RNN + LSTM, מידע נאסף, מועבד ונכנס למערכת למידה LSTM בעלת שתי שכבות בנות 50 נוירוינים לאחריו לשתי שכבות ברשת נוירונים 25, 1, בהתאמה במטרה לקבל פלט יחיד לכניסה מרובה, הפלט הינו מחיר הסגירה ליום המסחר הבא.   
    ניישם המודל על מספר מניות, נפעל להרחיב את כמות המאפיינים ונשנה בהמשך את קצב הלמידה, כמות המקבצים, כמות הנוירונים או אפילו מספר השכבות. במקביל נבחן לפחות מודל אחד נוסף להשוואה.
  + המידע יתכנס לפי ספריית קרס.



1. **תוכנית עבודה סופית מעודכנת:**



1. **ריכוז שינויים מאז דוח התכנון:**

* כריית מידע, בניגוד להגדרה של 20 טקסטים מ2 מקורות לפחות, כרגע היכולת הראשונית היא הרבה יותר גדולה.
  + כדי לתת למערכת בסיס טוב של למידה, כרגע המערכת יודעת למשוך מספר רב של ידיעות טוויטר בכל יום על כל מניה כאשר המגבלה למשיכה (כעת אל מול האתרים שנבדקים- stocktwits, gurufocus) הוא הזמן שלוקח למשוך את המידע. (באזור השלוש שעות ל2000 טוויטים)
    - צריך להגיד שכמות הטווטים היא גדולה אך פחות איכותית ממשתמש מנוסה או בעל ידע מקצועי המכיר את השוק מאשר אדם רגיל שמעלה ידיעה באופן שרירותי.
    - המערכת תדע עם הזמן לדרג את המשתמשים ולהבין למי לתת משקל גדול יותר על ידי סיווג של אמין/ לא אמין.
  + משיכת הדיווחים על מכירות וקניות של הבעלי עיניין מאתר gurufocus, מאוד פשוטים (לאחר ניתוח והבנה איך האתר בנוי), ניתן למשוך במספר שניות את העדכון של החודש האחרון, ובמספר דקות את כל ההיסטוריה של המידע.
* עיבוד מידע:
  + בהיבט זה, בחרנו בספריית NLTK, אשר מבצע את הפונקציונליות הדרושה לנו, תוך מתן ביצועים גבוהים ויעילים, ובעלת קהילת משתמשים רחבה. עיבוד המידע יכלול הן את ציון הסנטימנטלי לטקסטים, והן את ההיבט הסטטיסטי/ הניתוח הטכני.

מידול מידע:

* בחרנו במודל מסוג RNN+LSTM, מודל זה משלב מערכת למידה מזמן קצר - היום האחרון וחיזוי על פני חודשיים, לאחר מכן לרשת נוירונים אשר מוציאה פלט יחיד מחיר סגירה עבור היום הבא.
  + יבחנו בהמשך מודלים נוספים, לטווח קצר ולטווח ארוך בנפרד.
  + נבדקו בסקירת הספרות שלושה מודלים אחרים, מסורתיים יותר כגון SVM אשר הניבו תוצאות לא חד משמעיות, אנו נשתדל להפיק תוצאות חד משמעיות, אם כדאי או לא להשקיע במניה מסוימת.

1. **עדכון סטטוס ניהול וגידור סיכונים:**

כריית מידע:

* גישה למידע- כריית המידע מתבצעת על בסיס קוד האתר (html- source code), כאשר יש שוני בין אתר לאתר ומה שיעבוד על אתר אחד, לא יתאים לאתר השני. לכן יש להתאים את סגנון הכרייה מכל אתר בנפרד.כל אתר יכול גם לשנות את האופן בו הוא מציג מידע ולכן
  + יש צורך לבצע בקרה ומעקב באופן תקופתי וקבוע אחריי קוד האתר ותצורת כריית המידע (האם האתר שינה את העיצוב, צורת התצוגה או צורת האחסון), כי בלי כרייה נכונה המערכת לא תדע ללמוד את המידע.
  + עם הגשת דו"ח הבניים, אתר stocktweets הוסיף פרסומת קופצת פנימית כאשר ניגשים לכתובת המניה באתר, דבר המבטל את פעולת הכרייה.  
    בקוד המוצג פה בנספח של הקוד אין התייחסות לכך, אבל הקוד כבר תוקן והוכנסה מתודה הבודקת אם יש פרסומות קופצות ומאפשר לזחלן לעבוד כהלכה).
* זיהוי כרובוט- אופן כריית המידע רגיש מאוד אל מול כל אתר בנפרד, כל חברה או אתר מעוניינים להגן על השרת שלהם מפני מספר רב של בקשות אוטומטיות. בשלב מסויים ניתן לדמות לאחר הרבה בקשות את הדימיון למתקפת DDOS. לכן ישנה הגנה אוטומטית ברמה זו אחרת של כל אתר, אם האתר ידע לסמן את הזחלן כרובוט אוטומטי ולחסום אותו. כדי להימנע מהסיטואציה יש לבצע פעולות שונות אל מול כל בקשת מידע מהשרת. בעיקר ניסוי וטעיה ואלו הכללים המנחים:
  + השהיות יזומות בין הבקשות- דימוי פעולה של אדם ממוצע ולא רובוט העובד באופן אוטומטי ולפעמים מהר מדי עוד לפני שהדף נטען.
  + קצב לחיצות- בנוסף לשהיות בין הבקשות, לחיצות על הכפתורים וקישורים באתר בצורה איטית יחסית ולא מהירה. ובדיקה של כפתורים אמיתיים ולא ללחות על כפתור/קישור שקיים בקוד ולא באתר שהמשתמש רואה.
  + שינוי מזהה דפדפן- החלפת מדמה הדפדפן(driver)- Chrome, Firefox פרט מידע זה בנוסף לכתובת יכולה להגביר את החשד לזחלן רובוטי.
  + בקשות לשרת דרך שרת proxy- לשנות את הכתובת המבקשת את הבקשה מהשרת.
* שינוי ספריות- מעקב אחרי עידכונים ושינויים של הספריות איתם אנחנו עובדים. ספרייה מסויימת יכולה להפסיק שירות מסויים ויהיה צורך לשנות את הקוד בהתאם או לחליף לספרייה אחרת ולמצוא פתרון אחר.
* חוקיות כריית מידע- כמו שציון בתחילת המסמך, כרגע חוקית כריית מידע המוצג ללא גישה בתשלום מאושר לכרייה, אך עם זאת יש הגבלות (לא לזחלני כריית מידע, אבל כן לבוטים של חיפוש, כמו גוגל) שמכבדים את קובץ הrobots.txt של כל אתר לגבי הצגת האינדקס שלו באתרי החיפוש. מתוך השפע של המידע ופרויקט הלימודי הוחלט לעמוד גם בתנאים של הrobots.txt, וכמובן לאסוף מידע שזמין לציבור בחינם.

עיבוד מידע:

* המודל הטקסטואלי עלול להסתמך על טקסטים שאינם משמעותיים מספיק על מנת להשפיע על שוק ההון בפועל, להפיק ממנו תובנות שגויות, אשר יגרמו לביצוע פעולות קנייה/מכירה לא רצויות.
* המודל הטקסטואלי עלול לא להבין כראוי טקסטים בעלי אופן ציני או בעלי מאפייני רטוריקה אחרים, ולהפיק תובנות שגויות.
* המודל הטקסטואלי עלול להסתמך על אישים/כתבים/ארגונים שאינם אמינים.

מידול מידע:

* מבחינת מודל, יבחנו שני קריטריונים אשר יכולים להוות סיכונים, יחס רווח לסיכון ואחוז שגיאה.
  + הסיכון העיקרי הוא מידת הצלחה נמוכה בעת חיזוי, כאשר יחושב מדד שארפ, כלומר יחושב היחס בין רווח לסיכון וכך תמדד מידת ההצלחה.  
    כדי לצמצם את הסיכון, נבחן מעתה מספר מודלים נוספים ונפעל לבחור את המודל שיניב תוצאה גבוה יותר.
  + אי מציאת קשר ישיר בין עלייה/ירידה של מניה לניתוח ידיעות טקסט, הפתרון שמוצע הוא שימוש במידע סטטיסטי וטכני במקביל להשבחת התוצאות.
  + מבחינת אחוז שגיאה, נבחן אפשרויות שונות בתוך כל מודל, יחס למידה, כמות מקבצי נתונים, שכבות וכמות נוירונים ברשת.

לוח זמנים:

**מילואים נוספים של חבר הפרוייקט, כרגע עומדים מתחילת הפרוייקט על 84.5 ימים של מילואים.**

1. **רשימת מקורות :**

* ספרייה להצגת תרשימים:<https://matplotlib.org/>
* ספרייה בפייתון לסידור מקדים של המידע: <https://scikit-learn.org/stable/>
* ספרייה לשימוש מודלים: <https://keras.io/>
* ספריית פייתון למספרים נומרים: <https://numpy.org/>
* סקר ספרות לבחינת מודלים - <https://www.sciencedirect.com/science/article/pii/S157401371930084X>
* IKNOWFIRST - אתר הבית - [www.iknowfirst.com](http://www.iknowfirst.com/)
* סקר ספרות לבחינת מודל SVM - <https://www.sciencedirect.com/science/article/pii/S1877050916311619>
* is an open platform where readers find dynamic thinking, and where expert and undiscovered voices can share their writing on any topic. **,**[**www.medium.com**](http://www.medium.com/) ,
* A Medium publication sharing concepts, ideas, and codes.[**www.towardsdatascience.com**](http://www.towardsdatascience.com/)
* Selenium WebDriver, <https://www.selenium.dev/>
* Selenium documentaion#2, [**https://www.selenium.dev/documentation/en/**](https://www.selenium.dev/documentation/en/)
* BeautifulSoup, <https://www.crummy.com/software/BeautifulSoup/bs4/doc>
* Scrapy, [**https://docs.scrapy.org/en/latest/**](https://docs.scrapy.org/en/latest/)
* Pandas, [**https://pandas.pydata.org/docs/**](https://pandas.pydata.org/docs/)
* Html, [**https://html.spec.whatwg.org/multipage/**](https://html.spec.whatwg.org/multipage/)
* ChromeDriver, [**https://chromedriver.chromium.org/home**](https://chromedriver.chromium.org/home)
* HiQ Labs, Inc.v. Linkedln Corp, <https://law.justia.com/cases/federal/appellate-courts/ca9/17-16783/17-16783-2019-09-09.html>
* Victory! Ruling in hiQ v. Linkedin Protects Scraping of Public Data,[Victory! Ruling in hiQ v. Linkedin Protects Scraping of Public Data | Electronic Frontier Foundation (eff.org)](https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data)
* The web robots pages, [www.robotstxt.org](http://www.robotstxt.org)
* Can a robots.txt be used in a court of law?, <https://www.robotstxt.org/faq/legal.html#:~:text=Can%20a%20%2Frobots.,in%20a%20court%20of%20law%3F&text=There%20is%20no%20law%20stating,%2C%20but%20having%20a%20%2Frobots>.
* A note on google unsupported rules in robots.txt, <https://developers.google.com/search/blog/2019/07/a-note-on-unsupported-rules-in-robotstxt>

**קישור:** [**https://www.sciencedirect.com/science/article/pii/S1877050916311619**](https://www.sciencedirect.com/science/article/pii/S1877050916311619)

ציטוט: Prediction Models for Indian Stock Market

**Aparna Nayak, M. M. Manohara Pai, Radhika M. Pai,**

**Prediction Models for Indian Stock Market,**

**Procedia Computer Science,**

**Volume 89,**

**2016,**

**Pages 441-449,**

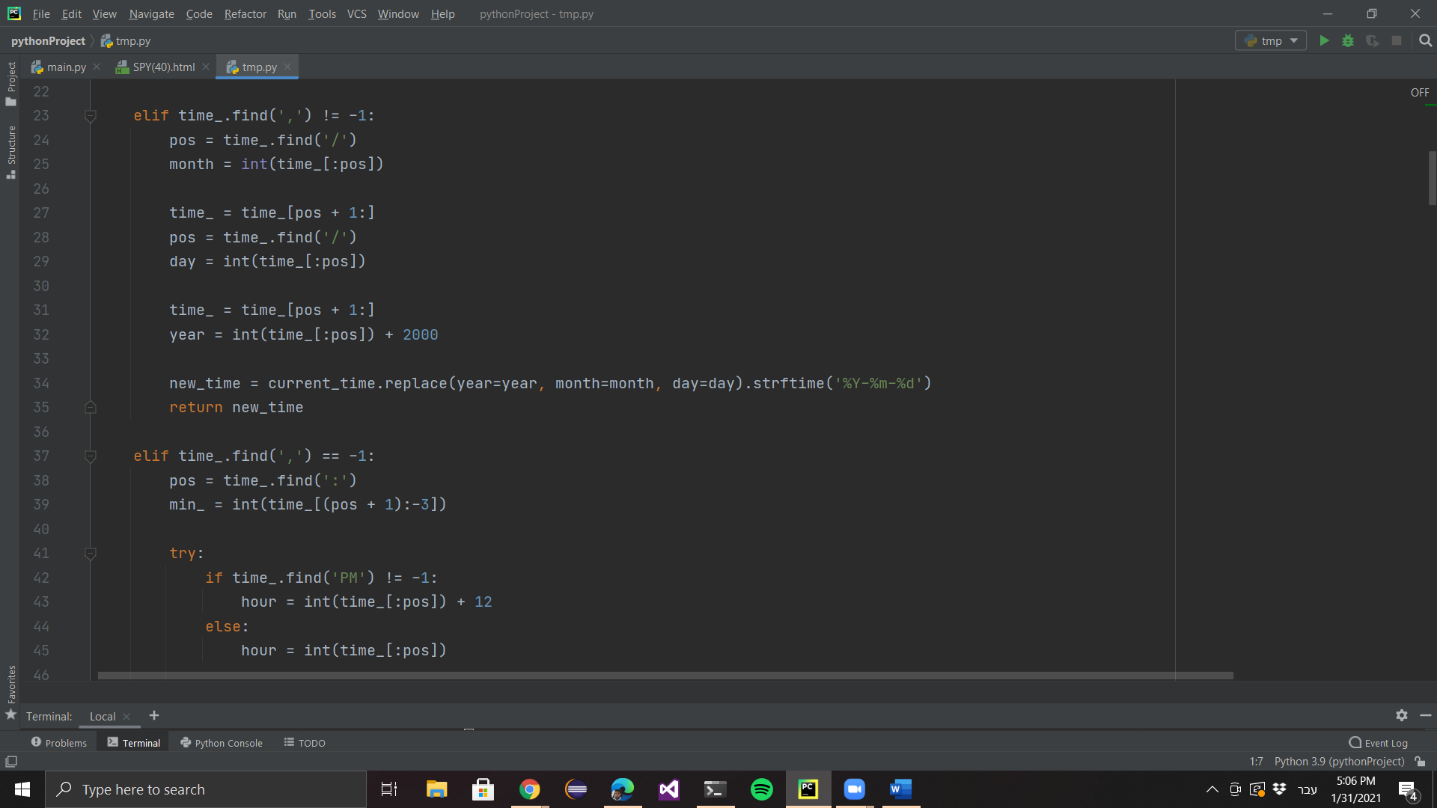
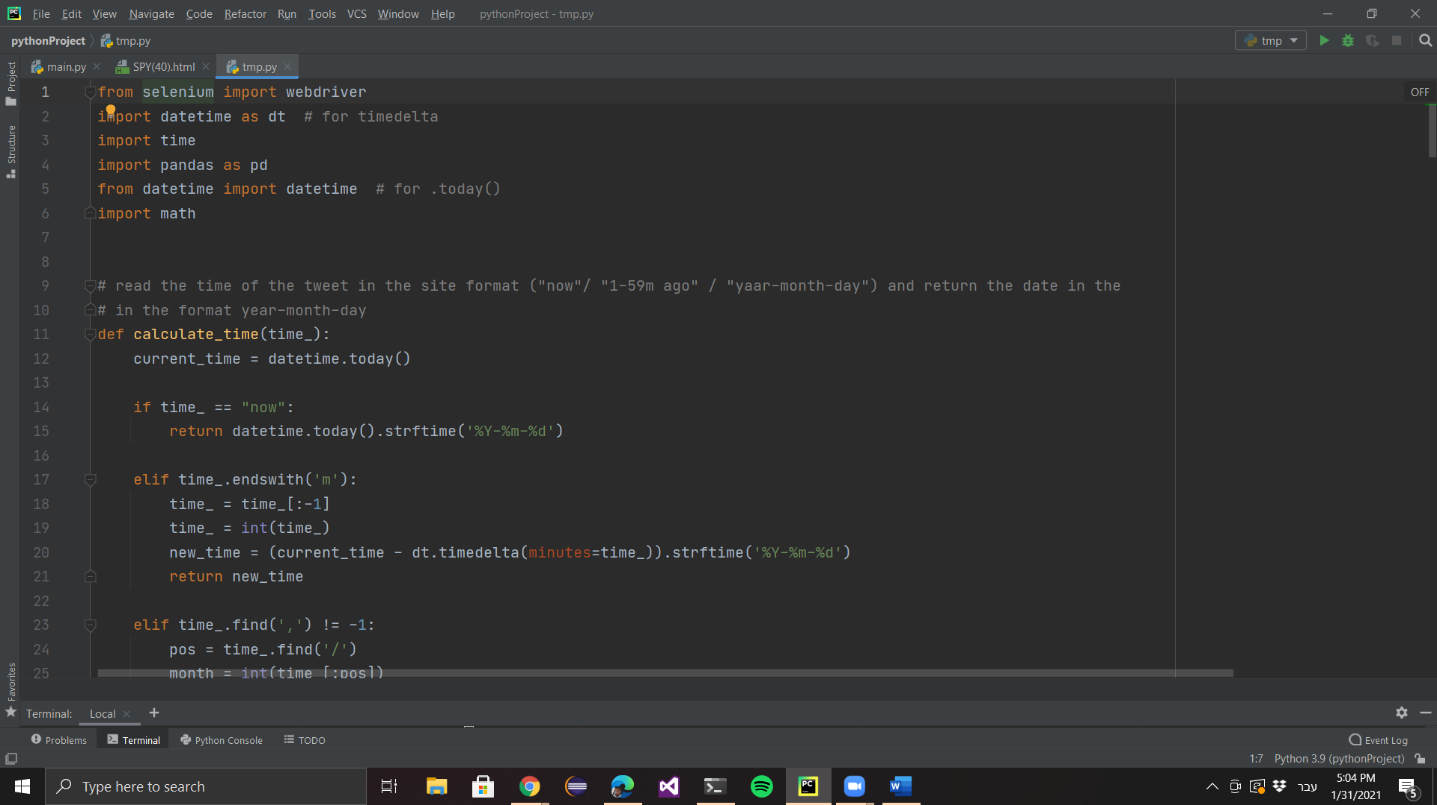
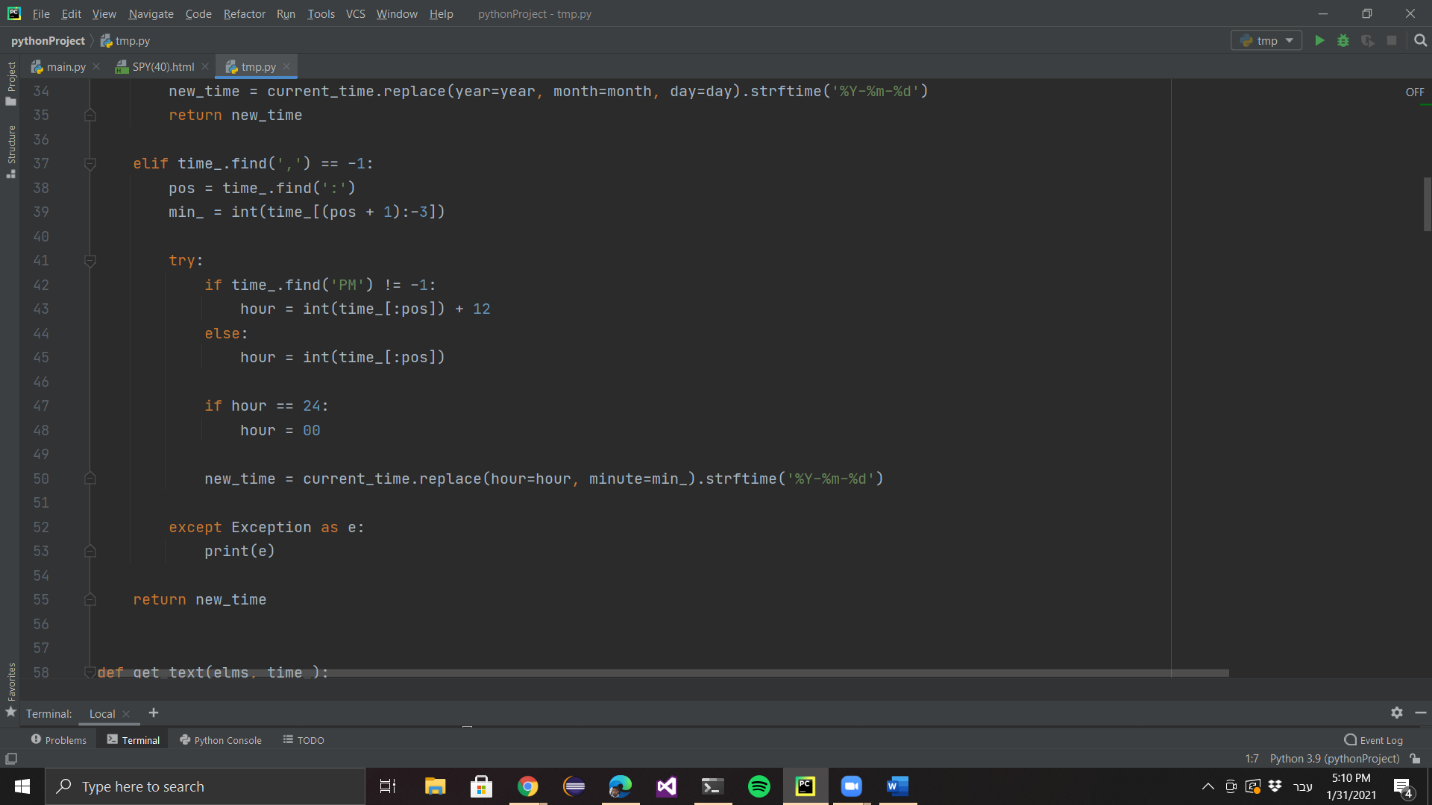
**ISSN 1877-0509,**

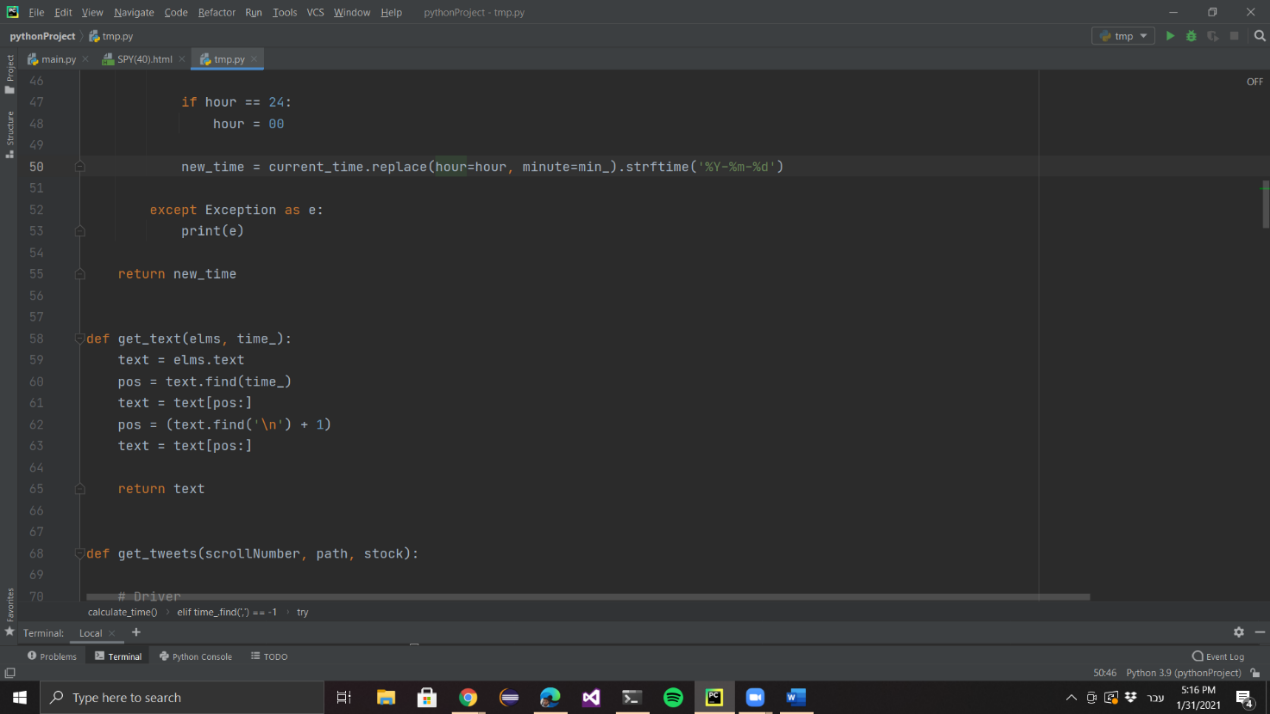
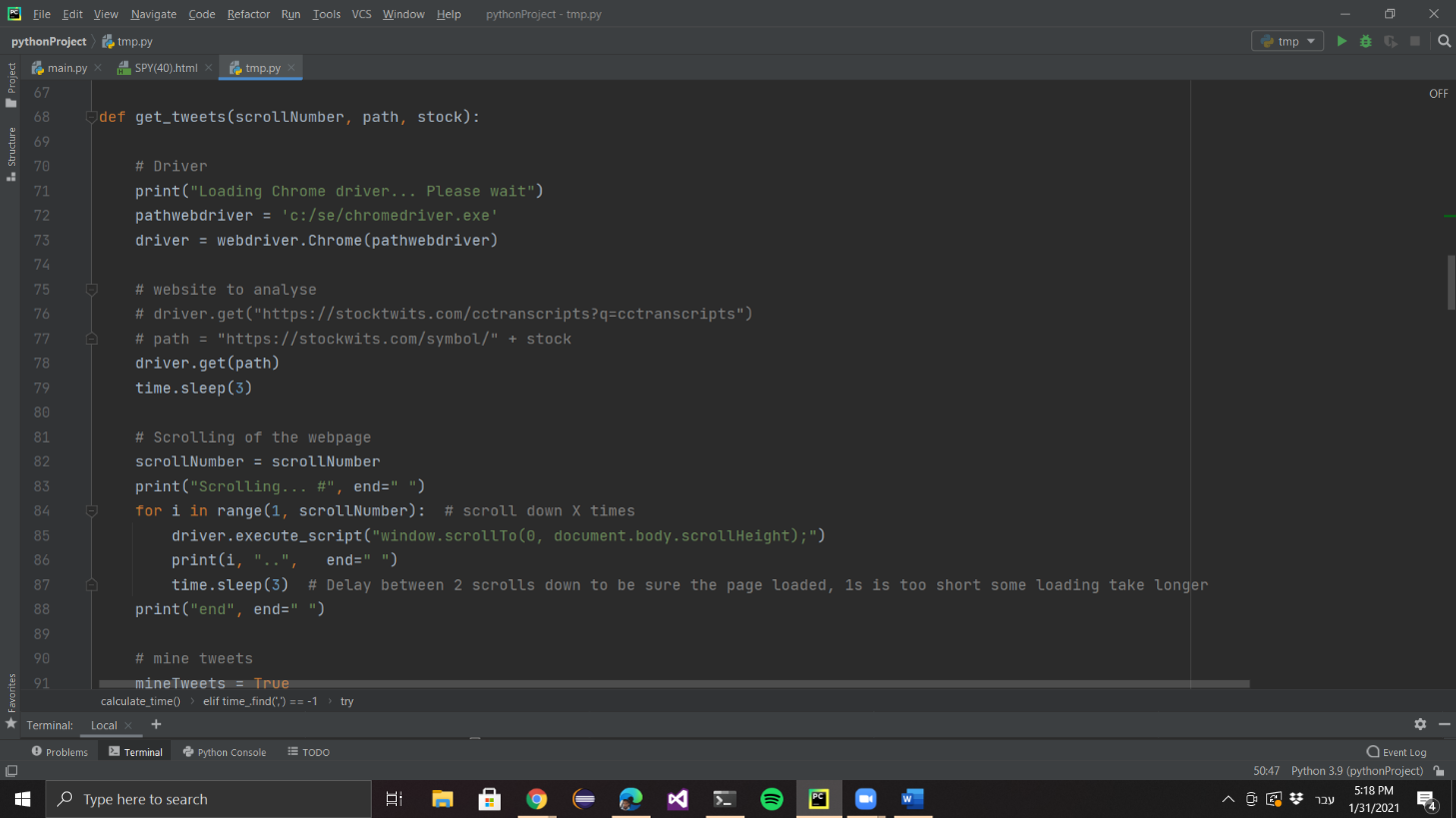
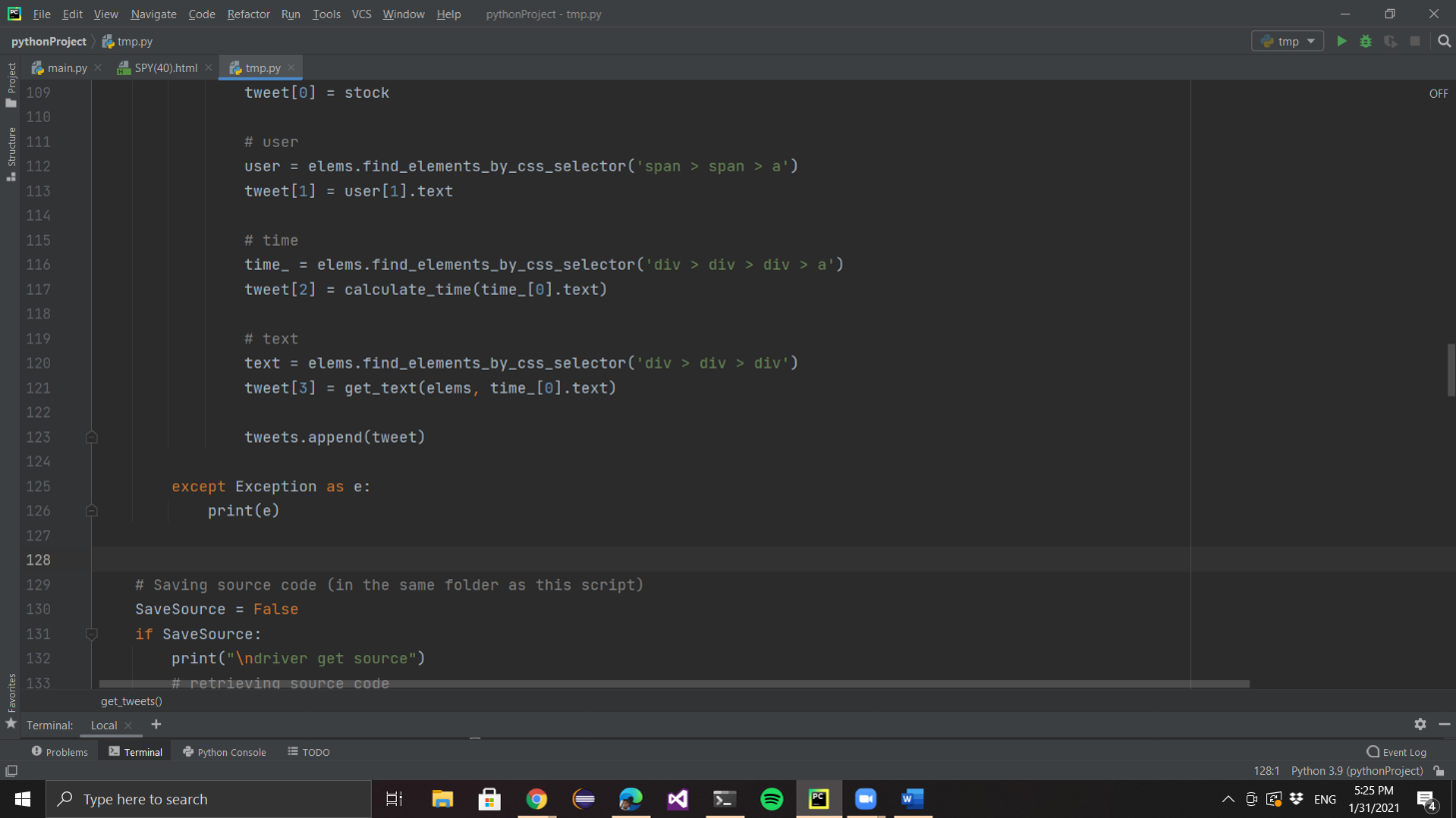
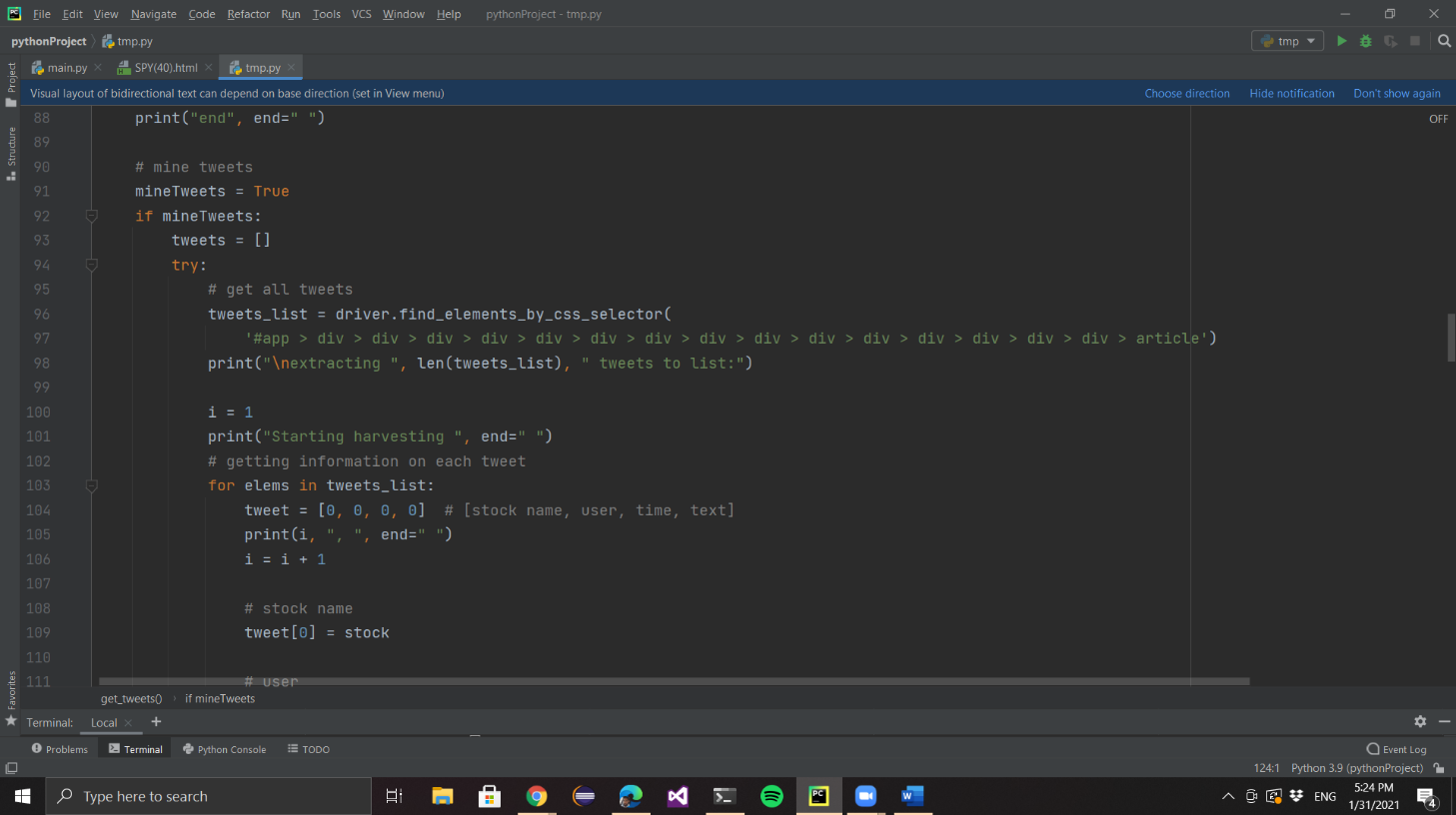
**https://doi.org/10.1016/j.procs.2016.06.096.**

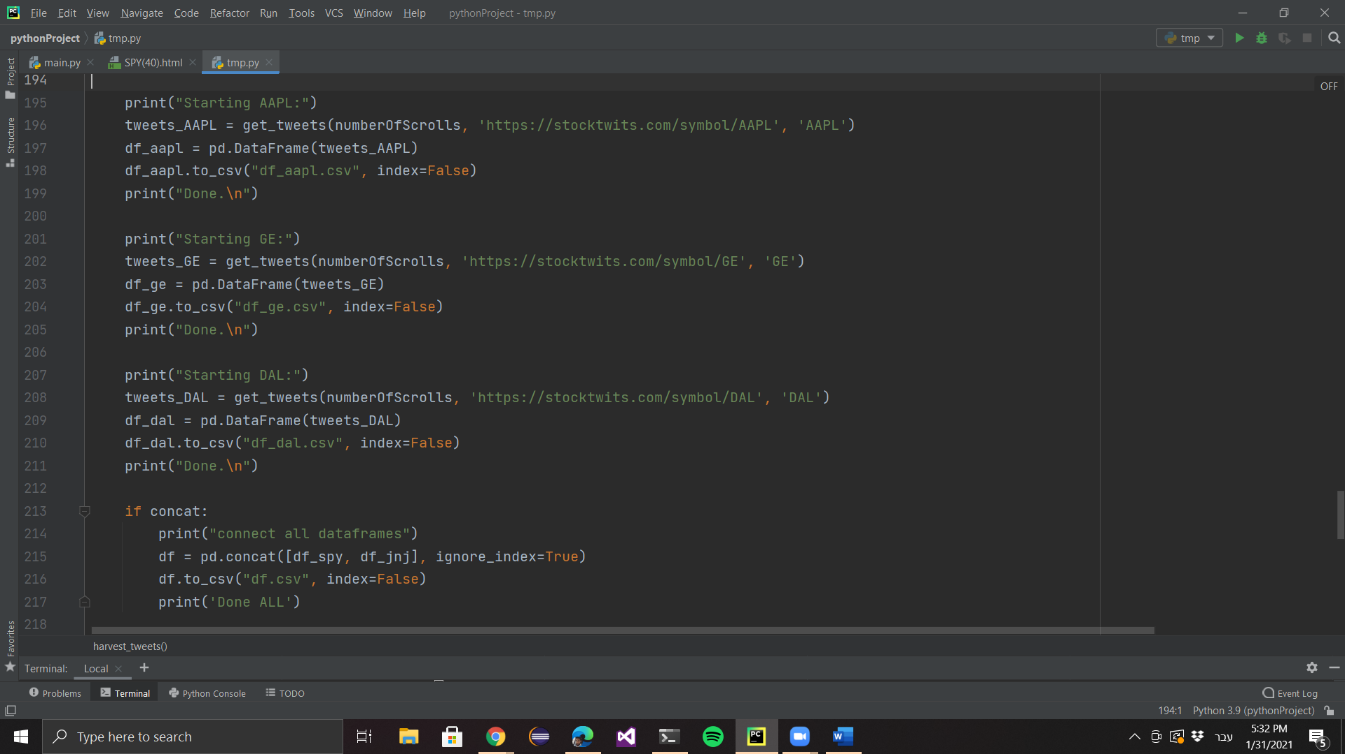
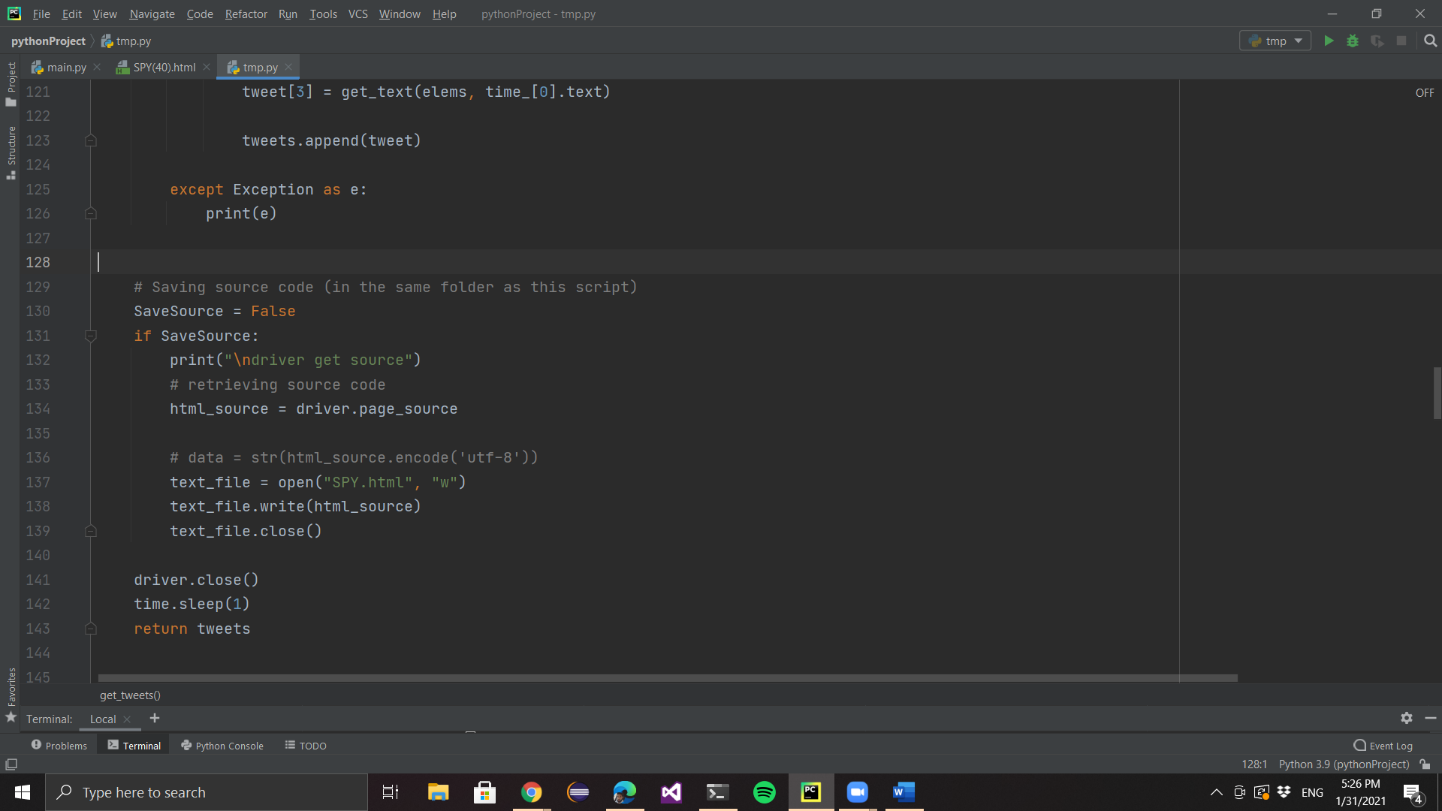
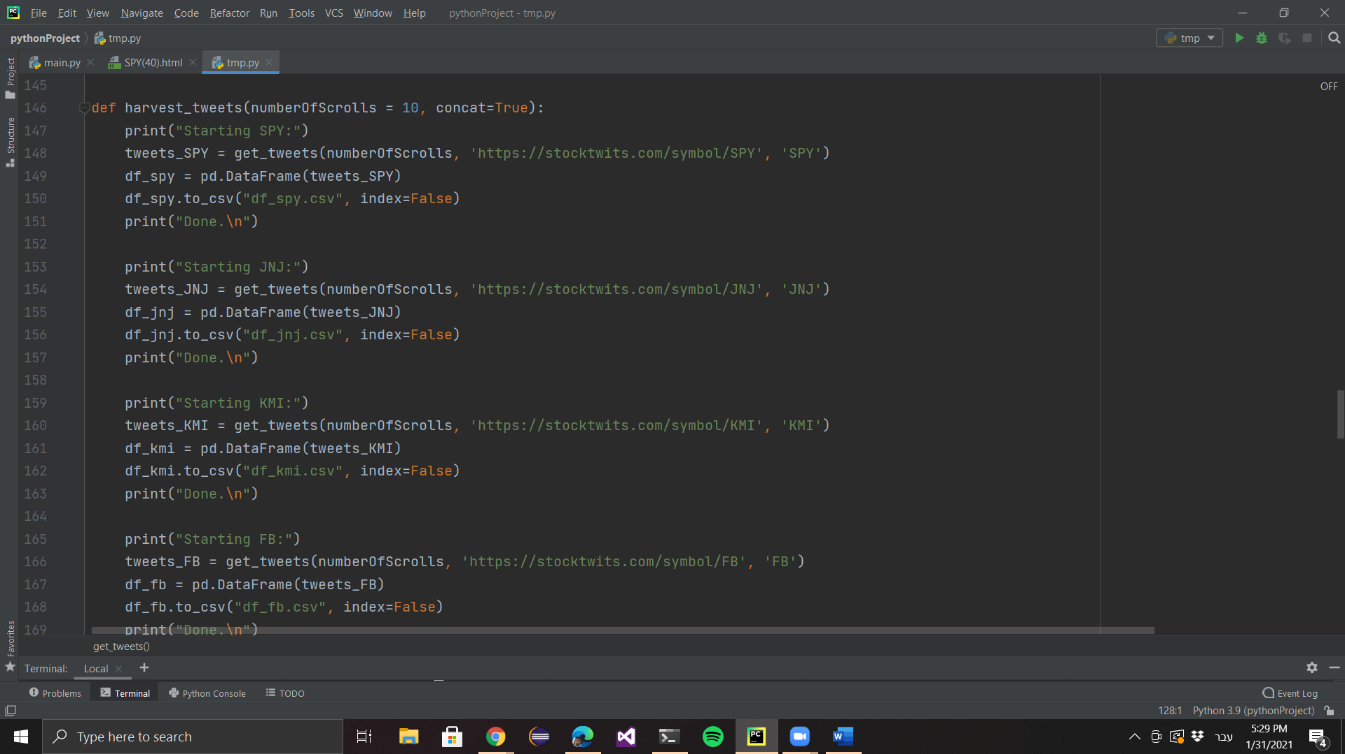
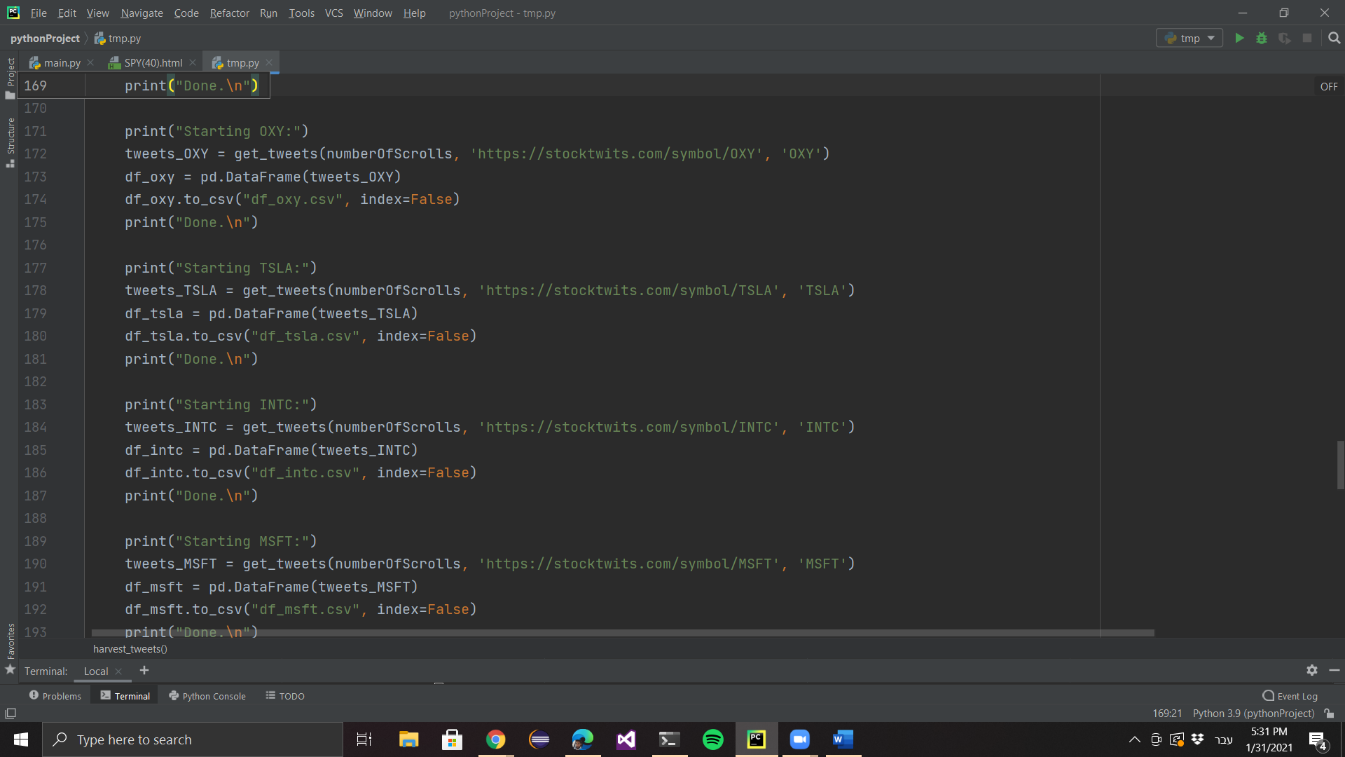
**(http://www.sciencedirect.com/science/article/pii/S1877050916311619)**

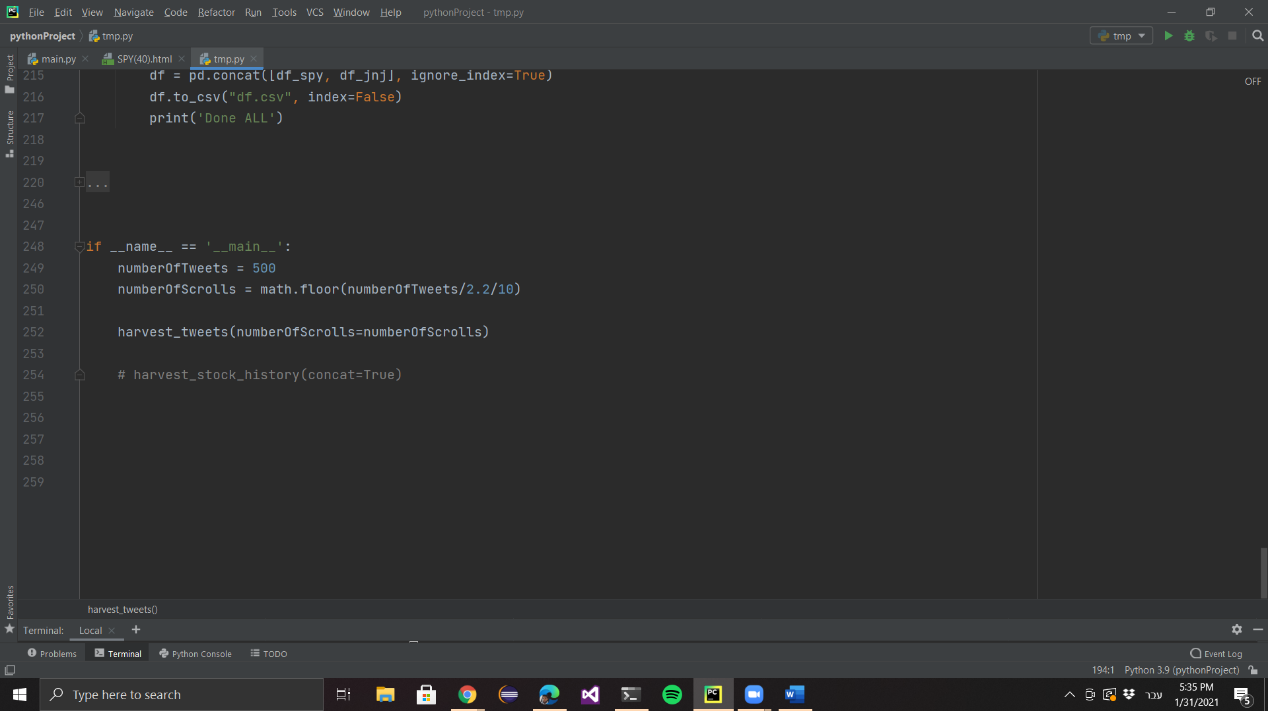
1. **נספחים :**

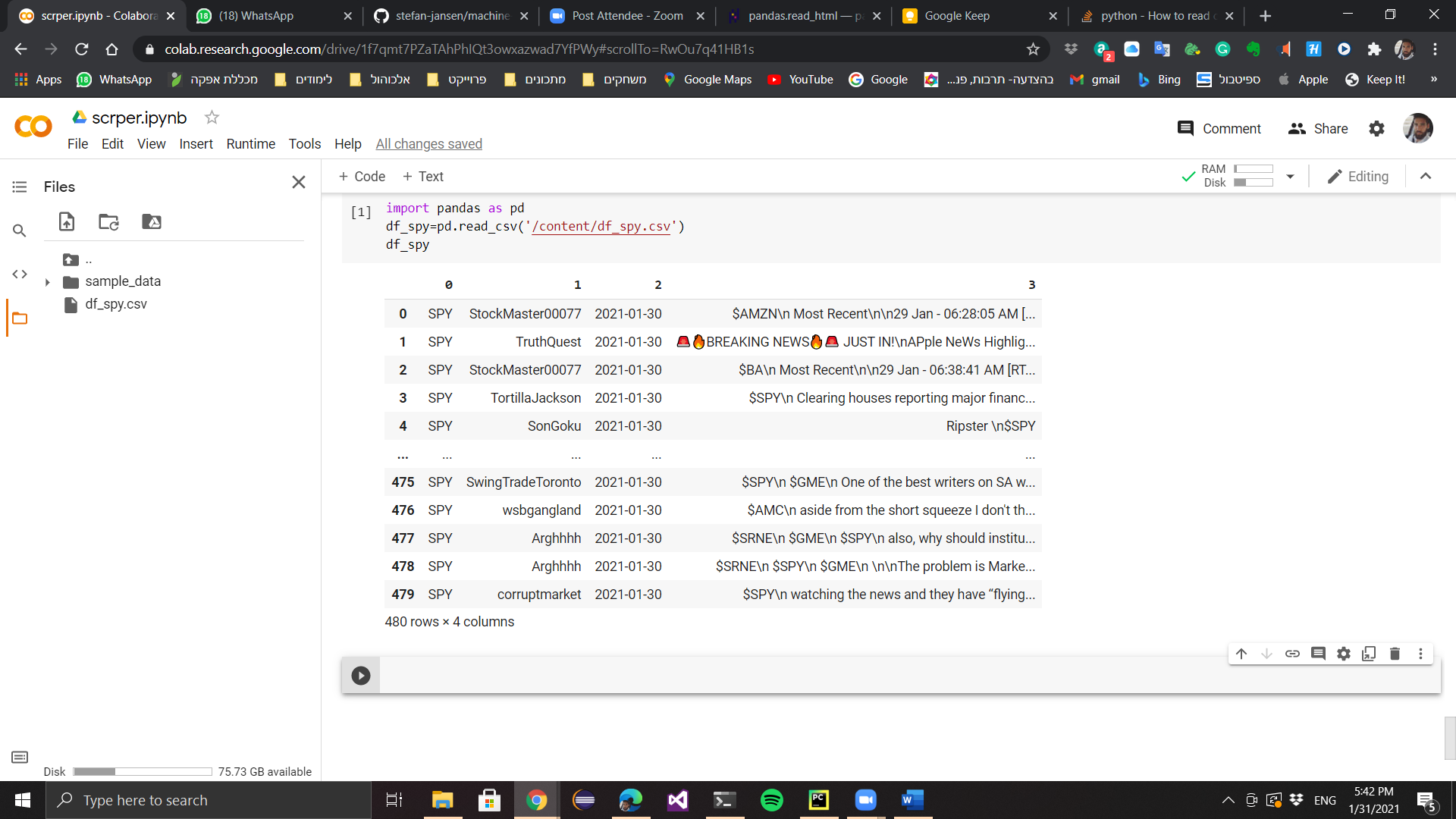
**קוד כרייה:**

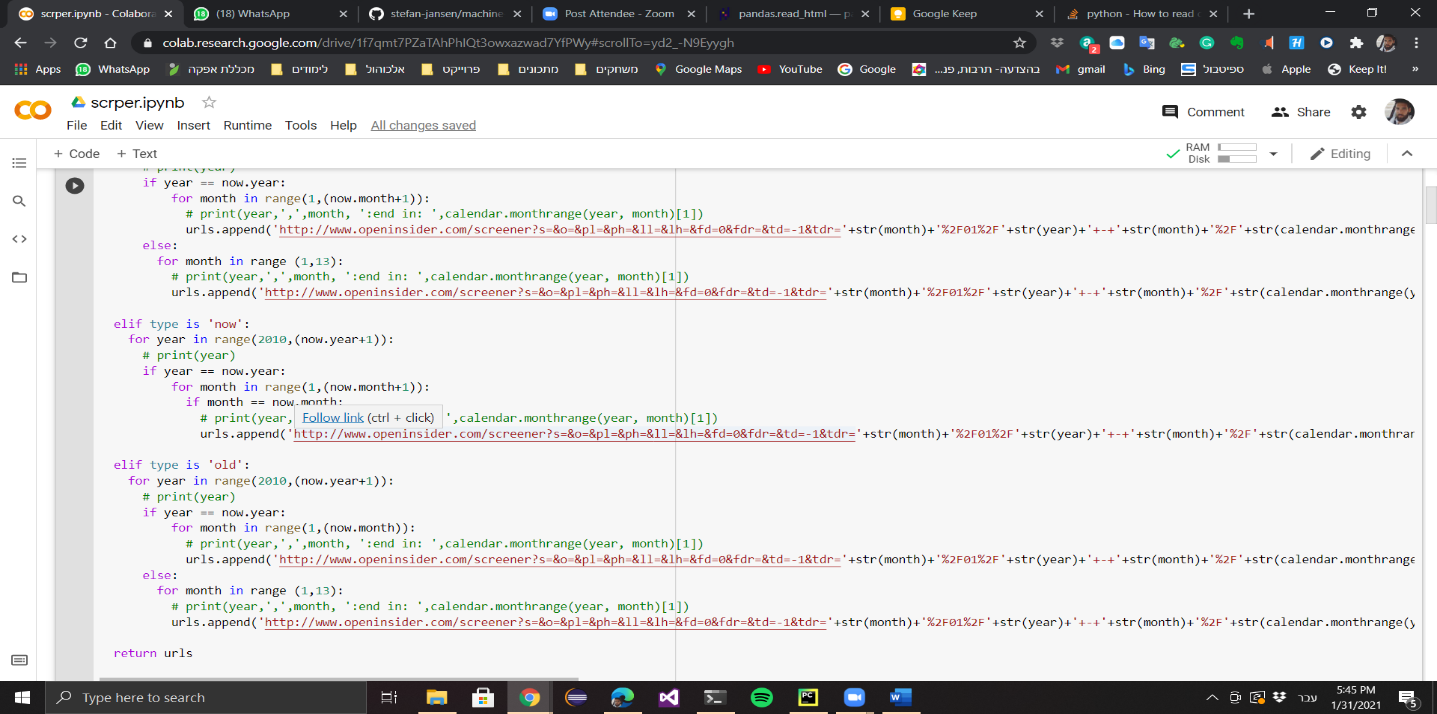
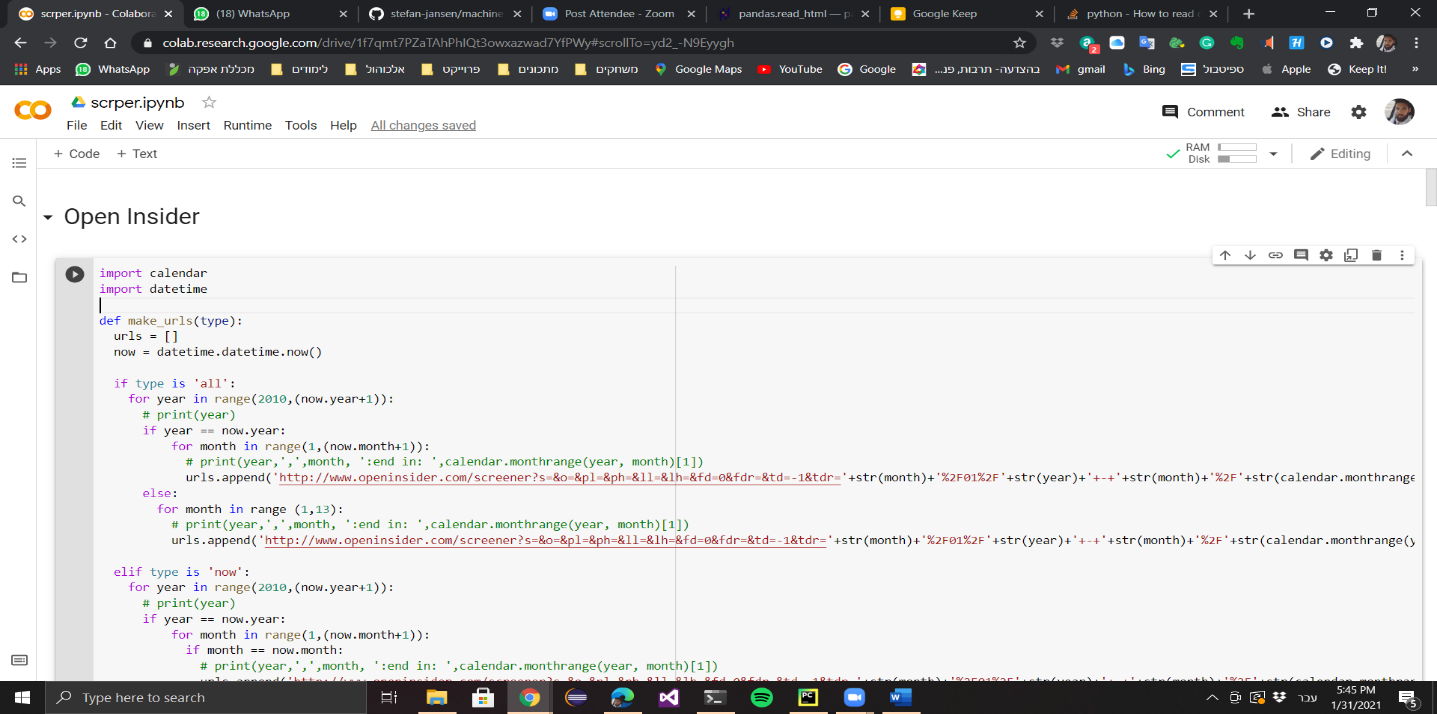
calculate\_time- מקבלת את הידיעה (tweet) ומחזירה את התאריך שבה היא פורסמה  
get\_text – מחזירה את ערך הטקסט של הידיעה 

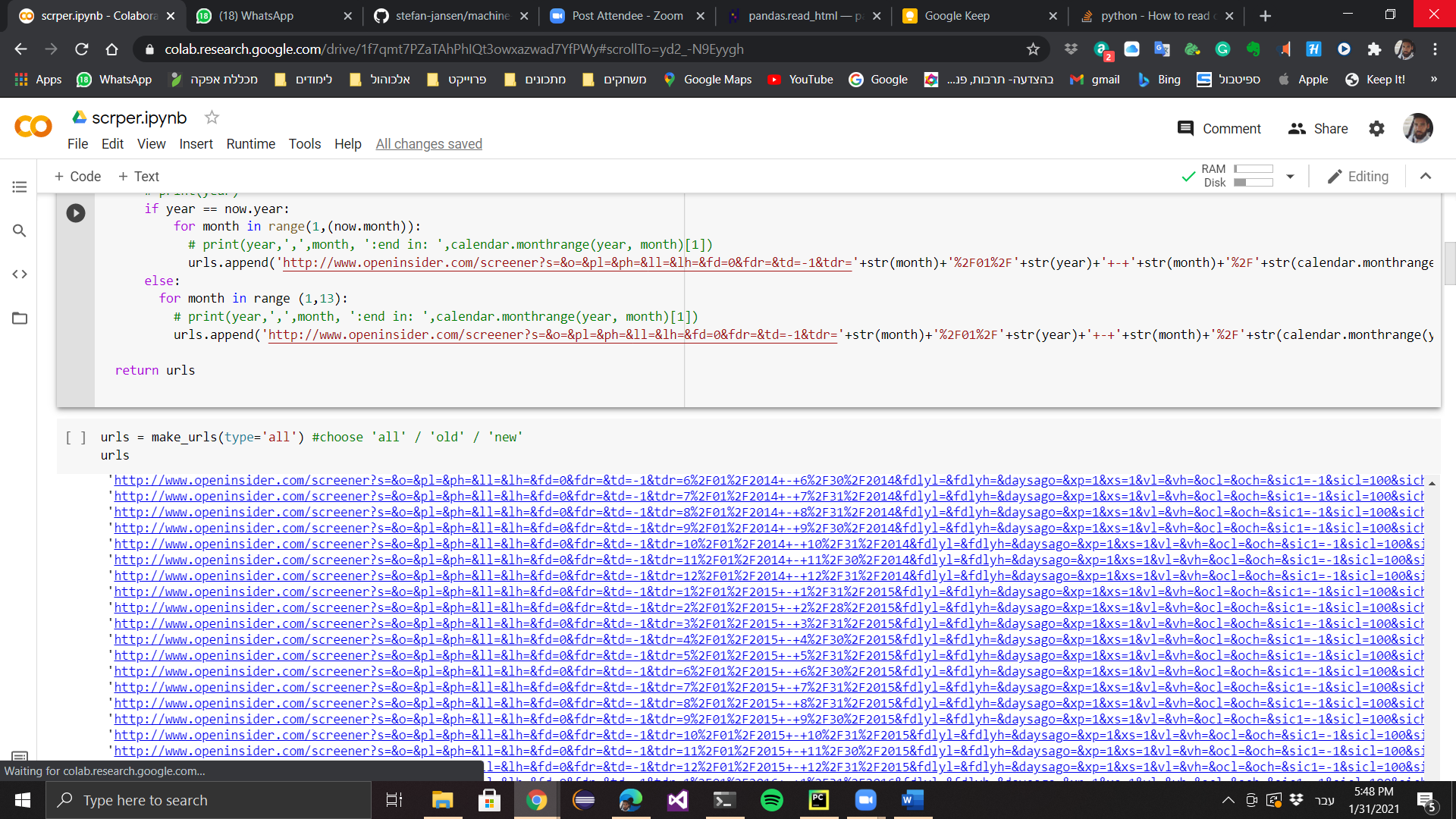
  
get\_tweets- **מעלה את הדרייב, טוענת את העמוד לפי כמות הטווטים שאנחנו רוצים לכרות, ולאחר מכן קוראת את הקוד ומושכת את המידע על כל ציוץ לפי סדר**  


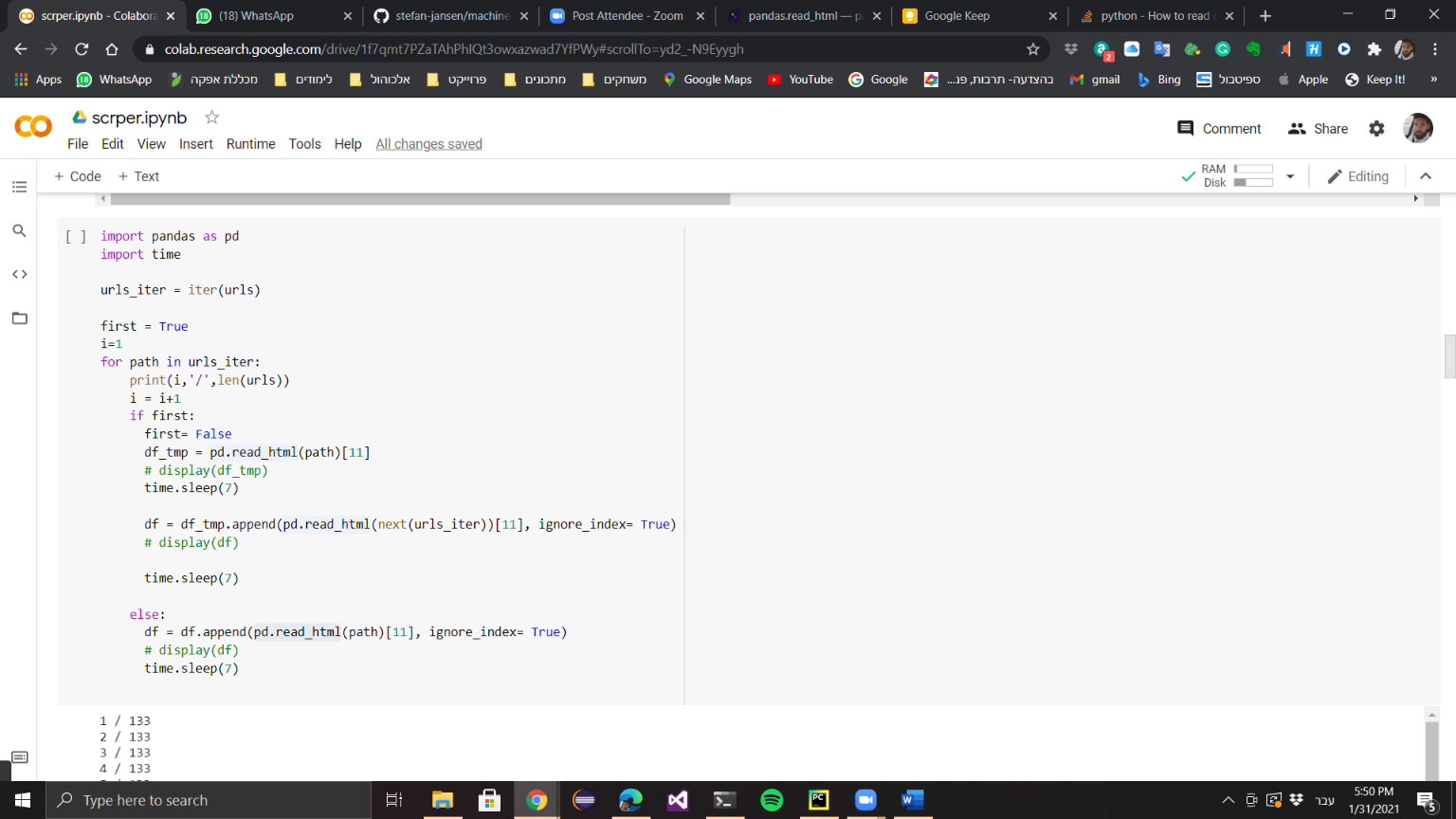
harvest\_tweets- **מגדיר את המניה והכתובת לכריית המידע, וכמות הידיעות שרוצים לקבל, שמירת המידע וייצוא לקובץ לטובת העברתו לחלק עיבוד המידע.**

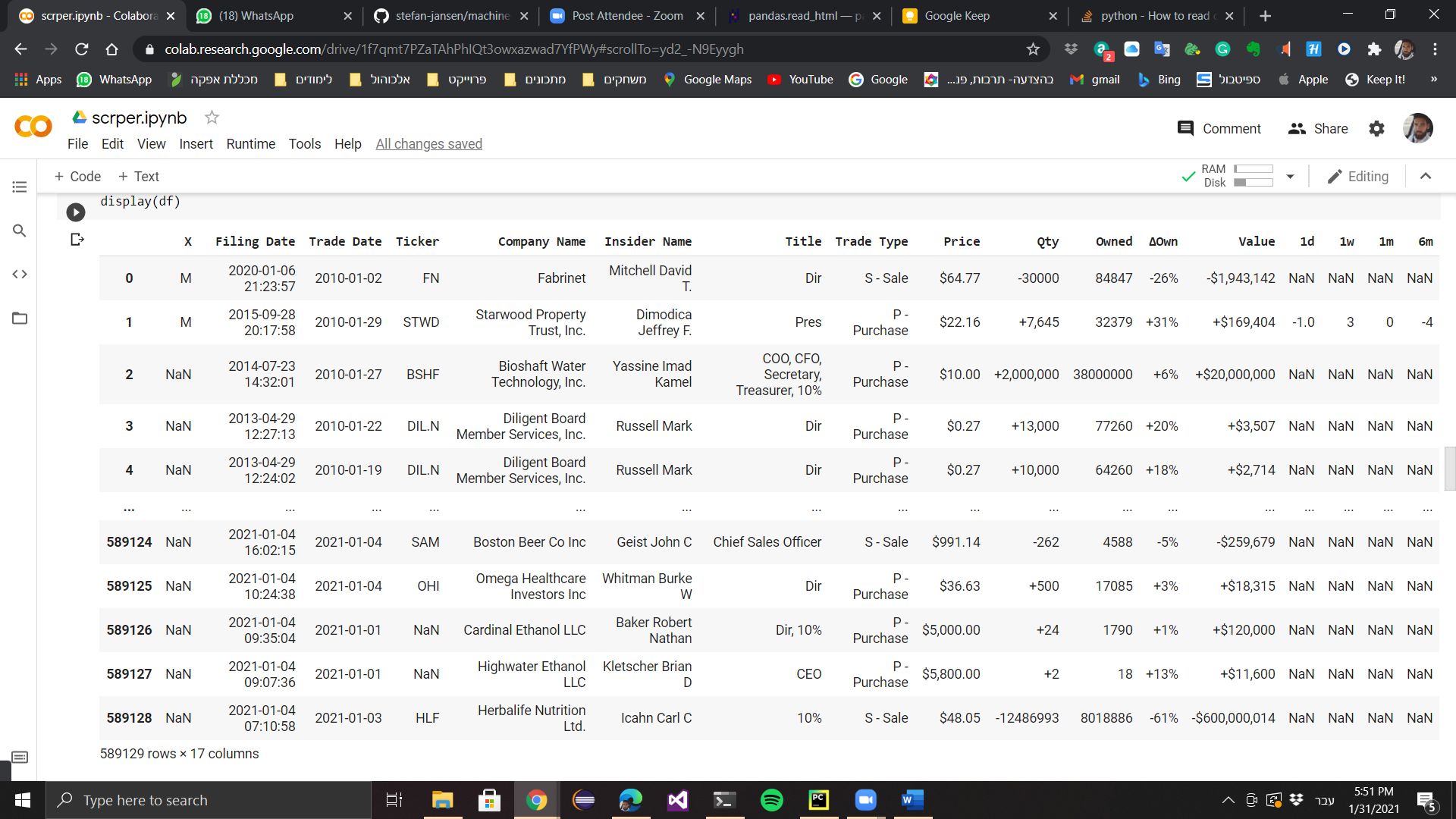
**Main-** הפעלת הכריית tweets, הגדרת כמות מבוקשת

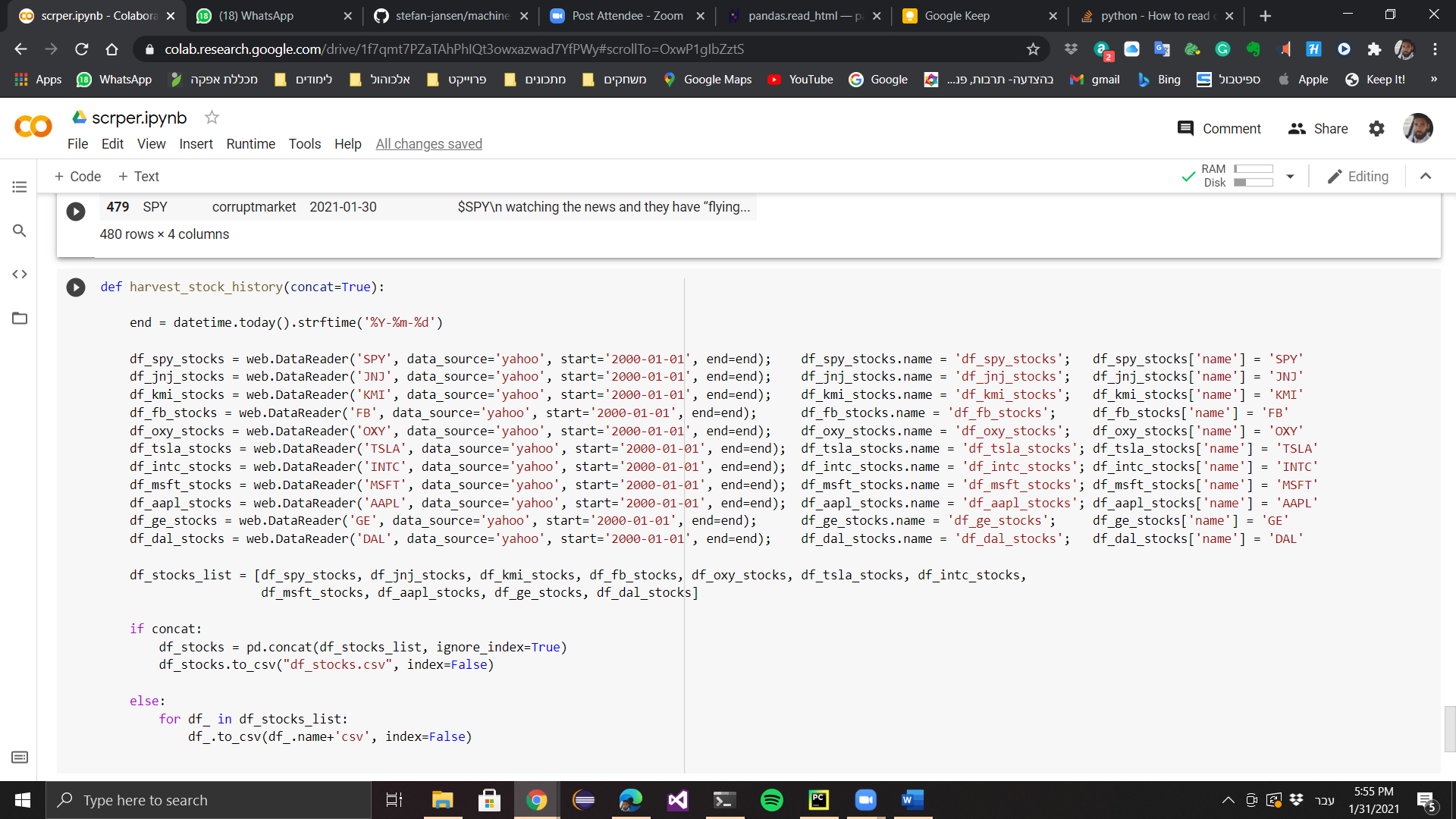
תוצר ציוצים על מניית SPY

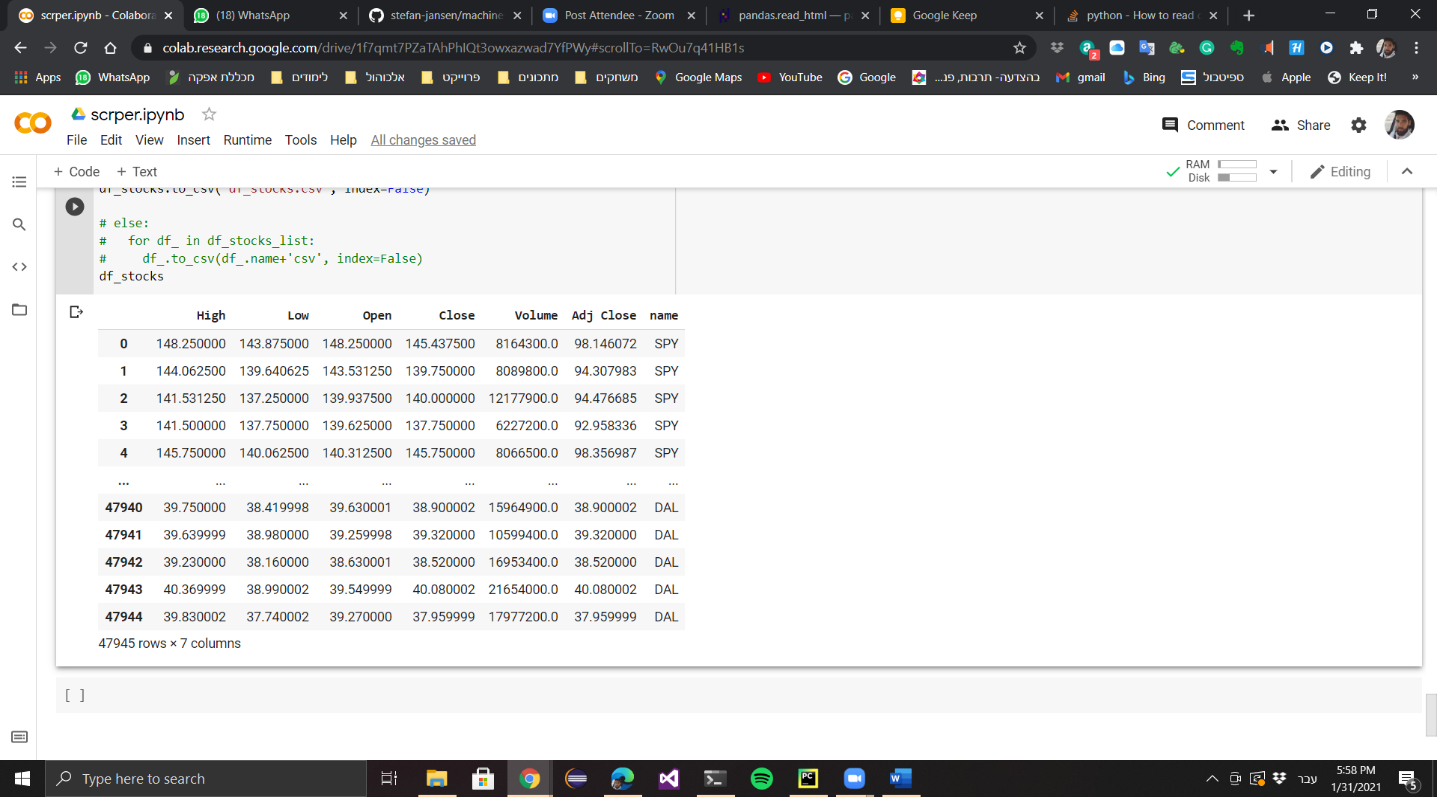
**Make\_url-** מייצרת את הכתובות לטובת תצוגת הטבלאות לפי תאריכים שנבחר (כל ההסיטוריה עד לפני חודש/ החודש הנוכחי/ הכל וכמובן שיהיה ניתן לחליט על כל תאריך שנבחר 

**דוגמא לפלט**

משיכת הטבלאות לdataframe

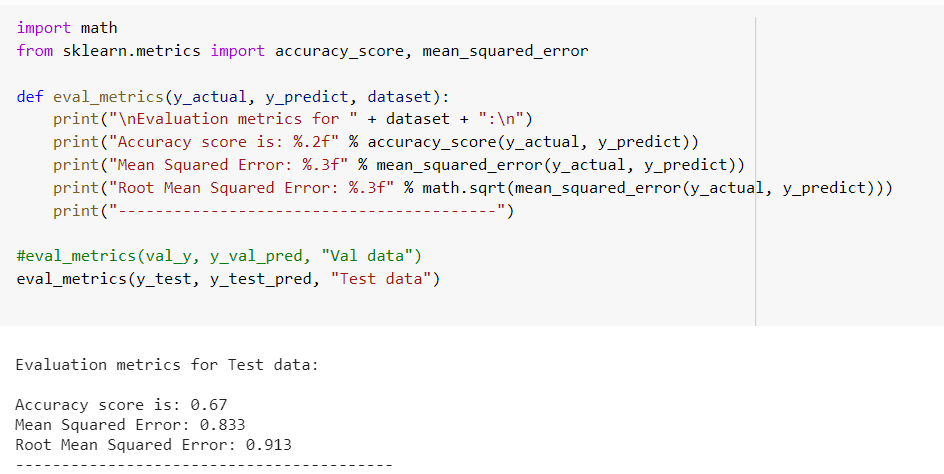
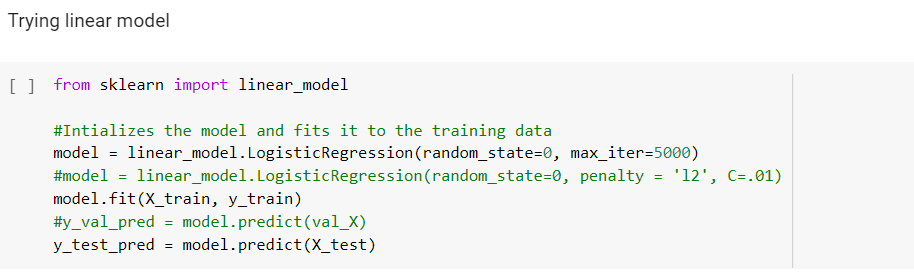
דוגמא לתוצר: 

**Harvest\_stock\_history-** מושך את מאפייני המניה הנבחרת לפי תאריכים 

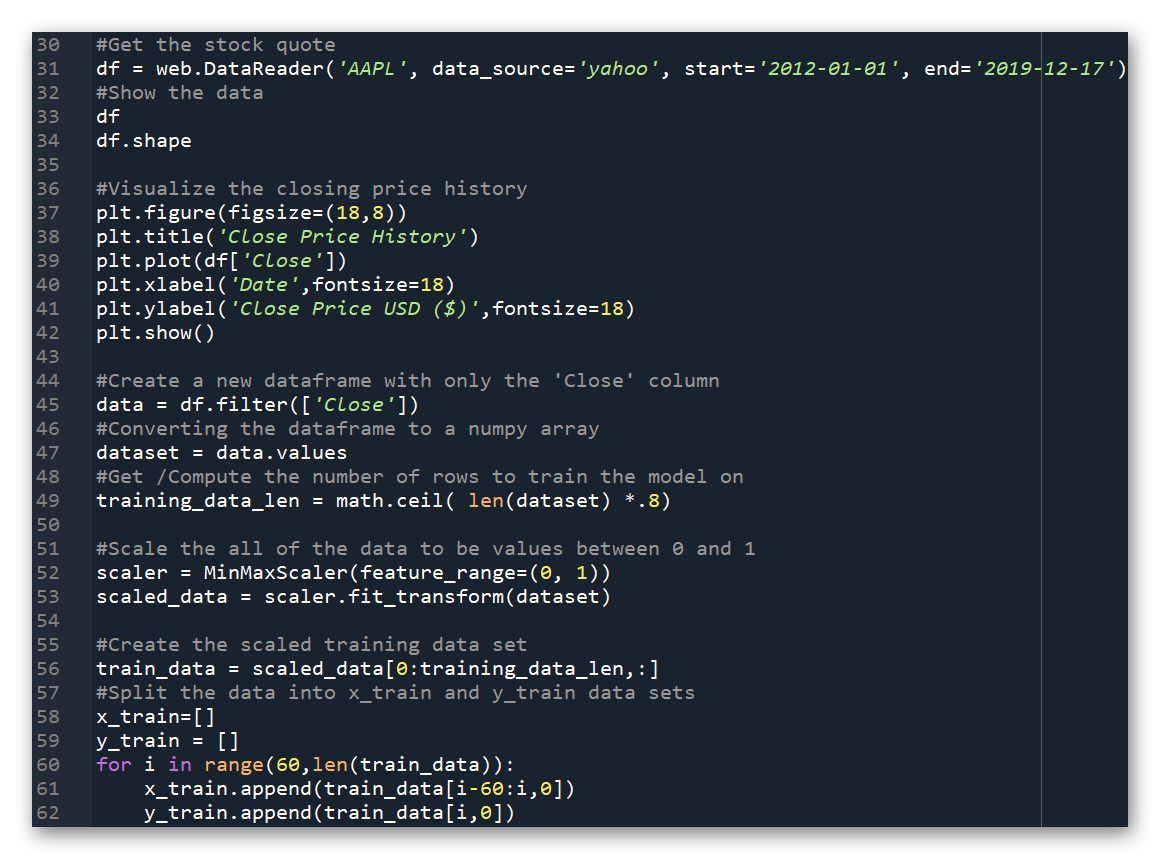
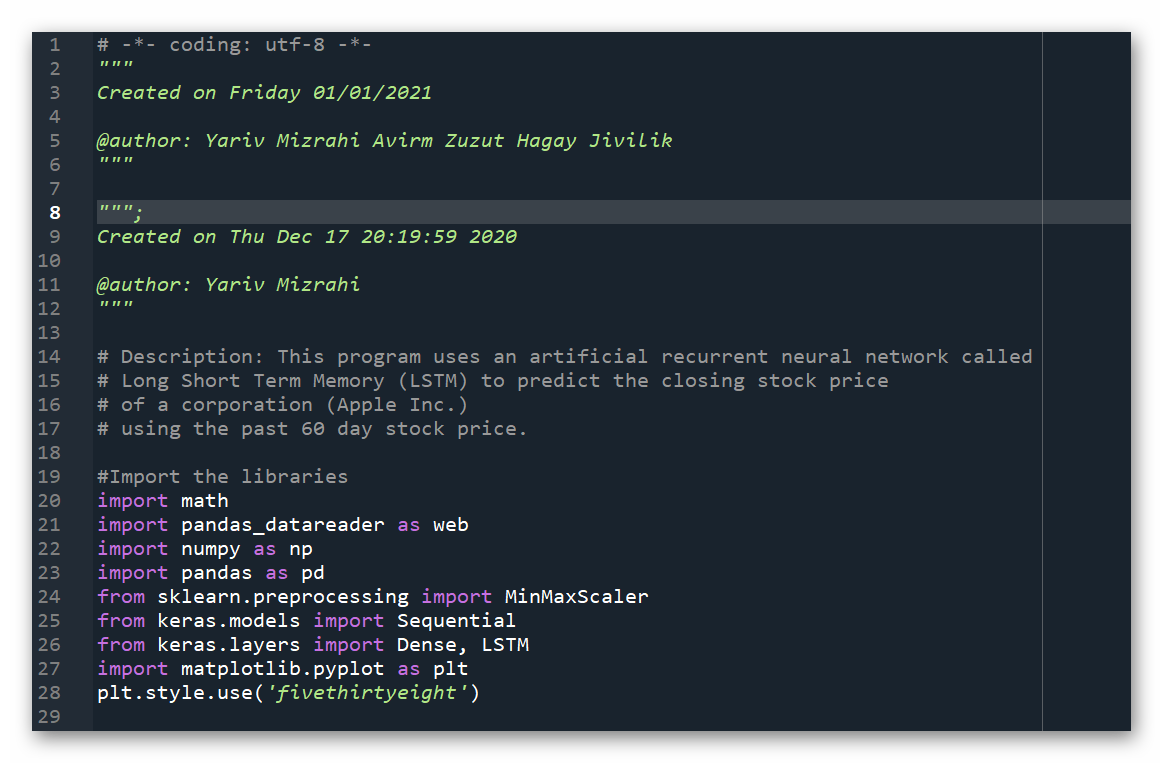
דוגמא לתוצר: 

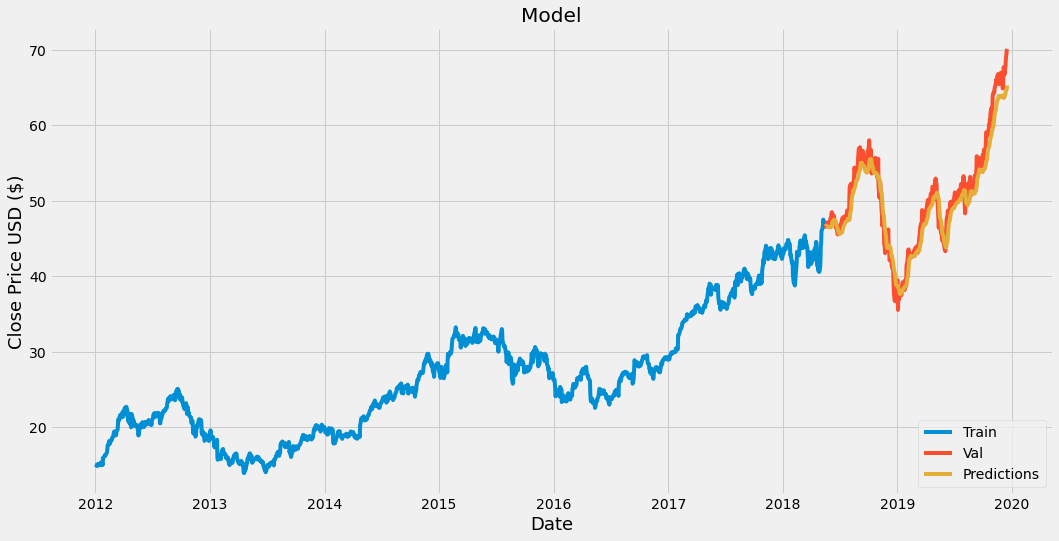
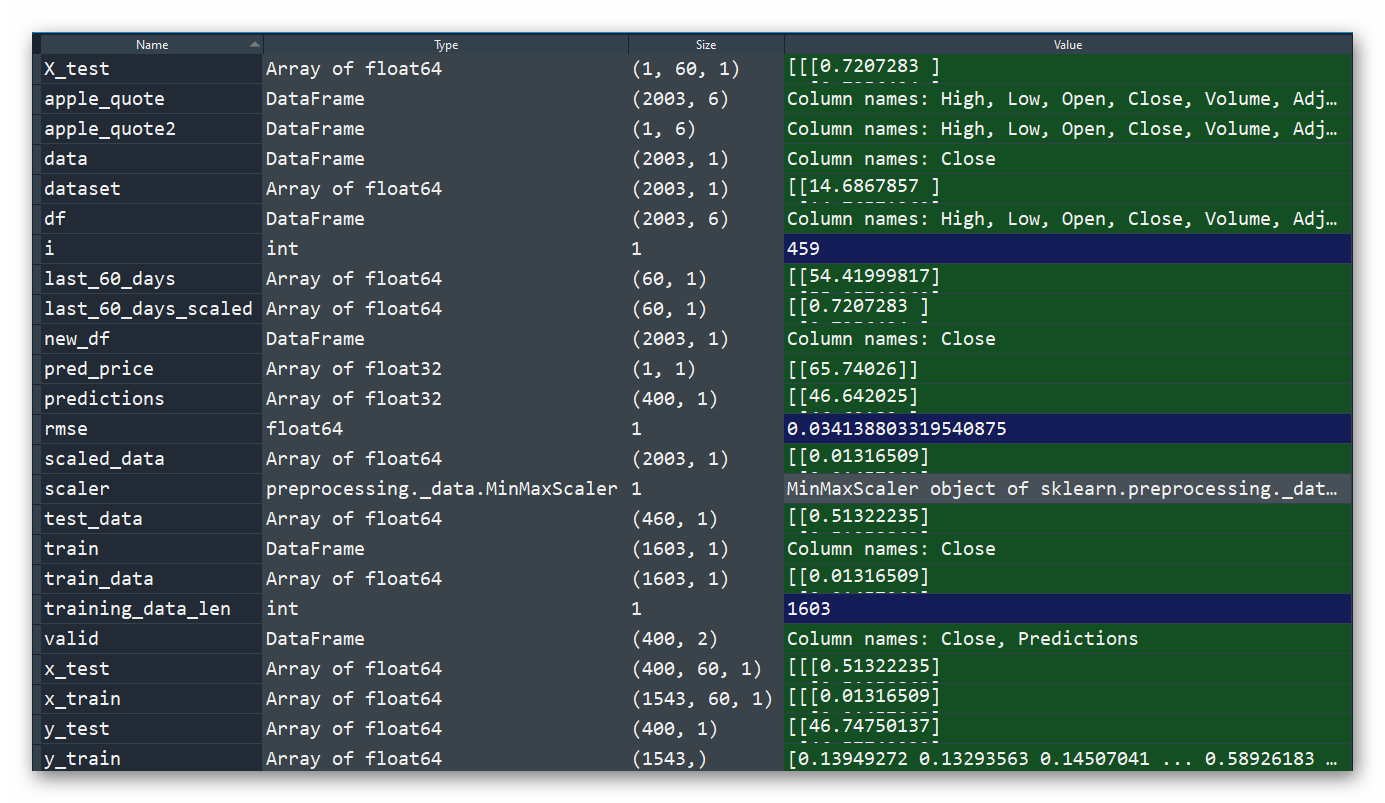
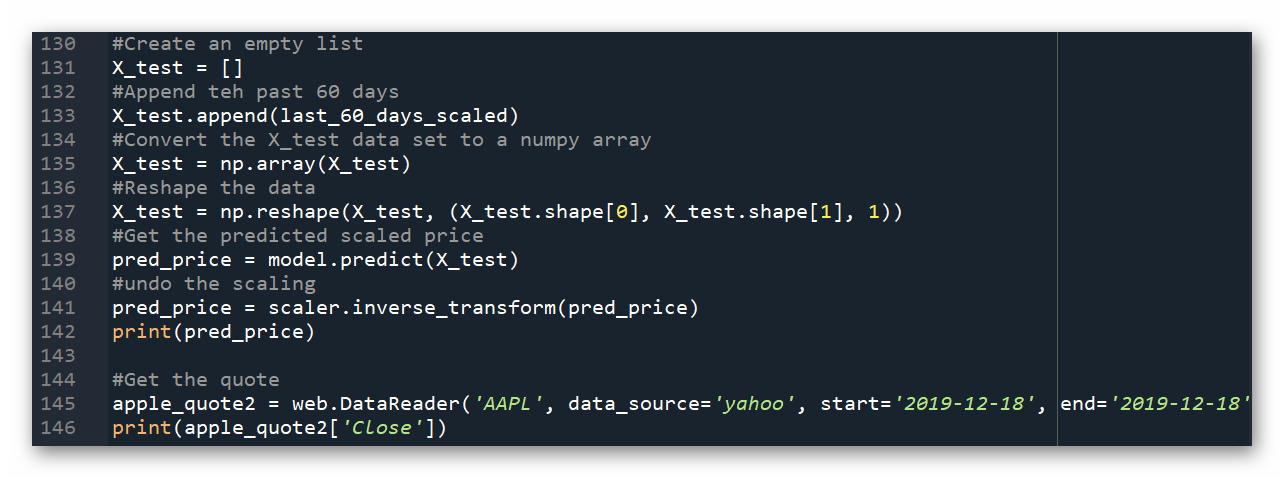
קוד עיבוד:



U,{c078479c-98f9-4bab-a0cb-40e8252d11f5}{2},10,6.66666651U,{c078479c-98f9-4bab-a0cb-40e8252d11f5}{5},10,6.66666651U,{c078479c-98f9-4bab-a0cb-40e8252d11f5}{8},10,6.66666651U,{c078479c-98f9-4bab-a0cb-40e8252d11f5}{11},10,6.66666651

**קוד מידול:**

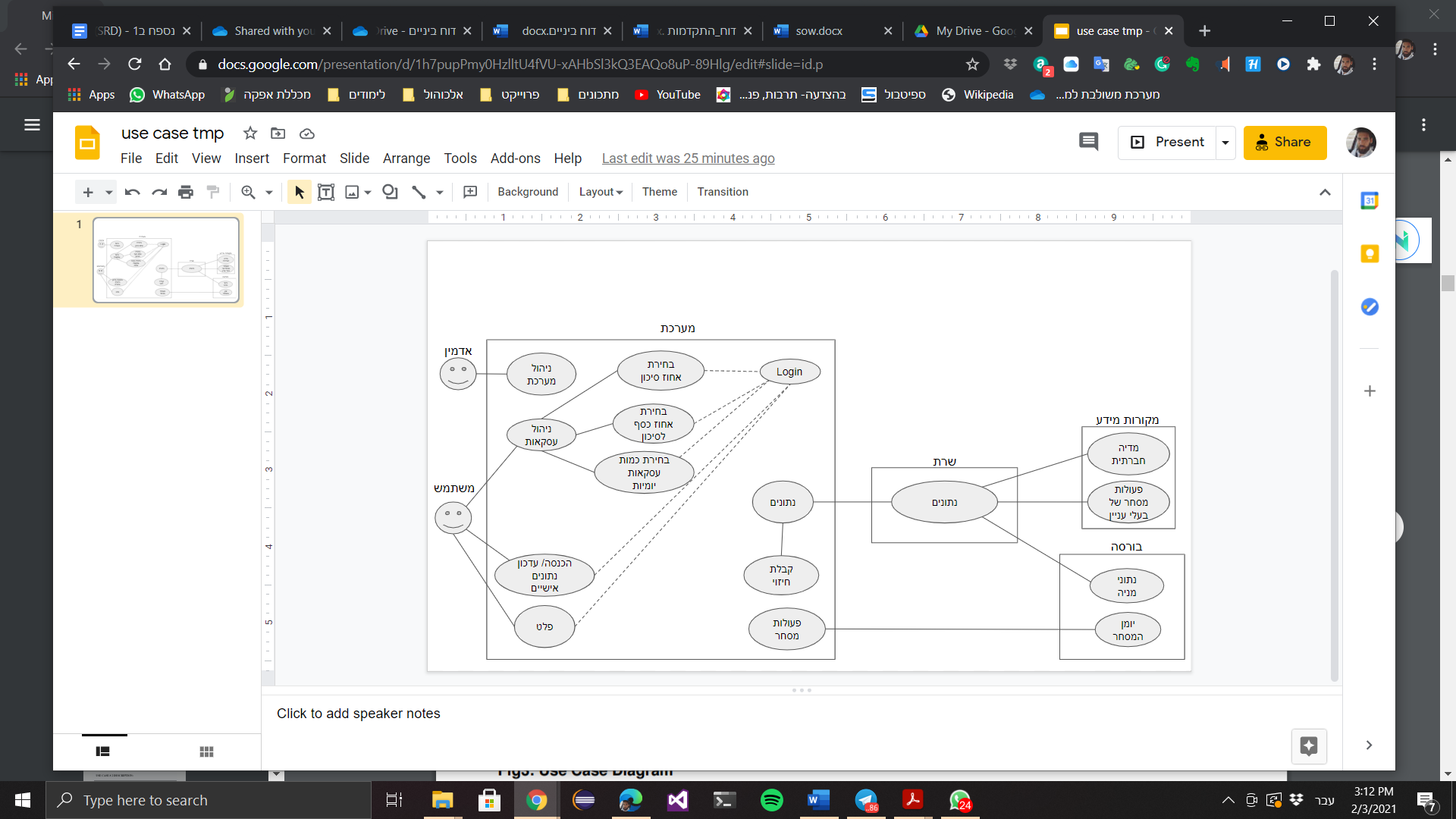




**נספח ב1 - (Software Requirements Document (SRD, מסמך דרישות למודולי תכנה**

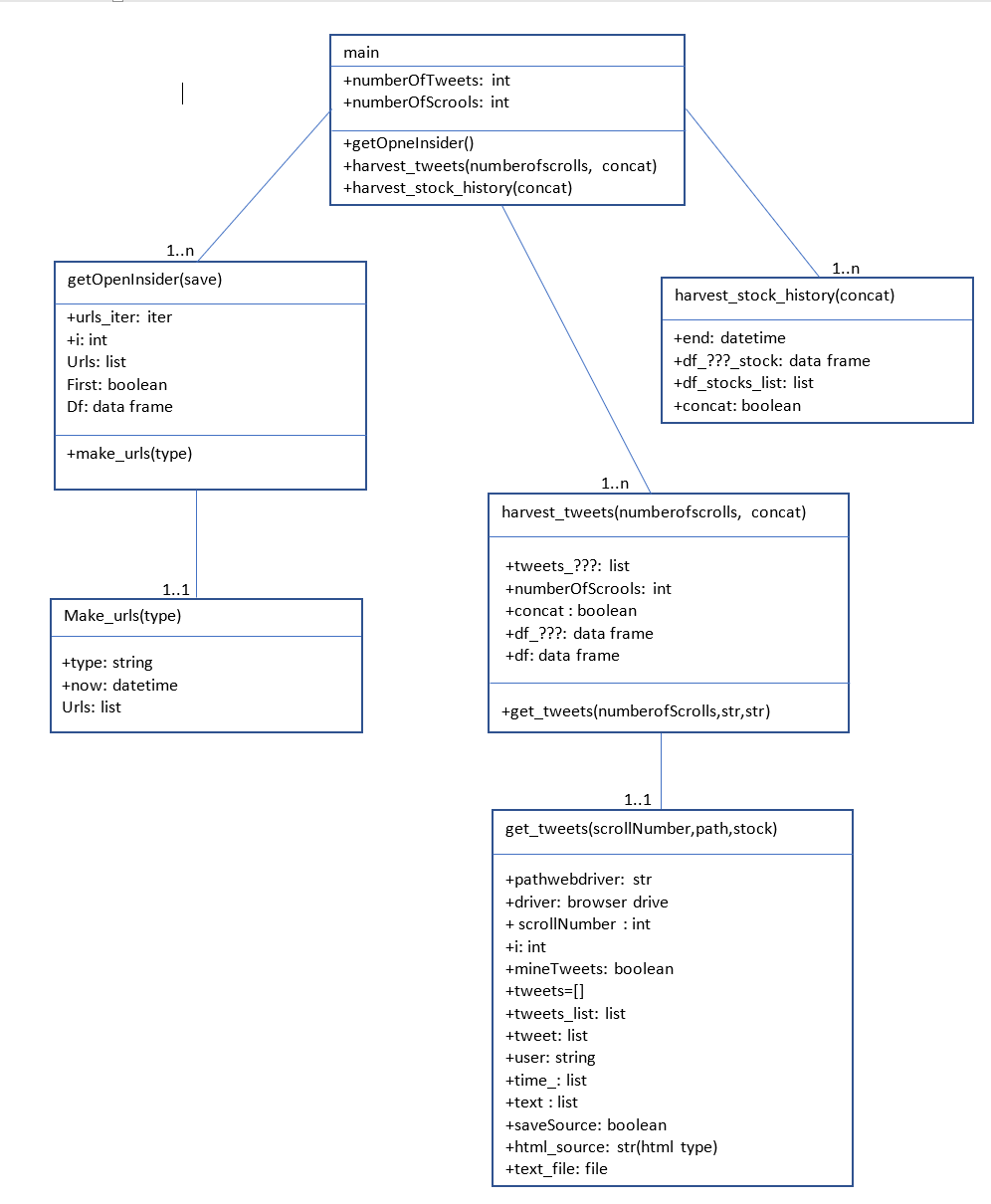
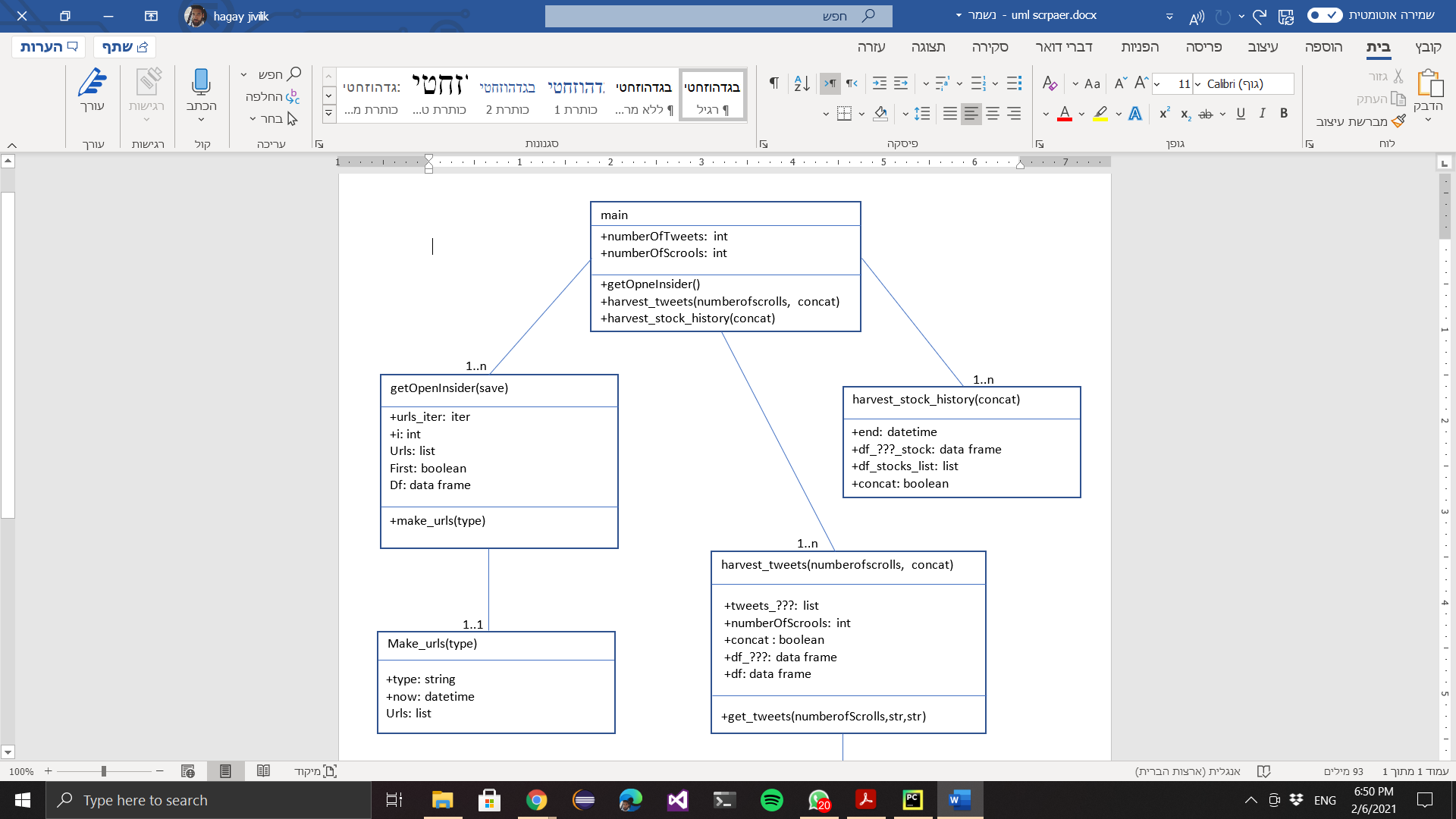
**הקדמה:**  
מטרת הפרויקט היא ליצור מערכת אוטומטית המבוססת על למידת מכונה, אשר תבצע שימוש וסחר בניירות ערך בשוק ההון. זאת תוך ניתוח מודלים פיננסיים, ניהול סיכונים וניתוח הודעות מתפרצות רלוונטיות ממקורות מידע ברשת ומדיה חברתית, כאשר שאיפת המערכת לייצר תשואה חיובית בשוק ההון, תוך מתן דגש על אוטומציה וחיסכון בזמן.

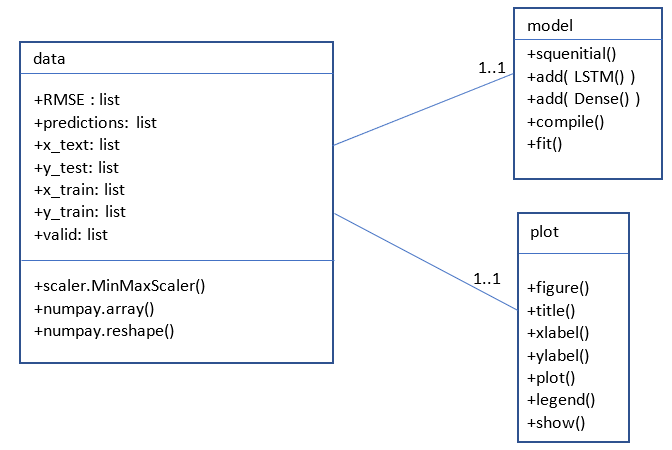
תיאור מפרט של מודלי התכנה והנתונים:  
משתמש- יכול לגשת למערכת בהינתן הזדהות בלבד, יכול לעדכן את הפרטים שלו ולקבל את הפלט מהמערכת, יכול לקבוע את ניהול העסקאות (אחוזי סיכון, אחוזי כסף ובחירת עסקאות)  
אדמין- בעל גישה לכל המערכת, תצוגת המשתמש , שינוי אלגוריתם הכריית מידע ומקורות המידע, אלגוריתם ניתוח הטקסט, אלגוריתם החיזוי, בחירת המודל הלמידה, ורשימת המניות.  
מערכת- מקבלת מידע על ידי כרייה ממקורות המידע, מנחת טקסטים ומבצעת חיזיוי למחיר המניות הנבחרות. תדרג את המניות הנבחרות, תנהל את המסחר ותנהל את סיכוני התיק.  
נתונים- משיכת נתונים ממחירי מניות בעבר, ציוצי מידע על מניות, פעילות בעלי עניין עם מניות החברה.



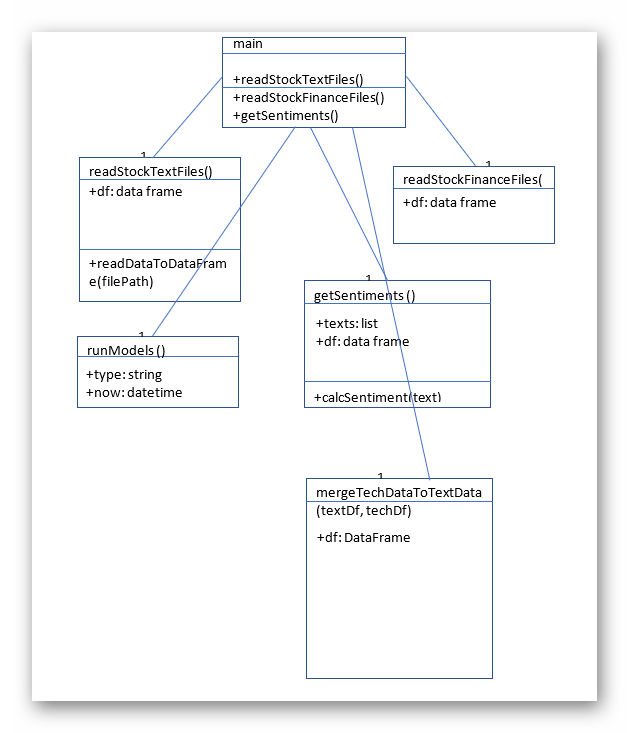
**דיאגרמת מחלקה:**

**זחלן:**

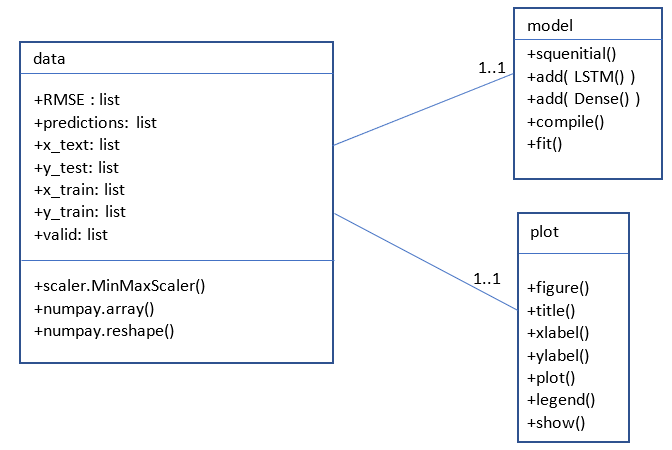
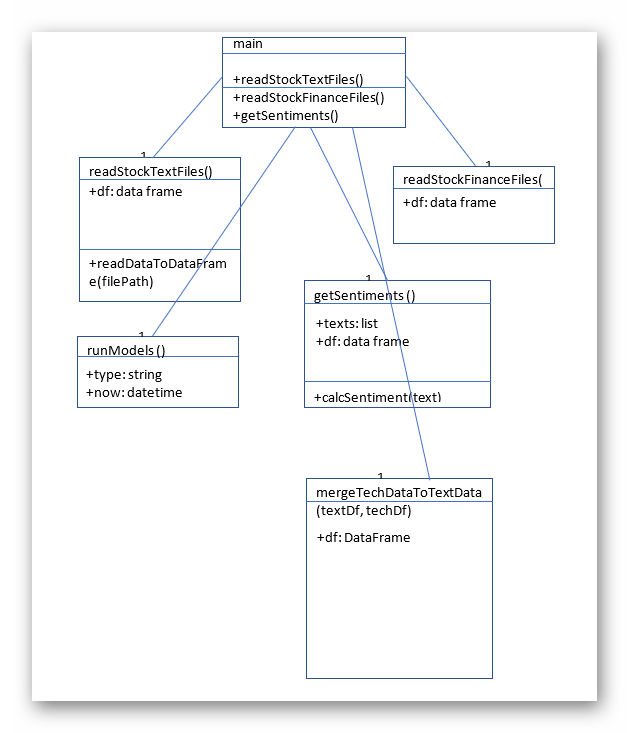
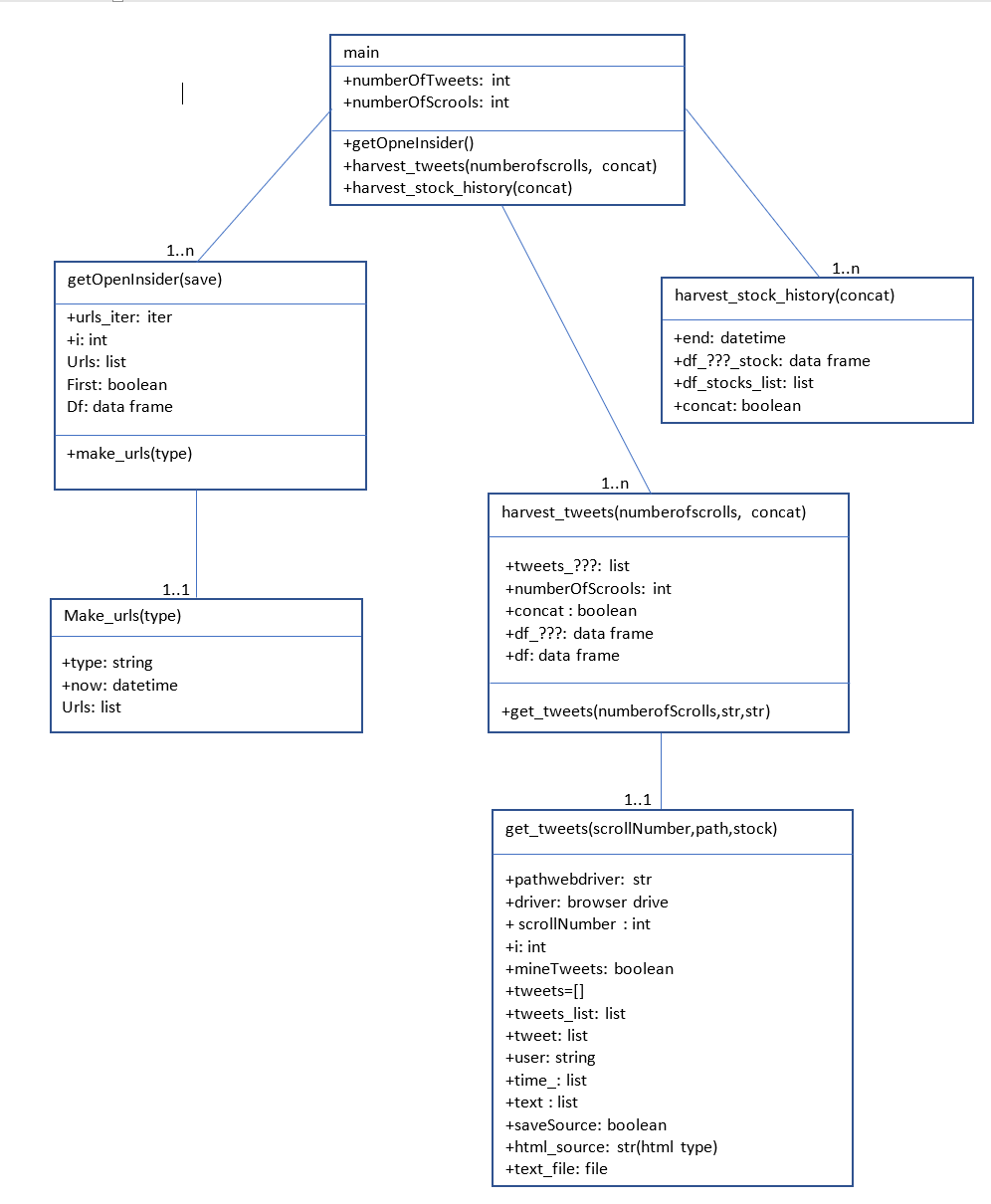
* 1. 



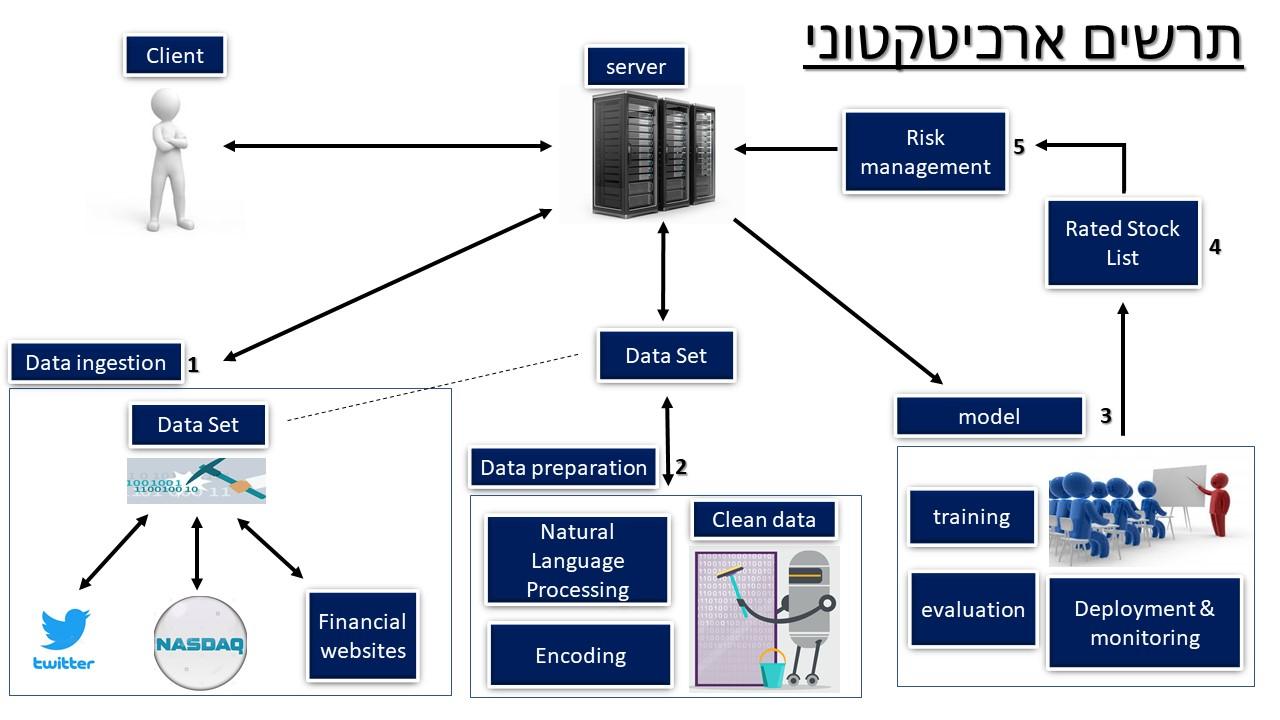
**נספח ב2 - (Software Design Description (SDD, מסמך עיצוב ותכן מודולי תכנה**

1. הקדמה: פירוט תכן על של מודולי התכנה   
   מופיע בסעיף תכן במסמך.
2. תיאור מפרט של מודלי התכנה והנתונים, בדגש ל:
   1. כריית מידע: אחראי להבאת הנתונים לצורך עיבוד וחיזוי, לצרוך כך יש שימוש במספר מודלים:
      1. **GetOpenInsider**- מתודה שנבנתה להבאת רשימת העסקאות של בעלי העיניין בתוקפה שמוגדרת מראש. הנתונים מגיעים בטבלה כאשר כל שורה מייצגת עסקה, והעמודות מפרטות את פרטי העסקה (תאריך העסקה, מי מבצע, כמה קנו/מכרו, איזה מניה, תפקיד המבצע...) יש שימוש ב:
         1. Make\_urls- מתודה שנבנתה המחזירה רשימה של urls, כל פריט ברשימה מסמל כתובת באתר המאפשר גישה לטבלה בתאריכים מסויימים עם פירוט העסקאות.
         2. Read\_html- מתודה מחזירה את הטבלה המבוקשת מהכתובת לבסיס נתונים של pandas (data frame) לשימוש נוח והעברה לעיבוד.
         3. To\_csv- המרה של טבלת הנתונים לקובץ csv כדי לעביר את הנתונים לעיבוד הנתונים.
      2. **Harvest\_stock\_history**- מתודה שנבנתה כדי לקבל את הנתונים על המניות הנבחרות (high, low, open, close, volume, adj close) בתאריכים נבחרים
         1. DataReader- מתודה של pandas למשיכת מידע מאתרים מוכרים כמו yahoo finance בצורה קלה על ידי, שם מניה ותאריכים.
         2. Concat- מתודה של pandas המאחדת בין הטבלאות שונות של כל מניה לטבלה אחת שלמה)
      3. **Harvest\_tweets-** מתודה שנבנתה לטובת כריית tweets מאתר stocktweets, המתודה יודעת לפי שם המניה להיכנס לאתר ולקבל את הציוצים האחרונים לפי הגדרת המערכת:
         1. Get\_tweets- מקבל רשימה של tweets, ציוצים על מניה נבחרת, המתודה מעלה דרייב של Chrome, מדמה דפדפן ומבצעת פעולות אוטומטיות על דפדפן כדי למשוך מידע:
            1. העלאת הדרייב, webdriver.Chrome- מאפשר לתת פקודות למדמה דפדפן לפעולות אוטומטיות
            2. Driver.get(path) - מאפשר לדפדפן לטעון את העמוד הרצוי.
            3. driver.execute\_script- מאפשר לדפדפן לגלול למטה לתחתית העמוד כדי לטעון עוד ציוצים על המניה.
            4. driver.find\_elements\_by\_css\_selector- מאפשר גישה לקוד הhtml באתר ולמצוא את כל הטווטים שהעמוד טען
            5. elems.find\_elements\_by\_css\_selector- מאפשר לטעון כל טוויט בנפרד ולבצע חיפוש נוסך בתוכו כדי לקבל את הערכים הרצויים (שם משתמש, תאריך, שם מניה, טקסט)
            6. Get\_text- מתודה שנוצרה כדי להתמודד עם הבעיה של חוסר העיקביות עם הטקסט בכל tweet, פתרון המותאם אישית לאחר למידה של התנהגות שמירת המידע על מנת לחלץ את הטקסט הרצוי.
            7. Calculate\_time- מתודה שנוצרה כדי להטפל בצורה נקודתית עם זמן המידע שמפורסם באתר, מכיוון שהאתר הוא דינמי ומשתנה מדקה לדקה בתצוגת כמה ישן הציוץ ביחס לזמן הנוכחי (עכשיו/ לפני דקה-59 דקות/לפני שעה-23 שעות / בשעה ויום אחר). המתודה יודעת לקרוא את הזמן לפי פורמט האתר המשתנה ולשנות אותו לתאריך יומי בלבד לפי צרכי המערכת.
            8. Driver.page\_source- מאפשר לשמור את הקוד הhtml של דפדפן למתשנה לוקלי (לטובת אפשרות עתידית של שמירת המידע של האתר לקובץ ולעבוד עליו ללא הצורך במדמה דפדפן לאחר המשיכה שלו)
            9. Open- פתיחת הקובץ לשמירת מידע האתר html.
            10. Write- כתיבת המידע לקובץ לוקלי.
            11. Close- סגירת הקובץ לטובת שימוש עתידי במידע המאוחסן.
         2. DataFrame- שימוש במתודה של pandas המאפשר להפוך רשימות של ערכים לשורה בטבלה לצורך צפייה ועיבוד קל של כל הנתונים יחד.
   2. עיבוד מידע:  
      **readStockTextFiles** - קריאת המידע הטקסטואלי שהתקבל מתהליך כריית המידע אל DATAFRAME, סידורו וניקוי/השלמת מידע.  
      **ReadStockFinanceFiles** - קריאת המידע הטכני שהקבל מהליך כריית המידע אל DATAFRAME, סידור וניקוי המידע.  
      **GetSentiment** - הפעלת פונקצית חילוץ סנטימנט בעזרת ספריית NLTK. מתן ציון משוקלל לכל טקסט.  
      **MergeTechDataToTextData** - מיזוג המידע הטקסטואלי והטכני, כך שלכל מניה יהיה גם ניתוח טכני וגם ציון נומרי על הטקסטים הרלוונטיים למניה.  
      **RunModels** - הרצת מס' מודלים לבקשתינו, מתן דיווח איזה מהמודלים הצליח בציון גבוה יותר לחזות בצורה הטובה ביותר את סנטימנט המניה והוצאת פלט של הקובץ אל שלב המידול. 
   3. מידול:
      1. MinMaxScaler- מנרמל משתנים לתחום מוגדר (אצלנו 0 עד 1), באמצעות חישוב מתמטי ערך מקור פחות ערך מינ' חלקי גודל קבוע(ערך מקס פחות ערך מינ')
      2. RMSE- -
      3. Numpay.array – בעקרון מערך הוא אוסף של מצביעים והגדרות, הפקודה הזו מבצעת פירוט מידע וייעול מבחינת בייטים, למשל מס' שורות/ עמודות וכמה בייטים נדרשים כדי לדלג בין איברים. הפקודה הזו חשובה כדי לייעל את הקוד ולהמשך הפקודות של חישובים נומריים.
      4. Numpay.reshape - משנה את המערך מבלי לשנות את המידע
      5. model = Sequential()- מאגד שכבות באופן לינארי, משמש אותנו לשכבות LSTM ורשתות נוירונים.
      6. LSTM - Long Short-Term Memory layer – מערכת לזמן קצר/ארוך
      7. Dense(Layer)- שכבה רגילה ברשתות נוירונים.
      8. Compile- מתאימה את הסט לאימון
      9. Fit – מאמנת את הסט בצורה איטרטיבית, לפי epochs - תקופה.
      10. Predict- מייצר תוצאת פלט עבור הכניסה.

דיאגרמות:



**תיאור הארכיטקטורה:**

השלב הראשוני הוא משיכת המידע:  
אתר stocktweet המאגד את כל הtweets לפי שם מניה  
אתר openInsider המאגד את כל הפעולות במניות של בעלי העניין באותה חברה (מנכל, סמנכ"ל, בעל מניות, דירקטוריון....) ומציין האם הפעולה הייתה מתוכננת מראש או עכשיווית, כמות המכירה/קניה, שם מבצע הפעולה, תפקידו בחברה  
אתר פיננסים yahoo- דרכו (דרך אחת מרבים) ניתן למשוך את היסטורית המחירים של מניות מסויימות משוק ההון האמריקאי  
לאחר משיכת המידע המערכת תבצע עיבוד למידע, ניקוי שלו, עיבוד הטקסטים המרת כלל המאפיינים השונים בקלט לערכים. לאחר המרת כל הנתונים לטבלת מאפיינים ותצפיות )הכוללת ערך נומרי לאהדה לטקסט שנאמר על מניה) העברתם למודל החיזוי היודע לנבות מחיר מניה המערכת תדע לדרג את המניות לפי חיזוי המודל ולאור הגדרות המשתמש המערכת תבחר על איזה מניות לבצע פעולות מסחר כאשר בסוף התהליך נוכל להשוות את התשואה של התיק אל מול פעולות המסחר באלגוריתם ללא ניתוח מדיה חברתית.

1. **Use case**
2. תרחיש 1
   1. שם התרחיש: הרשמה למע'
   2. שחקנים: לקוח(לקוחות פרטיים המעוניינים להשקיע), מערכת
   3. מטרה: לאפשר ללקוח להירשם למערכת
   4. מסלול עיקרי-
      * 1. לקוח נכנס למערכת ומבקש להרשם.
        2. לקוח מזין את פרטיו האישיים.
        3. לקוח בוחר הגדרות המותאמות אישית לצרכיו(נתונים אישיים, כמות עסקאות יומית, אחוז סיכון, אחוז כסף).
        4. המערכת שומרת את פרטי התחברות המשתמש והגדרותיו בבסיס הנתונים.
   5. מסלול חלופי 1-
      1. משתמש מזין פרטים לא נכונים (תקינות קלט)
      2. המערכת מודיעה למשתמש ומבקשת קלט חוזר
   6. מסלול חלופי 2-
      1. משתמש לא מגדיר הגדרות אישיות
      2. המערכת משתמש בהגדרות שנקבעו מראש.
4. תרחיש 2:
   1. שם התרחיש: שינוי פרטי לקוח
   2. שחקנים: לקוח(לקוחות פרטיים המעוניינים להשקיע), מערכת
   3. מטרה: לאשר לקוח לשנות פרטים והגדרות
   4. מסלול עיקרי:
      1. לקוח מבצע התחברות למערכת.
      2. הלקוח דורש לשנות את פרטי האישיים.
      3. הלקוח בוחר הגדרות המותאמות אישית לצרכיו(נתונים אישיים, כמות עסקאות יומית, אחוז סיכון, אחוז כסף).
      4. המערכת שומרת את פרטיו האישיים של הלקוח בבסיס הנתונים.
   5. מסלול חלופי 1:
      1. הלקוח משנה פרטים אישיים בקלט לא תקין
      2. המערכת מדווחת על קלט לא תקין ומבקשת להכניס את הנתונים בשנית
      3. אם הוא לא מכניס המערכת לא משנה את הנתונים הקודמים.
   6. מסלול חלופי 2:
      1. הלקוח משנה הגדרות מסחר לא תקין
      2. המערכת מדווחת על קלט לא תקין ומבקשת להכניס את הנתונים בשנית
      3. אם הוא לא מכניס המערכת לא משנה את הנתונים הקודמים.
5. תרחיש 3:
   1. שם התרחיש: התחברות למערכת
   2. שחקנים: לקוח(לקוחות פרטיים המעוניינים להשקיע), מערכת
   3. מטרה: לאפשר התחברות למערכת
   4. מסלול עיקרי-
      1. לקוח מבצע התחברות למערכת.
      2. מערכת בוחנת את פרטי ההתחברות שהזין הלקוח ומחפשת התאמה.
      3. המערכת מעבירה את הלקוח אל דף הבית המותאם אישית לצרכיו.
   5. מסלול חלופי-
      1. אין התאמה במערכת על נתוני ההתחברות
      2. המערכת מבקשת להכניס את הנתונים שנית כי אין התאמה
6. תרחיש 4:
   1. שם התרחיש: מסחר בבורסה
   2. שחקנים: לקוח(לקוחות פרטיים המעוניינים להשקיע), בורסה
   3. מטרה: לאפשר למתשמש לקבל תמונת מצב
   4. מסלול עיקרי:
      1. לקוח מבצע התחברות למערכת.
      2. המערכת סורקת את נתוני המשתמש ומחפשת הזדמנויות עסקיות עבורו בעזרת נתונים מהבורסה וממדיות חברתיות.
      3. המערכת מבצעת פקודות קנייה/מכירה באופן אוטומטי בהתאם להגדרות המשתמש.
      4. המערכת מעדכנת במסך הבית את המשתמש על פעולות שהיא מבצעת, כמות הרווח/הפסד בכל רגע נתון.
   5. מסלול חלופי:
      1. הלקוח לא מצליח להתחבר
      2. המערכת מודיעה על אי התאמה ומבקשת להתחבר בשנית
   6. מסלול חלופי 2:
      1. המערכת סורקת נתונים (כריית מידע) ולא מקבלת את המידע המבוקש
      2. המערכת מתריאה איזה מקור מידע לא תקין ולא ממשיכה במסחר עד לפתרון התקלה.
   7. מסלול חלופי 3:
      1. המערכת לא הצליחה לבצע את הפקודות קניה/ מכירה.
      2. המערכת מעדכנת את הפעולה שהיא לא הצליחה לבצע.
7. **פירוט דרישות**
8. כריית מידע:
   1. זמינות אתרים (ללא שינוי המידע- האופן בו התוכן נשמר ומוצג באתר)
   2. הצלחה בכריית המידע ומשיכת המידע
   3. אי חסימת הזחלן (גילוי כרובוט לא חוקי)
   4. שלמות מידע והפרדה ברורה של המאפיינים(שדות) שנכרו (טקסט, מחיר, משתמש, תאריך...)
9. עיבוד מידע:
   1. קבלת המידע לפי מה שנקבע
   2. הצלחת חיזוי של אפיון טקסט
   3. הצלחת המרת כלל המאפיינים לערכים נומריים
10. יצירת מודל:
    1. קבלת כלל המאפיינים לפי דרישה
    2. יצירת מודל עם מאפיינים ממדיה חברתית
11. מערכת:
    1. יכולת שמירת נתוני משתמש (הרשמה)
    2. יכולת בדיקה והשוואה של נתונים (התחברות וזיהוי)
    3. הגדרת ושינוי מאפייני מסחר
    4. הצגת הנתונים.
    5. הצלחה בפעולות מסחר.
12. פירוט דרישות בתחומי אבטחת מידע , בכל רמה של פונקציונאליות ותכן הנדסי
    * 1. כניסה למערכת (הצגה ושינוי של הנתונים) רק לאחר זיהוי אל מול המערכת.
      2. חיסיון פרטים אישיים.

**נספח ב3 - (Software Test Documentation (STD, מסמך תכנון בדיקות**

|  |  |  |  |
| --- | --- | --- | --- |
| מודל | תסריט בדיקה | מה ייחשב כתקין | מה ייחשב כלא תקין ודורש תיקון |
| כלל המודלים | שינוי מחלקות ושירותים מהמחלקות שיש בהן שימוש | אזהרה בעת הרצת / בניית הקוד  או שגיאה רלוונטית | קריסת המודלים עם שגיאה. |
| כריית מידע | שינוי במרקם האתר (פרסומות, תצוגת המידע, מיקום המידע, צורת אחסון, הוספת הגנות) | כריסת הזחלן והודעת שגיאה | ללא הודעת שגיאה והמשך עבודה שהכל תקין |
| כריית מידע, getOpenInsider | הרצה עם Url's תקינים | קבלת טבלה עם המידע מהאתר | יקרוס מכמה סיבות:  -שגיאה של list out of range  -טבלה עם 2 מאפיינים בלבד ו4 שורות (השורות התחתונות בדף) |
| הרצה עם רשימת url ריקה | לא יוגדר dataframe ותצא הודעת שגיאה (לא מעודכן בקוד כיום) | אין הודעת שגיאה והזחלן יקרוס ללא טבלה |
| הרצה עם רשימת url תקינה עם לפחות כתובת אחת לא תקינה | יוגדר dataframe כאשר רובו כתובות תקינות ימשכו מידע וכתובות לא תקינות לא ימשכו (עוד לא מעודכן בקוד) | לא תהיה הודעת שגיאה ותימשך טבלה כפולה (המצב כעת) |
| כריית מידע, harvest\_stock\_history | הרצה עם מידע תקין | קבלת טבלאות עם המאפיינים לפי כמות הימים שהוגדרה (תאריכים) | קבלת שגיאה של:  -זיהוי שם מניה לא תקין  -קלט לא תקין בתאריך  -תאריכים לא מסודרים כרונולוגית |
| הרצה עם תאריך לא תקין | שגיאה מתאימה לבעייה בקלט | החזרת שגיאה לא תקינה/ קריסת זחלן (קוד לא מועדכן לשגיאה עדיין) |
| הרצה עם שם מניה לא תקין | שגיאה מתאימה שאין התאמה לשם המניה | החזרת ערך של מניה אחרת/ קריסת זחלן (קוד לא מעודכן לשגיאה) |
| כריית מידע, harvst\_tweets | הרצה עם קלט חוקי | קבלת טבלה עם המידע של הtweets שנפתחו בדפדפן | קבלת שגיאה מתאימה (קיים בחלק מהקוד) |
| הרצה עם קלט לא חוקי במספר ציוצים (קטן או שווה לאפס) | לא יתבצע גלגול של הדף למטה לתחתית ויבצע כריית מידע רק של המידע שהוצג בדף | יתקבל ערך לא רצוי והדפדפן יבצע פעולת גלגול לא רצויה |
| הרצה עם קלט לא חוקי במספר ציוצים (תו במקום מספר) | שגיאת build | לא תופיע שגיאה והקוד ירוץ (כרגע מוגן על ידי הקומפיילר) |
| הופעת פרסומת באתר | סגירת חלון פרסומת והמשך כרייה (לא מטופל בקוד כעת) | אין הצלחה של סגירת חלון פרסומת ומשיכה של 2 ידיעות בלבד |
| Chrome driver לא תקין (תקינות קובץ, מיקום, טעינה) | הודעת שגיאה בהרצה או בבניה של הקוד | ללא הודעת שגיאה |
| לא נמצא הelement של הtweet בקוד הhtml | הודעת שגיאה שהרשימה ריקה ואין מידע שנמשך (כרגע לא קיימת שגיאה של רשימה ריקה) | המשך עבודה על רשימה ריקה וקריסת זחלן |
| מידול | קבלת טבלת מאפיינים ריקה/ חסרה | שגיאה על טבלה ריקה/ חסרה במאפיינים רלוונטיים | אין הודעת שגיאה והמשך הקוד למידול הנתונים הריקים |
| עיבוד מידע | ניתוח טקסט ציני | תדע לזהות שהטקסט שלילי ובמידהמוההתנהגות חוזרת על עצמה הדירוג של המתשמש תרד | לא יזהה ציניות ויתן דירוג גבוהה למשתמש |
| מערכת | הקלדת נתונים רישום לא נכונים | הודעת שגיאה | אין הודעת שגיאה והמשך קליטת מתשמש |
| התחברות עם נתונים לא נכונים | הודאת שגיאה ובקשה להכניס נתונים מתאימים לרישום | -אין הודעת שגיאה ואין כניסה  -הצלחת כניסה למערכת |
| מתשתמש לא מגידר הגדרות ניהול | הודעה מתאימה והמשך עבודה לפי בחירת ברירת מחדל | אין הגדרה ברירת מחדל/ הרצת מערכת עם הגדרות לא נכונות/ אין הודעת שגיאה |

1. תכנית עבודה ותיזמון הבדיקות: 