סמסטר א – 2021 / 2020

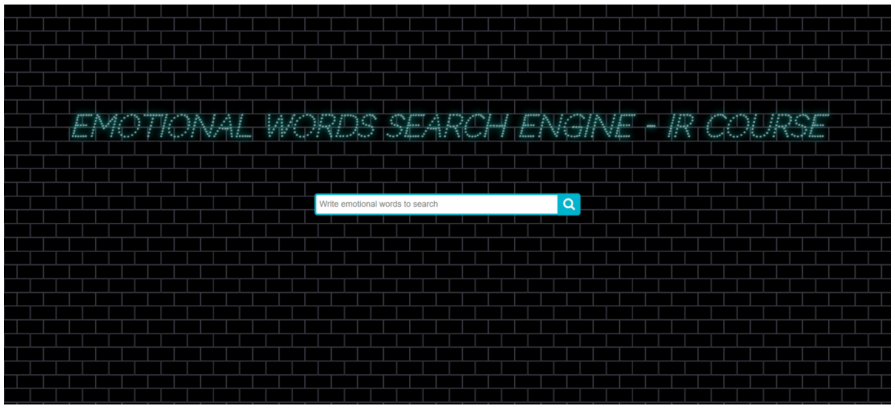
# משימה מעשית מסכמת

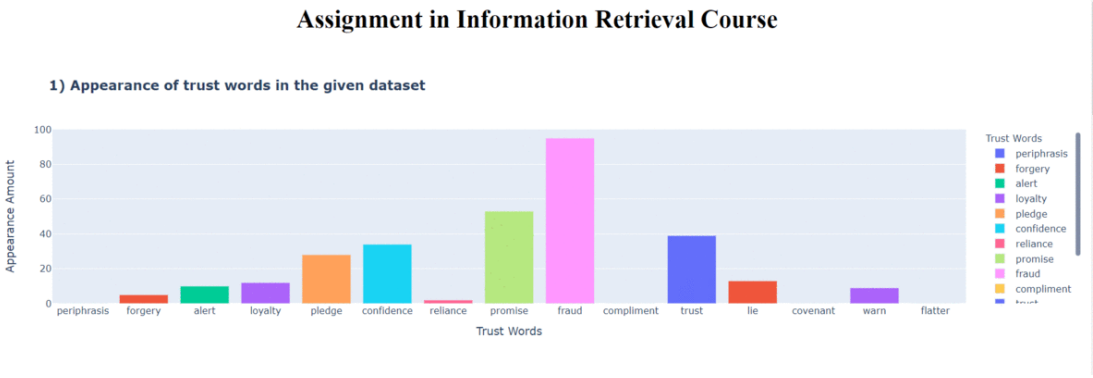
## קורס אחזור מידע


**מגישים:**

- מוחמד עביד - 313413346
- מוחמד חטיב - 208525220

**מרצה הקורס:** פרופ' צבי קופליק

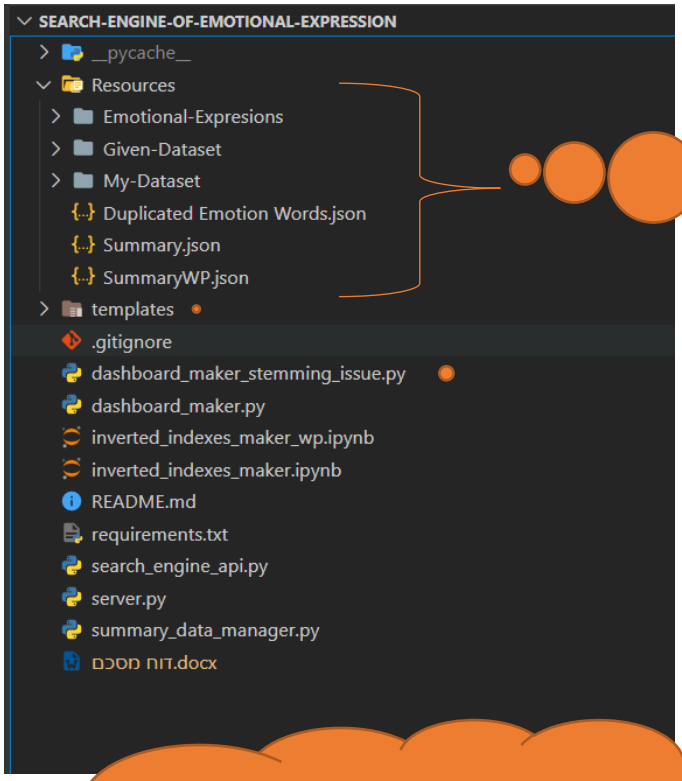
# תוכן עניינים

## הקדמה

### טכנולוגיות וסביבת פיתוח

- הפרויקט פותח בשפת Python (Python 3.8.6).
- Visual Studio Code IDE כסביבת פיתוח.
- Python Flask Server לשימוש כשרת למנוע החיפוש.
- Plotly, Dash, Pandas Modules לצורך מימוש ובנית ב Interactive Dashboards בפרויקט.

### מבנה הפרויקט

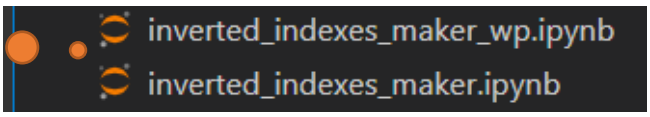


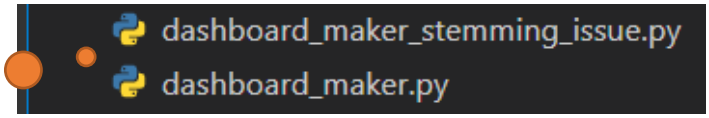

משאבים – תיקיה שמכילה את המסמכים המקוריים שקבלנו שנמצא בתת התיקיה Given Dataset, תת תיקיה שמכילה מילות האימון ומילות הרגש (Emotional-Expressions), תת תיקייה שמכילה את המסמכים אחרי עיבוד וניקוי וגם האינדקסים שבנינו (My-Dataset), בנוסף לשלושה מסמכים מסוג JSON, מסמך מסכם ללא Stemming מסמך אחר מסכם עם פעולות Stemming, ומסמך מסכם למילות משוכפלות אחרי stemming כלומר כל המילים שנופלות על אותו ביטוי אחרי stemming

תיקייה שמכילה דפי HTML שפותחו ב Python לצורך הצגת ושימוש מנוע החיפוש

**קוד Python:**

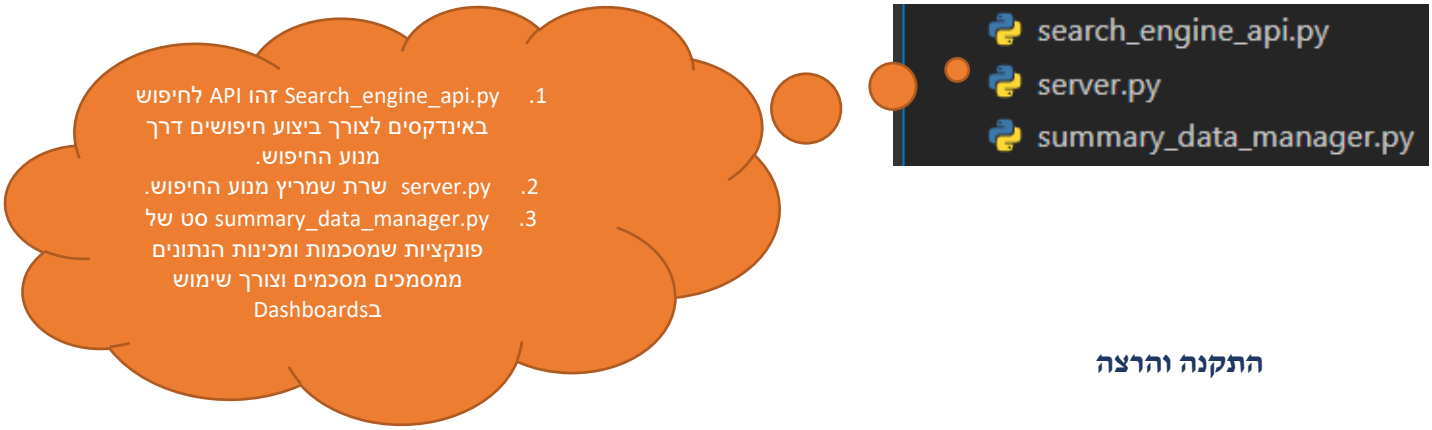
שני מסמכים שמייצרים דרכם את המאגר החדש וכל האינדקסים והמסמכים המסכמים הן עם ביצוע Stemming והן ללא Stemming , פותח בעזרת Google Colab לצורך שימוש ביצועים מהירים יותר בבניית המאגר




שני מסמכים שמייצרים דרכם את ממשקים ויזואליים והצגת התוצאות והנתונים, המסמך הראשון להצגת הסיבות להימנע מ Stemming

המסמך השני להצגת הנתונים בהנחה ולא עושים Stemming




## התקנה והרצה

- **הורדת הפרויקט:**
  - להוריד את הפרויקט למחשב האישי.
  - לבצע Unzipping.
  - לפתוח אותו דרך Visual Studio Code.

- **התקנת ה Packages הרלוונטיים דרך:**
  - pip install -r requirements.txt

    **OR**

  - pip3 install -r requirements.txt

- **הרצת מנוע החיפוש – יש לפעול לפי השלבים הבאים:**
  - יש להריץ את server.py.
  - נכנסים לדפדפן וגולשים ל http://127.0.0.1:5002/.

- **הרצת ה Dashboards - יש לפעול לפי השלבים הבאים:**
  - **להצגת ה Dashboard של הבעייתיות של Stemming:**
    - יש להריץ את dashboard_maker_stemming_issue.py.
    - נכנסים לדפדפן וגולשים ל http://127.0.0.1:8051/.
  - **להצגת ה Dashboard ללא Stemming:**
    - יש להריץ את dashboard_maker.py.
    - נכנסים לדפדפן וגולשים ל http://127.0.0.1:8050/.

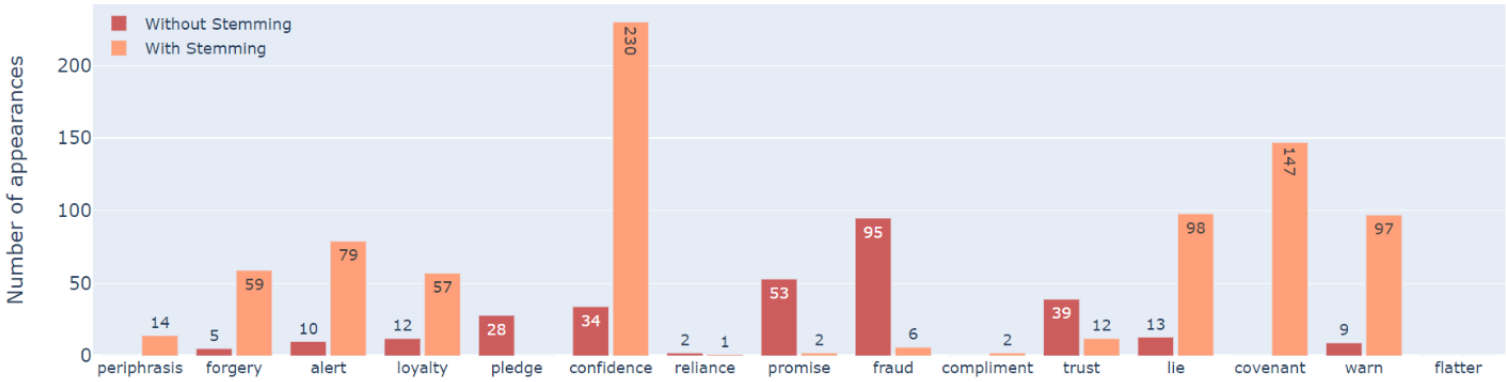
## פיתוח, מסקנות והחלטות

- הורדנו את המאגר שקבלנו, לאחר מכן ביצענו ניקוי Html Tags ושמרנו אותם במאגר חדש בתיקייה

```
Resources/My-Dataset/Clean-DB
```

- לגבי Stopwords לא ביצענו כי מילות האימון ומילות הרגש הן שנשמרים במילון באינדקס שבנינו ומשום שאין בהן מילות stopwords אז דלגנו על תהליך זה.
- לגבי Stemming ראינו שיש בעיות בביצוע אותה על מילות האימון ומילות רגש, אז בנינו והצגנו בצורה ויזואלית הבעייתיות (dashboard_maker_stemming_issue.py).
  - **בעיות Stemming:**
    - מילות שונות בעלות משמעויות שונות נופלות באותו תא של ביטוי אחרי stemming חלק מהן או כולם לפעמים הן לא מילות אימון או רגש ונופלות עם מילת אימון או רגש ולכן מקבלים המון False Positive

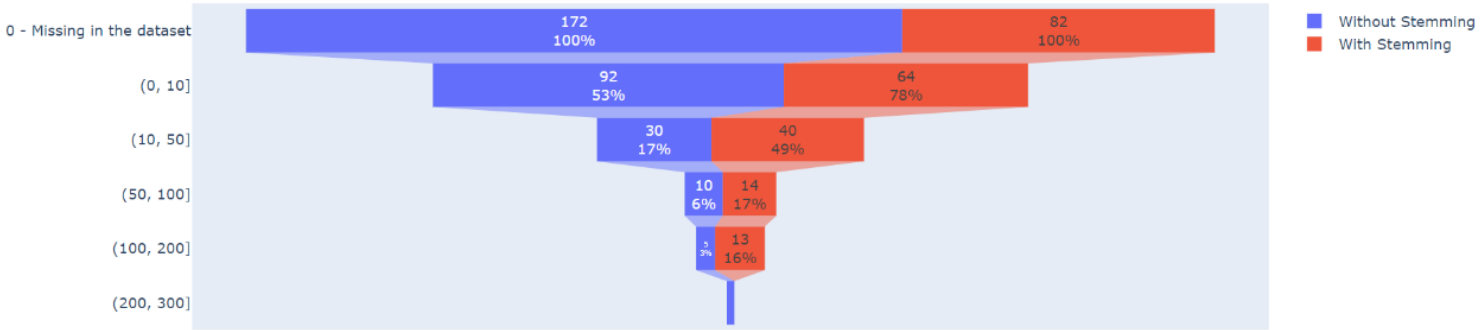
**1) Comparing between trust words with and without stemming**




**לדוגמה:** מילת האימון confidence במאגר הנתון מופיעה 34 פעמים וכאשר עושים stemming מספר ההופעות גדל וקופץ ל 230, דוגמה נוספת, מילת האימון covenant בכלל לא מופיעה במאגר אבל אחרי ה stemming נראה שמספר ההופעות קופץ מ 0 ל 147, אז בגלל שמילים אחרות נופלות על אותו ביטוי אחרי stemming מקבלים נתונים מזויפים.

- לא מעט מילים שלא היו מופיעות במאגר הנתון למרות אחרי stemming יצא שהן כן מופיעות:

**3) Comparing between emotion words with and without stemming**



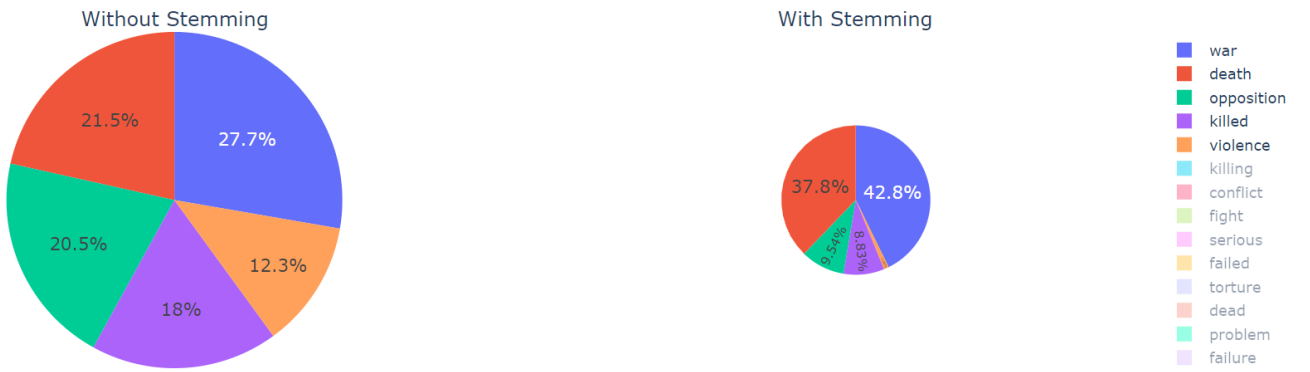

אז בטווח הופעות של ( 0 ) רואים שבמאגר האמיתי יש 172 מ 313 מילות רגש לא מופיעות בכלל, אך אחרי stemming רואים שמספר זה ירד ל 82 כלומר יש 90 מילים שכן אחרי stemming מופיעות במאגר.

דוגמה נוספת בטווח הופעות [ 10 , 0) ישנם 92 מילים אבל אחרי stemming יורד המספר ל 64.

להפך ממה שהיה בשתי דוגמאות קודמות בטווח הבא קרה ההפך נקבל False Positives ישנם 10 מילים שאחרי stemming נופלות עם אותו ביטוי של מילות הרגש ובכך יש זיופים בתוצאות.

- בגרף הבא רואים עד כמה ההשפעה של stemming על מילות הרגש וההתייחסויות שלהם במאגר, ניתן לראות עד כמה תהליך זה ישפיע על התוצאות ובחלק מהמילים נפסיד תוצאות ובחלק אחר נקבל תוצאות שגויות.

**4) Comparing emotion words with and without stemming**



- לצורך סיכום בעייתיות ה stemming הכנו טבלה מסכמת קצרה הבאה:

**6) Summary**

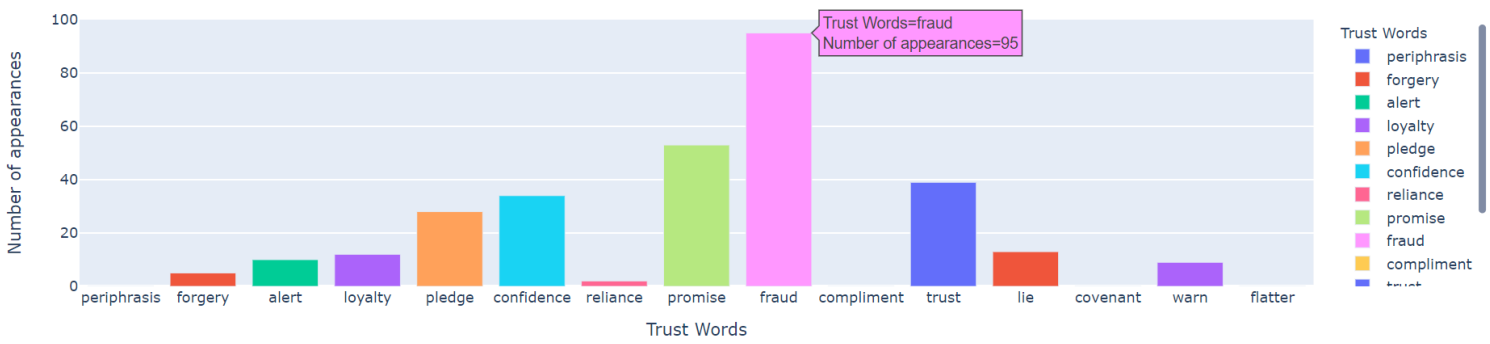
| Subject | Exisiting In The Dataset | Total | Percentage |
|---|---|---|---|
| Files Numbers | 773 | 773 | 100% |
| Sentences Numbers | 24174 | 24174 | 100% |
| Emotion Words | 142 | 224 | 63.39% |
| Trust Words | 13 | 15 | 86.67% |
| Trust-Emotional Pairs | 182 | 3360 | 5.42% |
| Duplicated Emotion Words | 94 | 224 | 41.96% |

רואים שרק 244 מילים רגש מופיעות במאגר מכלל 313 מילים רגש נתונים אחרי stemming ו 142 מהן מופיעות עם משפטים במאגר שמהווים 63% למרות שבאמת הן מהוות 45% ( רואים את זה ב dashboard_maker.py ), רואים שיש 94 מילות רגש שנופלות עם אותו ביטוי אחרי stemming, שמהווים כמעט 42%, כלומר מילות רגש אילו יש להן אותו שורש.

- **מבט וסטטיסטיקה על הנתונים האמיתיים – ללא stemming.**

# Assignment in Information Retrieval Course

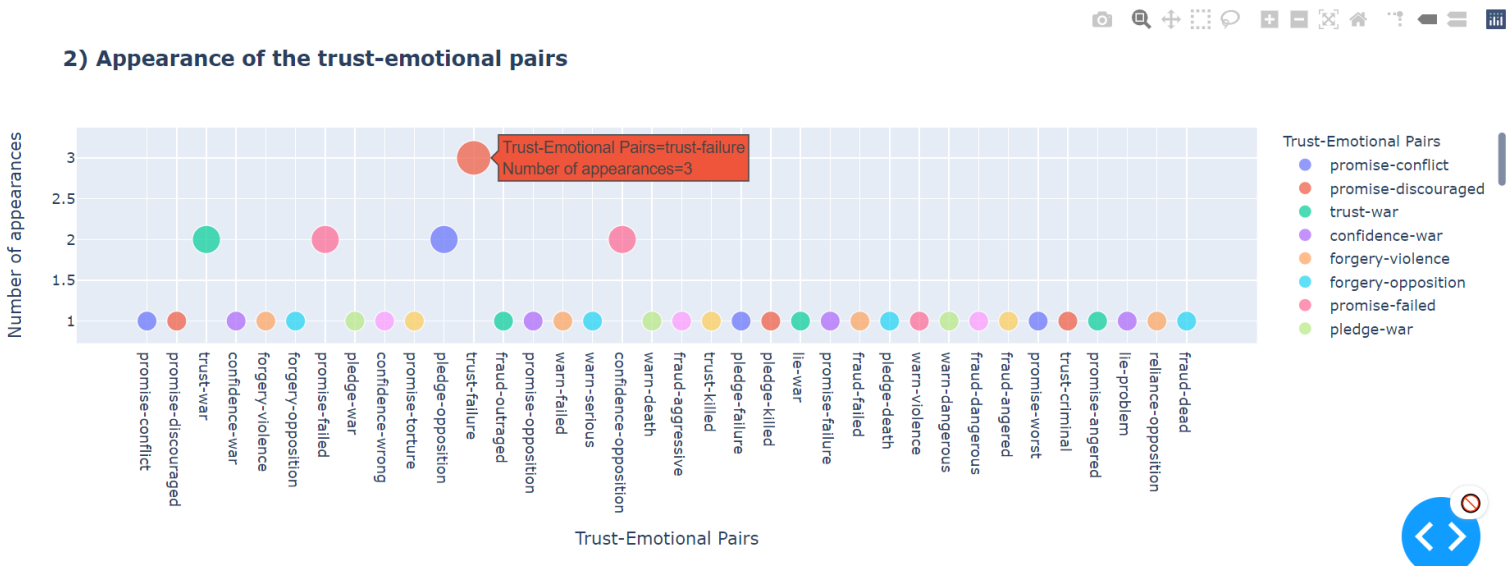
**1) Appearance of trust words in the given dataset**




בגרף זה רואים שכיחות מילות האימון ה 15 במאגר שקיבלנו.

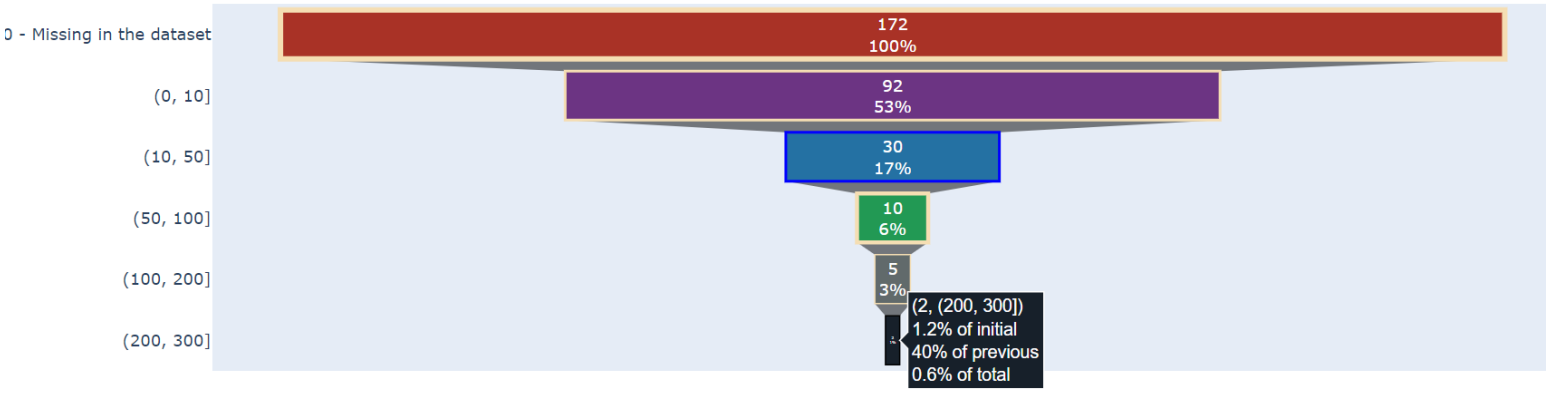
רואים שיש 4 מילה מכלל 15 שהן בכלל לא נמצאות במאגר, ורואים שהמילה fraud מופיעה 95 פעמים במאגר במסמכים שונים.

**2) Appearance of the trust-emotional pairs**




גרף זה מראה שכיחות הופעת מילות האימון עם מילת רגש, רואים שלמשל הזוג trust-Failure מופיעות ביחד 3 פעמים.

**3) Appearence of Emotion Words By Appearence Range**

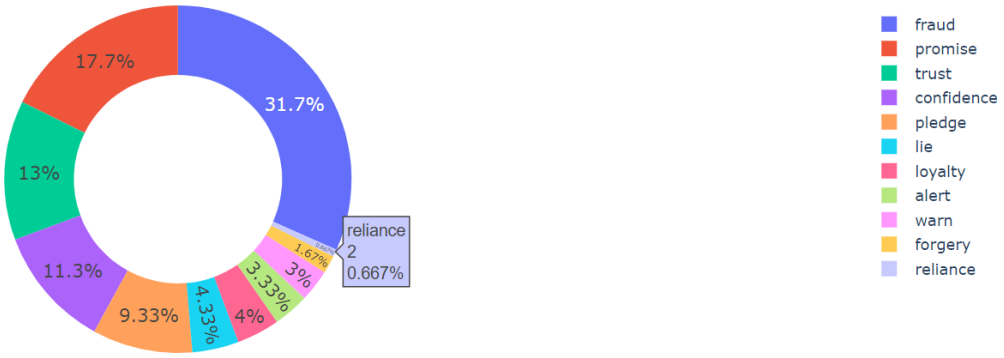



גרף זה מראה כמות ההופעות של מילות הרגש במאגר לפי טווחים, למשל רואים שיש 172 מילים שבכלל לא מופיעות במאגר, ו92 מופיעות בין 1 – 10 בעמים, ו 30 מילים מופיעות בין 10 ועד 50 פעמיים וכו.

**4) Percentage of existing emotion words in the given dataset**


גרף זה מראה התפלגות מילות הרגש במאגר, רואים שהמילה war מופיעה 10.8% ומילת הרגש Harassment מופיעה 13 פעמים.

**5) Appearence of trust words in the given dataset**




גרף זה מראה התפלגות מילות אימון שמופיעות במאגר ( שהן 11 מתוך 15), רואים ש fraud מופיעה 31.7% ו promise מופיעה 17.7% ו reliance מופיעה 2 פעמים שזה מהווה 0.667%.

### 6) Summary

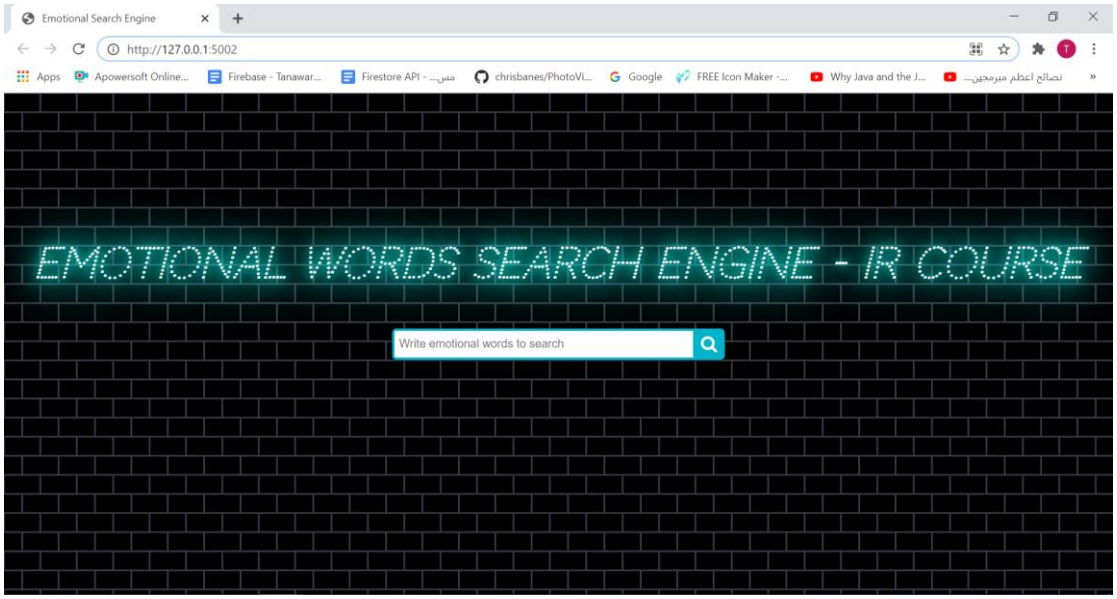
| Subject | Exisiting In The Dataset | Total | Percentage |
|---|---|---|---|
| Files Numbers | 773 | 773 | 100% |
| Sentences Numbers | 24174 | 24174 | 100% |
| Emotion Words | 141 | 313 | 45.05% |
| Trust Words | 11 | 15 | 73.33% |
| Trust-Emotional Pairs | 36 | 4695 | 0.77% |

טבלה זו מסכמת את מילות רגש ומילות אימון במאגר, רואים שבמאגר:

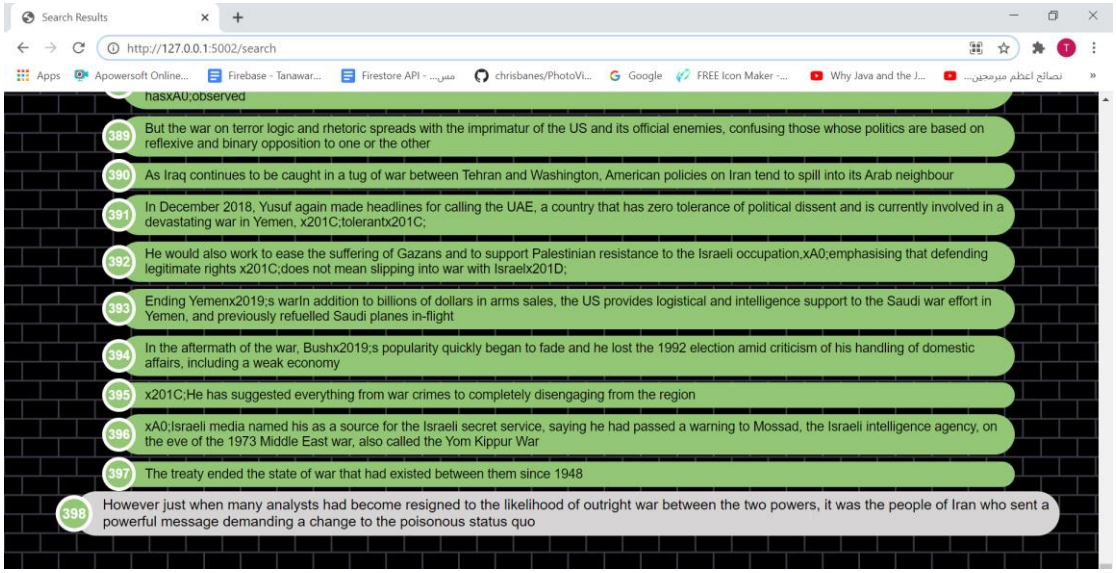
- יש 773 מסמכים.
- 24174 משפטים.
- נתונים 313 מילות רגש כך ש מעל 45% מהן מופיעות במאגר.
- נתונים 15 מילות רגש כך ש יותר מ 73% מהן מופיעות במאגר.
- רואים שיש 4695 קומבינציות להופעת מילות הרגש והאימון יחד במאגר כך ש 0.77% מהן כן מופיעות במאגר.

## מוצר

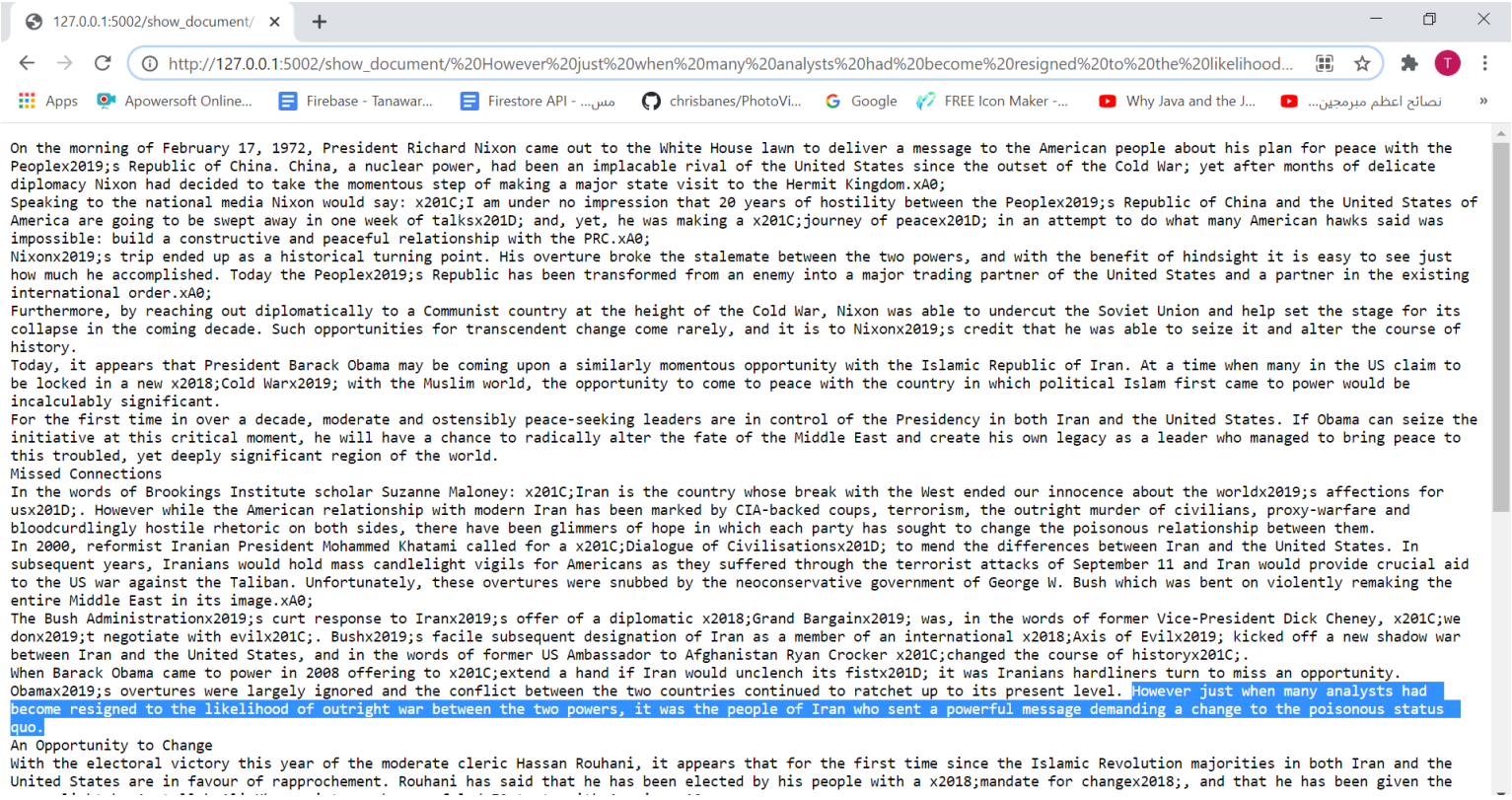
- פיתחנו גם מנוע חיפוש על אותה תשתית שהסקנו ממנה את הסטטיסטיקות שהראינו קודם.




למשל נחפש את המילה war



רואים שיש 398 תוצאות למילה זו ואם נלחץ על תוצאה כלשהי נעבור למסמך שמכיל אותה.




- Dashboard של בעייתיות Stemming בו. ( dashboard_maker_stemming_issue.py )
- Dashboard של התשתית שבנינו. ( dashboard_maker.py )