

## משימה מעשית מסכמת

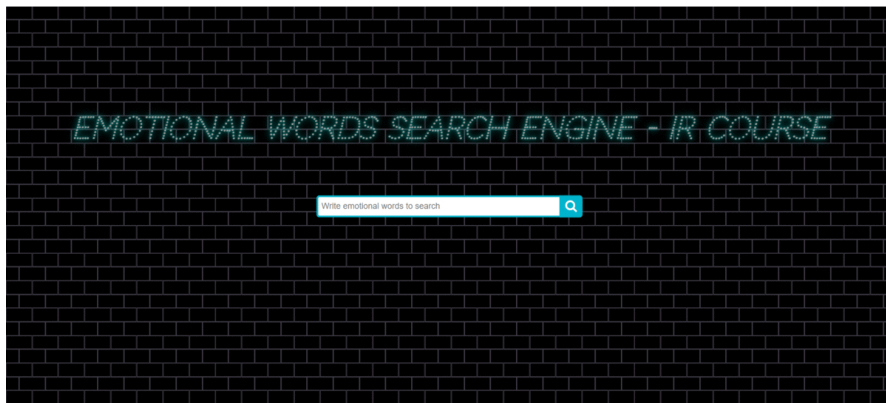
### קורס אחזור מידע

#### מגישים:

• מוחמד עביד - 313413346

• מוחמד חטיב - 208525220

מרצה הקורס: פרופ' צבי קופליק



### Assignment in Information Retrieval Course

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



## תוכן עניינים

הקדמה.....2

טכנולוגיות וסביבת פיתוח.....2

מבנה הפרויקט.....2

התקנה והרצה.....3

פיתוח, מסקנות והחלטות.....4

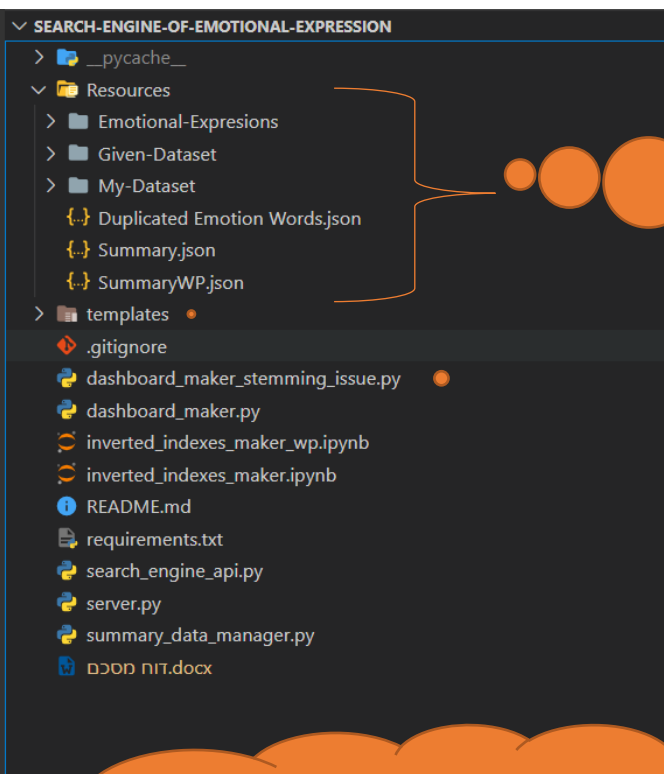
מוצר.....10

## 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 – תיקייה שמכילה את המסמכים המקוריים שקבלנו שנמצא בתת התיקיה Given Dataset, תת תיקייה שמכילה מילות האיומן 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 משוכללות 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:

```
inverted_indexes_maker_wp.ipynb  
inverted_indexes_maker.ipynb
```

```
dashboard_maker_stemming_issue.py  
dashboard_maker.py
```

שני מסמכים שמייצרים דרכם את המאגר החדש וכל האינדקסים והמסמכים המסוממים 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

1. Search\_engine\_api.py זהו API לחיפוש באינדקסים לצורך ביצוע חיפוש דרך מנוע החיפוש.  
 2. server.py שרת שמריץ מנוע החיפוש.  
 3. summary\_data\_manager.py סט של פונקציות שמסכמות ומכינות הנתונים ממסמכים מסכמים וצורך שימוש בDashboards

```
search_engine_api.py
server.py
summary_data_manager.py
```

## 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:

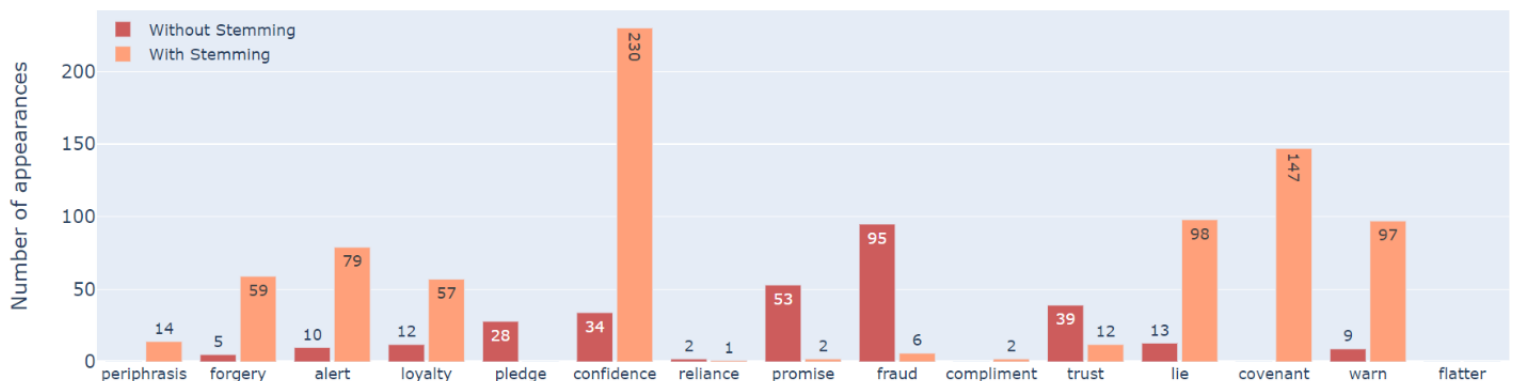
- להצגת ה Dashboard של הבעייתיות של Stemming:
  - יש להריץ את dashboard\_maker\_stemming\_issue.py.
  - נכנסים לדפדפן וגולשים ל <http://127.0.0.1:8051/>.
- להצגת ה Dashboard ללא Stemming:
  - יש להריץ את dashboard\_maker.py.
  - נכנסים לדפדפן וגולשים ל <http://127.0.0.1:8050/>.

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

[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

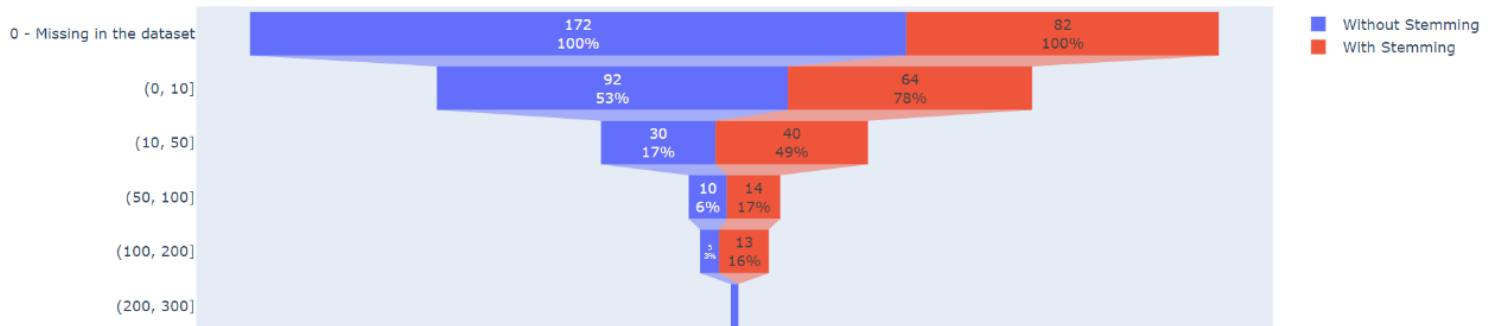
### 1) Comparing between trust words with and without stemming



**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



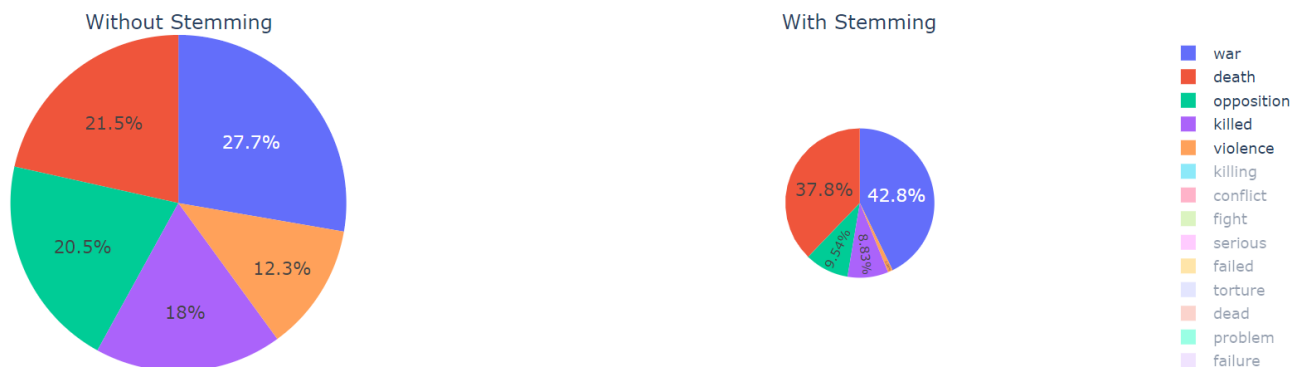
אז בטווח הופעות של (0) רואים שבמאגר האמיתי יש 172 מ 313 מילות רגש לא מופיעות בכלל, אך אחרי stemming רואים שמספר זה ירד ל 82 כלומר יש 90 מילים שכן אחרי stemming מופיעות במאגר.

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:

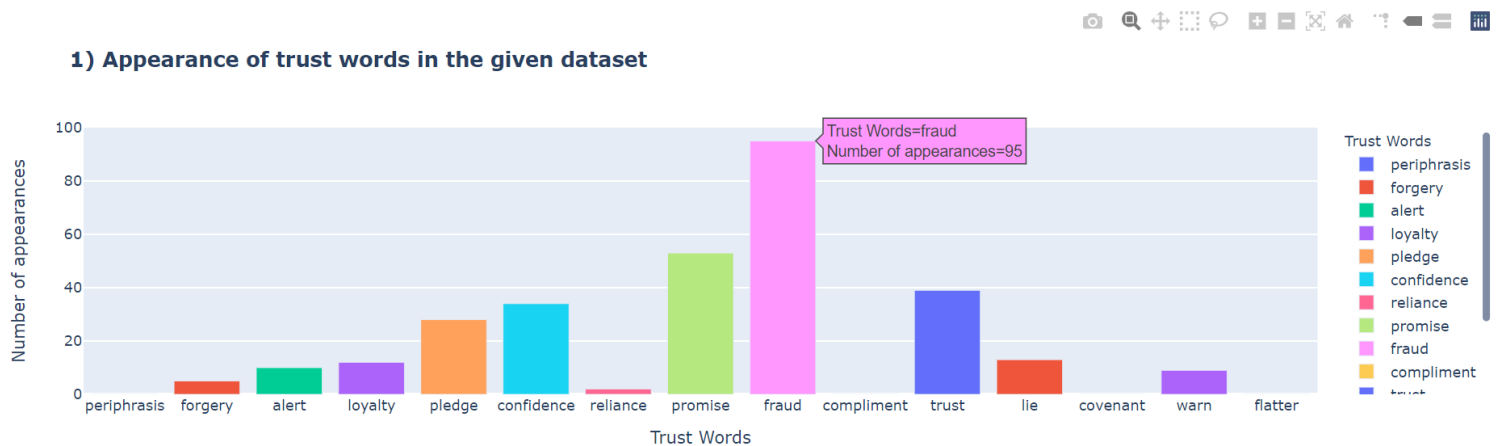
## 6) Summary

Subject	Existing 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.

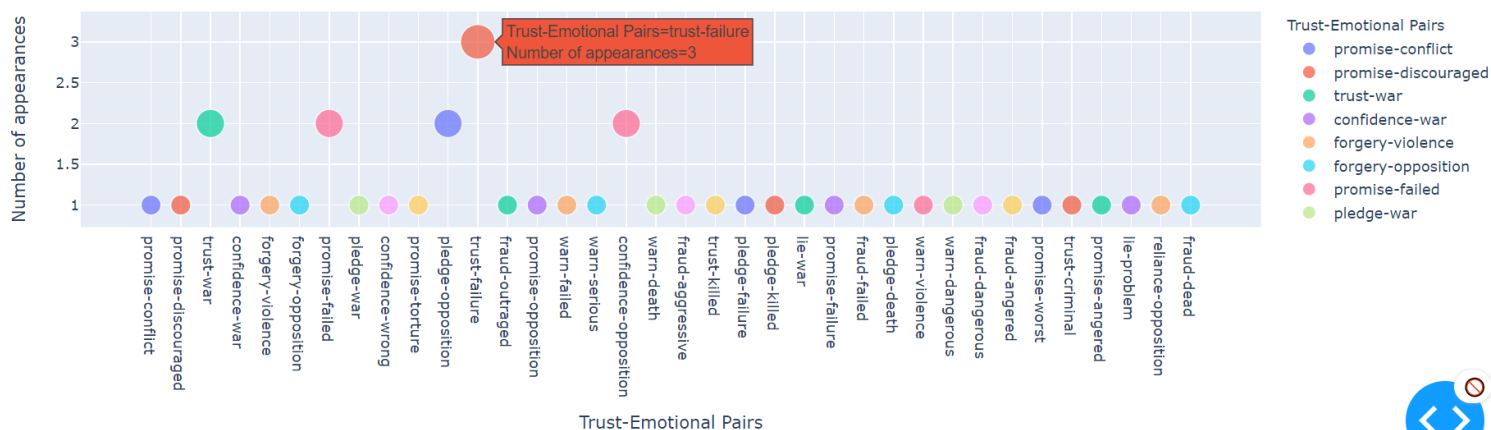
- מבט וסטטיסטיקה על הנתונים האמיתיים – ללא stemming.

## Assignment in Information Retrieval Course



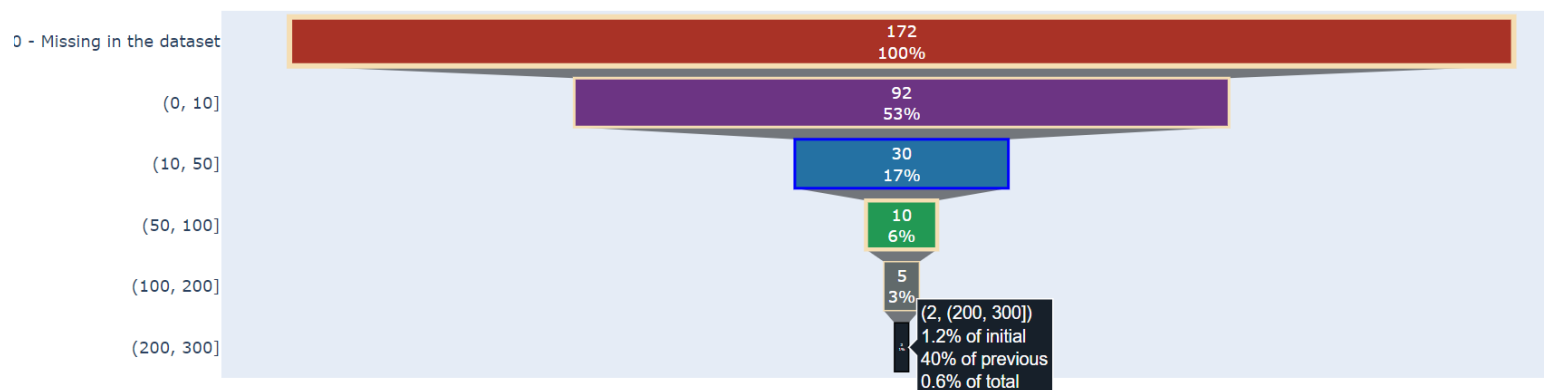
בגרף זה רואים שכיחות מילות האימון ה 15 במאגר שקיבלנו. רואים שיש 4 מילה מכלל 15 שהן בכלל לא נמצאות במאגר, ורואים שהמילה fraud מופיעה 95 פעמים במאגר במסמכים שונים.

## 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.

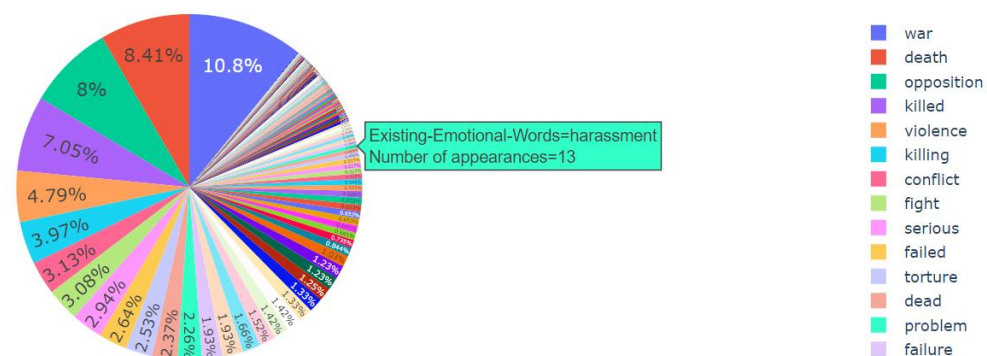
## 3) Appearance of Emotion Words By Appearance Range



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.

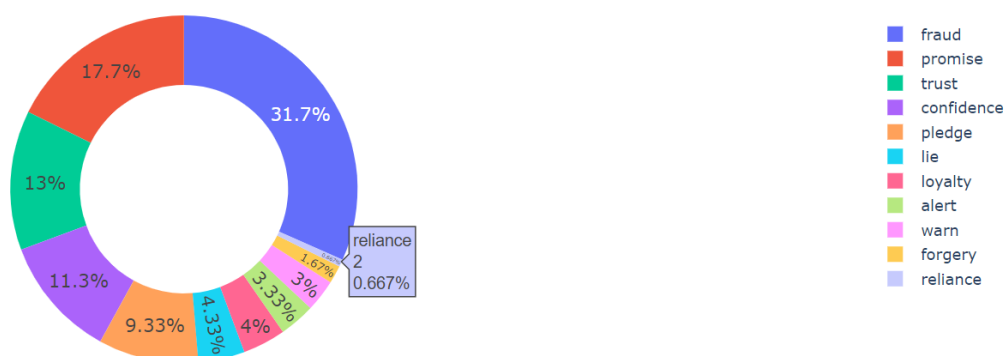


#### 4) Percentage of existing emotion words in the given dataset



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

#### 5) Appearance of trust words in the given dataset



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

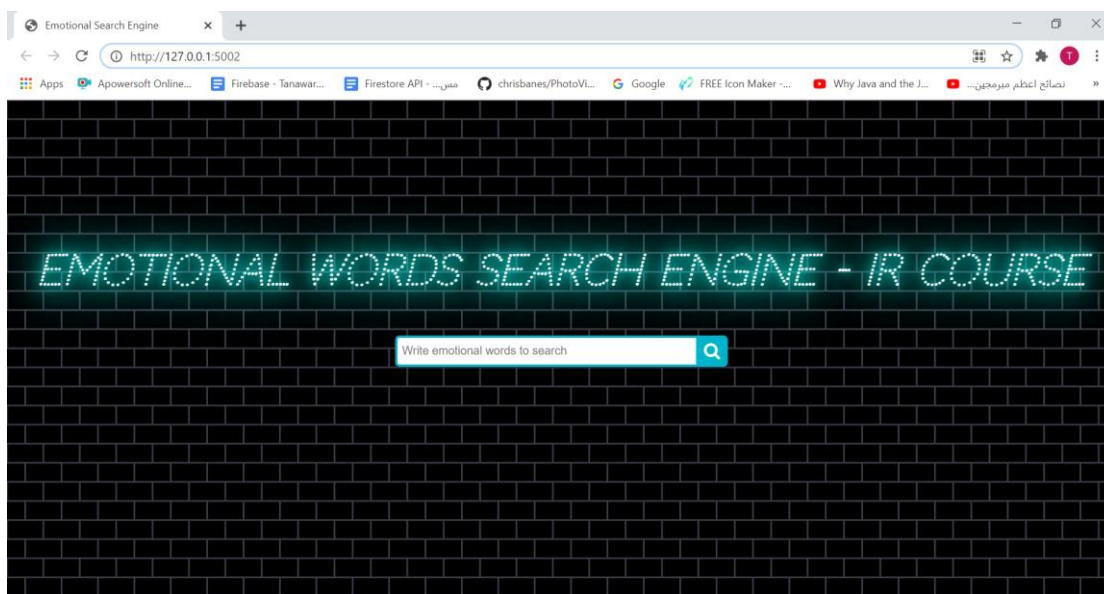
6) Summary

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%

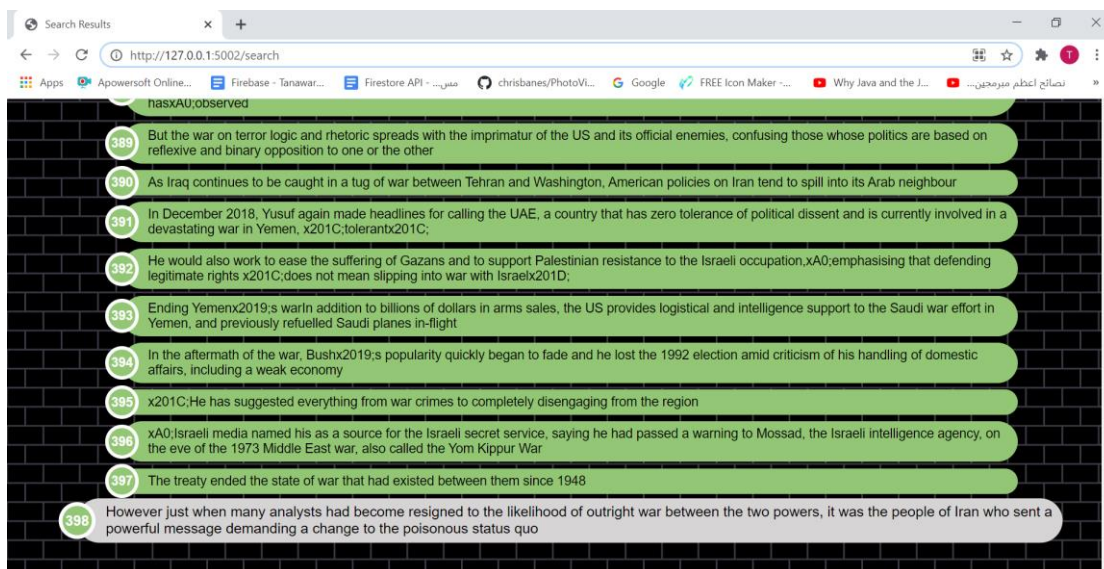
טבלה זו מסכמת את מילות רגש ומילות אימון במאגר, רואים שבמאגר:

- 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.

- 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.

127.0.0.1:5002/show\_document/ x +

http://127.0.0.1:5002/show\_document/%20However%20just%20when%20many%20analysts%20had%20become%20resigned%20to%20the%20likelihood...

Apps Apowersoft Online... Firebase - Tanawar... Firestore API - ... מסי... chrisbanes/PhotoVi... Google FREE Icon Maker ... Why Java and the J... نصاب اعظم ميرمحين...

On the morning of February 17, 1972, President Richard Nixon came out to the White House lawn to deliver a message to the American people about his plan for peace with the People's Republic of China. China, a nuclear power, had been an implacable rival of the United States since the outset of the Cold War; yet after months of delicate diplomacy Nixon had decided to take the momentous step of making a major state visit to the Hermit Kingdom. Speaking to the national media Nixon would say: "I am under no impression that 20 years of hostility between the People's Republic of China and the United States of America are going to be swept away in one week of talks; and, yet, he was making a journey of peace; in an attempt to do what many American hawks said was impossible: build a constructive and peaceful relationship with the PRC. Nixon's trip ended up as a historical turning point. His overture broke the stalemate between the two powers, and with the benefit of hindsight it is easy to see just how much he accomplished. Today the People's Republic has been transformed from an enemy into a major trading partner of the United States and a partner in the existing international order. Furthermore, by reaching out diplomatically to a Communist country at the height of the Cold War, Nixon was able to undercut the Soviet Union and help set the stage for its collapse in the coming decade. Such opportunities for transcendent change come rarely, and it is to Nixon's credit that he was able to seize it and alter the course of history. Today, it appears that President Barack Obama may be coming upon a similarly momentous opportunity with the Islamic Republic of Iran. At a time when many in the US claim to be locked in a new Cold War; with the Muslim world, the opportunity to come to peace with the country in which political Islam first came to power would be incalculably significant. For the first time in over a decade, moderate and ostensibly peace-seeking leaders are in control of the Presidency in both Iran and the United States. If Obama can seize the initiative at this critical moment, he will have a chance to radically alter the fate of the Middle East and create his own legacy as a leader who managed to bring peace to this troubled, yet deeply significant region of the world.

Missed Connections

In the words of Brookings Institute scholar Suzanne Maloney: "Iran is the country whose break with the West ended our innocence about the world's affections for us. However while the American relationship with modern Iran has been marked by CIA-backed coups, terrorism, the outright murder of civilians, proxy-warfare and bloodcurdlingly hostile rhetoric on both sides, there have been glimmers of hope in which each party has sought to change the poisonous relationship between them. In 2000, reformist Iranian President Mohammed Khatami called for a Dialogue of Civilisations; to mend the differences between Iran and the United States. In subsequent years, Iranians would hold mass candlelight vigils for Americans as they suffered through the terrorist attacks of September 11 and Iran would provide crucial aid to the US war against the Taliban. Unfortunately, these overtures were snubbed by the neoconservative government of George W. Bush which was bent on violently remaking the entire Middle East in its image. The Bush Administration's curt response to Iran's offer of a diplomatic Grand Bargain; was, in the words of former Vice-President Dick Cheney, "we don't negotiate with evil. Bush's facile subsequent designation of Iran as a member of an international Axis of Evil; kicked off a new shadow war between Iran and the United States, and in the words of former US Ambassador to Afghanistan Ryan Crocker "changed the course of history. When Barack Obama came to power in 2008 offering to extend a hand if Iran would unclench its fist; it was Iranians hardliners turn to miss an opportunity. Obama's overtures were largely ignored and the conflict between the two countries continued to ratchet up to its present level. However just when many analysts had become resigned to the likelihood of outright war between the two powers, it was the people of Iran who sent a powerful message demanding a change to the poisonous status quo.

An Opportunity to Change

With the electoral victory this year of the moderate cleric Hassan Rouhani, it appears that for the first time since the Islamic Revolution majorities in both Iran and the United States are in favour of rapprochement. Rouhani has said that he has been elected by his people with a mandate for change; and that he has been given the

- Dashboard של בעייתיות Stemming בו. ( dashboard\_maker\_stemming\_issue.py )
- Dashboard של התשתית שבנינו. ( dashboard\_maker.py )