Natural Language Processing – Final Project

Yuval Yarden - 318763455
Rom Eisenberg - 207811472

מבוא

מטרת הפרויקט היא לקבץ בקשות משתמשים שונות לקבוצות דומות (Clusters) על בסיס המשמעות שלהן, לבחור נציגים (Representatives) לכל קבוצה, וליצור תווית (Label) משמעותית לכל קלאסטר שתייצג אותו באופן הטוב ביותר. בפרויקט זה השתמשנו בטכניקות שונות של עיבוד שפה טבעית (NLP) כגון embedding, clustering, וניתוח תחבירי (POS tagging) על מנת להשיג את המטרה.

חלק א׳ – חלוקה לקלאסטרים (Clustering)

בשלב זה השתמשנו באלגוריתם איטרטיבי שמחלק את הבקשות (Requests) לקלאסטרים (Clusters) על סמך דמיון הווקטורים (Embeddings) של הבקשות. העקרונות המרכזיים של השלב הזה הם:

- **מבנה הנתונים:**

clusters: כל קלאסטר מיוצג כרשימה שמכילה:

"indices" – רשימת האינדקסים (מה-dataset) של הבקשות השייכות לקלאסטר.

"centroid" – הצנטרואיד של הקלאסטר, המחושב כממוצע של וקטורי הEmbeddings של הבקשות בו.

- **תהליך החלוקה:**

1. עבור כל בקשה (על פי סדר מסוים, אשר יכול להיות רנדומלי) אנו מחשבים את הדמיון (באמצעות מכפלה פנימית, שהיא ייצוג של cosine similarity בהנחה שהווקטורים מנורמלים) בין וקטור הבקשה לבין הצנטרואידים של כל הקלאסטרים הקיימים.
2. אם הדמיון הגבוה ביותר גבוה מסף מסוים (similarity_threshold), אנו משייכים את הבקשה לאותו קלאסטר, ונעדכן את הצנטרואיד בהתאם.
3. אם לא נמצא קלאסטר קיים שקרוב מספיק, אנו יוצרים קלאסטר חדש עם אותה בקשה.
4. התהליך חוזר שוב ושוב עד שהגענו לאיטרציה שבה אין יותר שינויים או עד למספר איטרציות מקסימלי.

- **הפרדת בקשות "לא משוייכות" (Unclustered):**

בסיום האיטרציות, אנו מגדירים קלאסטרים תקינים כאלו שיש להם לפחות מספר מינימלי של בקשות (min_cluster_size = 10).

כל הבקשות שלא שייכות לקלאסטר תקין, כלומר כאלו שנשארו בקלאסטרים קטנים נספרות כ"בקשות שלא משוייכות לקלאסטר" (unclustered).

הערה - שמנו לב שהוקטורים של ה-embeddings מנורמלים כבר ולכן אורכם 1, וכאשר השתמשנו בcosine similarity אמנם קיבלנו דיוק גבוה, אבל ראינו שאנחנו מחלקים ב1.000001 או 0.999999 שכן משפיעים במידה מסוימת על הדיוק ולכן החלטנו להשתמש במכפלה פנימית ובעצם לחשב רק את המונה.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

תוצאות:

banking

```
Converged after 7 iterations.
clusters in 1st and 2nd solution: 65 and 65
unclustered requests in 1st and 2nd solution: 262 and 248
rand score: 0.9909370596388278
adjusted rand score: 0.8474280562338228
```

covid19

```
Converged after 5 iterations.
clusters in 1st and 2nd solution: 36 and 35
unclustered requests in 1st and 2nd solution: 590 and 595
rand score: 0.9784992956431712
adjusted rand score: 0.9350084364651388
```

חלק ב׳ – בחירת נציגים (Representative Choosing)

בשלב זה אנו רוצים לבחור מספר בקשות מייצגות מתוך כל קלאסטר, כך שהן יהיו שונות זו מזו וייצגו את המגוון שבקלסטר. מטרת תהליך זה היא לבחור נציגים כך שישמרו גם גיוון בתוך הקלאסטר וגם יתפסו את העיקר שבו.

תיאור תהליך בחירת הנציגים:

הקוד בוחר נציגים לכל קלאסטר במטרה לייצג את הבקשות בקלסטר, תוך שמירה על איזון בין רלוונטיות לתוכן של הבקשות בקלסטר וגיוון (Diversity). הבחירה מתבצעת בשני שלבים מרכזיים:

בחירת נציגים בתוך כל קלאסטר (פונקציה get_cluster_representatives)

המטרה היא לבחור **שלושה** נציגים המייצגים גיוון בתוך הקלאסטר:

1. **בחירת הנציג הראשון:**
   - סופרים את כמות המופעים של כל הבקשות בקלאסטר.
   - הבקשה שמופיעה הכי הרבה פעמים תבחר בתור הנציג הראשון - בחירה זאת תבטיח ייצוג טוב של הקלאסטר מכיוון שמבוססת על הבקשה שחזרה הכי הרבה.

2. **פרמטרים לחישוב הscore:**
   - לכל בקשה שטרם נבחרה מחשבים:
     - **רלוונטיות** – המרחק שלה לצנטרואיד - ככל שקרובה יותר תקבל score גבוה יותר.
     - **גיוון** – ממוצע המרחקים בין הבקשה לכל אחד מהנציגים שכבר נבחרו - ככל שרחוקה יותר משאר הנציגים תקבל score גבוה יותר (באופן הכי משמעותי).
     - **תדירות** – כמות המופעים של הבקשה בקלאסטר לפי נרמול יחסית למספר המופעים הגבוה ביותר של בקשה (מספר בין 0-1) - ככל שהופיעה יותר פעמים היא תקבל score גבוה יותר.

3. **חישוב ניקוד ובחירת שאר הנציגים:**
   - ה-score מחושב לפי הפונקציה הבאה:

$$score = (3 \times diversity) + (1.5 \times frequency) + (0.5 \times relevance)$$

נתנו חשיבות גבוהה ל-diversity כפי שנדרש תוך כדי מתן דגש לרלוונטיות כדי לא לבחור outliers שלא יאפיינו נכונה את הקלאסטר. בנוסף, העלנו את הניקוד למשפטים שחוזרים הרבה פעמים כי הם מייצגים בצורה טובה את הקלאסטר.

דוגמאות:

```
"cluster_name": "Transfer to an account",
```

```
"representatives": [
    "i couldnt transfer money from my account.",
    "how do i do a successful transfer to an account?",
    "i tried to transfer cryptocurrency into my account but was denied"
]
```

```
"cluster_name": "Coronavirus test",
```

```
"representatives": [
    "testing spots for coronavirus?",
    "coronavirus test kit",
    "who do i contact if i want to get tested for covid 19 infection?"
]
```

```
"cluster_name": "Activate my card",
```

```
"representatives": [
    "how can i activate the new card i got?",
    "i need to actuate my card.",
    "how long does it take to activate the card, as it does not look like its working?"
]
```

ניתן לראות שבאמת כל הנציגים רלוונטיים לתוכן הקלאסטרים אבל גם מגוונים בדרך הבקשה שלהם.

תהליך הפתרון - שיפור האלגוריתם:

בתחילה ניסינו להשתמש באלגוריתם אחר שהתעלם מכמות המופעים וחיפש בעיקר diversity גבוה גם על חשבון רלוונטיות.
בתור נציג ראשון בחרנו את הבקשה שהכי רחוקה מהצנטרואיד ואז את שאר הנציגים בחרנו באמצעות הפונקציה: $score = -relevance + diversity$ וקיבלנו לדוגמא:

שם הקלאסטר (Label):

```
"cluster_name": "Activate my card",
```

המשפט שהכי קרוב לצנטראויד:

```
"how do i activate my card",
```

הנציגים:

```
"representatives": [
    "how can i make my card ready to use?",
    "i couldn't complete card activation.",
    "how long should i wait ti activate my card"
]
```

ולמרות שהבקשות שנבחרו היו בהחלט מגוונות, שמנו לב שבחרנו בעיקר בoutliers שלא באמת מייצגים את הקלאסטר ולכן החלטנו להתחשב גם ברלוונטיות וגם בתדירות של הבקשה בקלאסטר.

חלק ג׳ – תיוג הקלאסטרים (Labeling)

המטרה כאן היא ליצור לייבלים קצרים, תקינים תחבירית ואינפורמטיביים לכל קלאסטר. הגישה שבחרנו היא להפיק קבוצה של n-grams מתוך הבקשות של הקלאסטר ולדרג אותם על סמך כמה קריטריונים:

1. **יצירת n-grams:**
   - השתמשנו ב־CountVectorizer להפקת n-grams (בין 2 ל־5 מילים) מהטקסטים של הקלאסטר.
   - השארנו את כל המילים כולל stop words, כך שיש לנו את מלוא המשפט, אבל וידאנו שהמשפט לא מכיל בעיקר stop words אלא מילים משמעותיות עם קונטקסט בעזרת דירוג ה-n-grams.
2. **דירוג n-grams:**
   - **תדירות:** ככל שה-ngram מופיע יותר פעמים, הוא מקבל ניקוד גבוה יותר.
   - **אורך:** מוסיפים משקל קטן מ-1 לאורך שלו (מספר המילים בחזקת ערך מסוים) – כך ש-ngramים ארוכים מידי לא יקבלו העדפה.
   - **בונוס מילים משמעותיות:** כדי לוודא שn-grams המכילים את המילים המשמעותיות ביותר בקלאסטר מקבלים עדיפות הוספנו עוד בונוס ל-score. בהתחלה ביצענו עיבוד של כל הטקסטים בקלאסטר כדי שכל המילים יהפכו לאותיות קטנות ויעברו תהליך למטיזציה (המרה לצורת הבסיס שלהן, למשל *currency* → *currencies*) כדי שיספרו עבור אותה מילה. לאחר מכן הורדנו את ה-stop words כדי לא לספור מילים נפוצות ללא קונטקסט משמעותי.
     ביססנו את הבחירה על חישוב ישיר של שכיחות המילים והתמקדות בשמות עצם (NN) ובפעלים (VB), שהם לרוב המילים המשמעותיות ביותר בהבנה של כוונת המשפט ויותר נותנות מידע על הקונטקסט של ה-request.
   - **בונוס תבנית POS:** אנו בודקים תבניות תחביריות (למשל, אם המשפט מתחיל בפועל ואחריו שם עצם) ומוסיפים בונוס אם התבנית מתאימה.
   - **פסילה של n-grams:** נבצע סינון לביטויים פוטנציאליים לתוויות הקלאסטרים, בהתבסס על המבנה הדקדוקי שלהם. ביצענו זאת כדי למנוע בחירת ביטויים לא תקינים מבחינה תחבירית, במיוחד כאלה שמתחילים או מסתיימים במילים שלא מתאימות להופיע באופן עצמאי בכותרת - כמו למשל הlabel:

```
"cluster_name": "To transfer some money to",
```

   משפט שמסתיים ב-to הוא לא תקין תחבירית.
3. **בחירת התווית:**
   - לאחר חישוב ה-Score עבור כל המועמדים, בוחרים את ה-ngram עם ה-Score הגבוה ביותר והופכים אותו ללייבל של הקלאסטר.

תהליך הפתרון:

בהתחלה ניסינו לסווג באופן בסיסי רק עם שימוש ב-ngrams, אבל התוצאות שקיבלנו לא היו מספיק טובות, לדוגמא:

```
"cluster_name": "New card",
"requests": [
    "can i order a new card to china?",
    "how to get a new card in china",
    "how do i get a new card sent to china?",
    "where can i order a card when i am in china?",
    "can i have a new card shipped to china?",
    "if i am in china, can i still order a new card and if so, how?",
    "i am overseas in china, can i get a replacement card?",
    "is it possible to get a new card in china?",
    "want to get a new card in china",
```

ניתן לראות שאכן ה-ngram הנפוץ ביותר הוא new card אבל חסר פה מידע משמעותי על המיקום - china שמשותף לרוב המוחלט של ה-requests בקלאסטר הנ"ל. ולכן, לאחר מכן ניסינו לתת עדיפות למשפטים ארוכים יותר בעזרת score גבוה יותר עבור אורך המשפט וקיבלנו:

```
"cluster_name": "Get new card in china",
"requests": [
    "can i order a new card to china?",
    "how to get a new card in china",
    "how do i get a new card sent to china?",
    "where can i order a card when i am in china?",
    "can i have a new card shipped to china?",
    "if i am in china, can i still order a new card and if so, how?",
    "i am overseas in china, can i get a replacement card?",
    "is it possible to get a new card in china?",
    "want to get a new card in china",
```

כלומר פתרנו את הבעיה הזאת, אבל יצרנו בעיות אחרות:

```
"cluster_name": "Before it goes through",
"requests": [
    "i messed up a transfer and need to reverse it.",
    "i need to cancel my recent transfer. i made a mistake. please help quickly before the transfer goes through.",
    "it is extremely important that i cancel the transfer i made yesterday and put the money into a different account.  is thi
    "i need to cancel my recent transfer immediately, i made a mistake there, please help quickly before it goes through",
    "is there any way to cancel a transfer?",
    "a transfer has been made in my account that i need to cancel.  this was a mistake on my part and i hope we can get this r
    "what do i have to do to cancel a transfer?",
    "i just initiated a transfer but i'd like to cancel it.",
    "i have arranged a transfer, but no longer need to send it. how do i stop the transfer?",
    "please cancel my most recent transfer, it was a mistake. this is an emergency. it needs to be canceled before it goes thr
    "cancel my last transfer",
```

```
"cluster_name": "My card is about to",
"requests": [
    "my card is about to expire. what do i need to do to keep using my card?",
    "since my card is about to expire, what do i do to get a new one?",
    "what is the process when my card is due to expire?",
    "my card is about to expired what will happen?",
    "soon my card will expire, how do i get a new one?",
    "what happens after my card expires?",
    "what do i do if my card is about to expire?",
    "my card's expiring, what happens now?",
    "what do i do when my card expires?",
```

```
"cluster_name": "Allow me to do it",
"requests": [
    "i have been trying to exchange this for crypto but the app won't work could you please help me?",
    "hi, i'm trying to by some crypto and the app isn't allowing it. i really want to exchange this, what am i doing wrong?",
    "i wanted to purchase some crypto. the app will not let me. what's going on here? i really want to make this exchange.",
    "hello i am unable to exchange the crypto currency through the app, i am very keen to buy. please let me know the issue.
    "hello - i'm on the app and trying to purchase crypto. it's not going through. what am i doing wrong?",
    "hey i want to buy some crypto but the app doesn't allow me to! what's the issue, i really want to exchange this",
    "hi,  is there any problem with app? as i am facing some issues with the exchange of crypto currency. i am extremely inter
    "i was trying to purchase some crypto and the app won't allow me to do it. what's the deal with this? i really wanted to c
    "i really was trying to complete this exhange. i was trying to purchase some crypto but the app won't allow me to do it. c
    "can you please help me with this exchange?  i am trying to get crypto and the app won't let me.",
    "i'm trying to purchase crypto via the app. i haven't been able to get it to go through. am i doing something wrong?",
    "hi, i am interested buying crypto currency but unable to purchase it through the application. i do want to do the exchang
```

העדפנו משפטים ארוכים יותר, מה שגרם לתעדף ngrams ארוכים גם אם הם פחות נפוצים ולכן קיבלנו הרבה labels חסרי משמעות (דוגמא 1), לא נכונים תחבירית (דוגמא 2) או לא קשורים לקונטקסט המרכזי של הקלאסטר (דוגמא 3).

ולכן כדי לפתור זאת היינו צריכים להגדיר מה נחשב label טוב, ולאחר ניתוח של הlabels הגענו למסקנות הבאות:

1. **אורך ה-label:** נעדיף לייבלים באורך של 2-4 מילים ורק במקרים חריגים 5 מילים, כך נקבל משפט עם מספיק מידע אבל מצד שני לא ארוך מידי שמכיל מילים מיותרות.

2. **מילים משמעותיות:** כדי לבחור label משמעותי, נרצה להתמקד במילים שמייצגות בצורה הברורה ביותר את התוכן של הקלאסטר:

- נרצה להימנע מ-Stop Words כמו *"the", "is", "to", "for", "my"* אלא אם יש להן משמעות קריטית בצירוף.
- נביא עדיפות לשמות עצם ופעלים שחוזרים הרבה בקלאסטר, כי אלו מילים שקובעות את המשמעות במשפטים.
- מילים שחוזרות שוב ושוב בטקסטים של הקלאסטר הן לרוב בעלות חשיבות גבוהה, ולכן השתמשנו בפונקציה **get_top_k_significant_words** כדי לזהות מילים שחוזרות בתדירות גבוהה במיוחד.

3. **תחביר תקין:** תחביר הוא קריטריון קריטי, ולכן כדי להבטיח איכות תחבירית, נתנו עדיפות לתבניות הבאות:

א. מבנה ציווי (Imperative Sentence):

במבנה זה, המשפט מתחיל בפועל בצורת ציווי (Imperative) כמו Activate, Cancel, Change, דבר שמצביע על הנחיה לפעולה. לאחר הפועל, ניתן למצוא (Determiner) כמו "my" ,"the" או "a", ולאחר מכן שם עצם (Noun) שמגדיר את הפעולה. לדוגמא: "Activate my card".

ב. צירופי שם עצם (Noun Phrase)

מבנה נפוץ נוסף הוא צירוף שם עצם (Noun Phrase), שבו המשפט מורכב רק משמות עצם ותארים. מבנים אלו עדיפים, כי הם קצרים, אינפורמטיביים, ונוחים לקריאה. לדוגמא: "Exchange fee".

ג. צירופים קצרים של פועל + שם עצם

כאשר ה-label מכיל פועל + שם עצם, הוא קצר ונותן משמעות ברורה לפעולה שיש לבצע. לדוגמא: "Cancel transfer".

לאחר שהטמענו את השיטה של ה- score שמנו לב שה-labels שהתקבלו דווקא ארוכים יותר ברוב המקרים ויותר שואפים לכיוון ה4-5 מילים מכיוון שהעדפנו label עם כמות גדולה של מילים משמעותיות. ולכן הורדנו את ה-weight של אורך ה-label למספר שקטן מ-1 כדי לקבל labels מעט קצרים יותר.

קיבלנו ברוב המקרים labels שעונים על הדרישות (עדיין יש כמה labels בודדים שלא עונים על חלק מהדרישות):

```
"cluster_name": "Transfer some money",
"requests": [
    "i needed to send my friends some money urgently.  i tried multiple times to transfer the money this morning, but the tra
    "the transfer keeps rejected, i tried to transfer some money to friends but it keeps getting rejected for some reason, wo
    "the transfer keeps failing - i tried to transfer some money to friends this morning but it keeps getting rejected for so
    "please help me as i am continuously facing the issue in transferring money to my friends, as all my transactions are get
    "hi i'm sending some money over to an investment bank i do business and the transaction is being rejected. i'm not sure w
    "i've tried to send money to my friends and i keep getting rejected. can you please tell me why, i don't understand.",
    "the transfer keeps failing , i tried to transfer some money to friends this morning but it keeps getting rejected for so
    "hello. i'm trying to transfer some money to a friend but it keeps getting rejected. can you help me?",
```

```
"cluster_name": "Open an account",
"requests": [
    "how old do i need to be to open an account?",
    "at what age can a person open an account?",
    "how old does one have to be to have an account with the bank?",
    "at what age can i open an account?",
    "how old do you need to be to open an account?",
    "how young can i be to open my own account?",
    "how old do you need to be to use the banks services?",
```

```
"cluster_name": "Sore throat",
"requests": [
    "i have sore throat",
    "if i only have a sore throat is it coronavirus?",
    "my throat hurts and i have a headache as well",
    "i have a sore throat",
    "i have a sore throat",
    "ihave sore throat",
    "my throat and head hurt",
```

הערה - חשוב לציין שכל שינוי עדין שביצענו בscore השפיע משמעותית על הLabels שקיבלנו ולאחר תהליך של ניסוי וטעייה הגענו למסקנה שמאוד קשה שכל הlabels יענו על הדרישות באופן מושלם, שכן כל שינוי קטן אמנם שיפר לייבלים מסוימים אך פגע באחרים, ולכן בסופו של דבר חיפשנו את האיזון הכי טוב.

סיכום

- **חלוקה לקלאסטרים:**
בנינו אלגוריתם איטרטיבי שמקצה בקשות לקלאסטרים על סמך דמיון לצנטרואיד, ומפריד בין קלאסטרים תקינים (מעל סף הבקשות המינימלי) לבין בקשות "לא משוייכות".
- **בחירת נציגים:**
עבור כל קלאסטר בחרנו נציגים של בקשות מהקלאסטר באמצעות אלגוריתם שבנינו, שמבטיח גיוון בין הנציגים וגם רלוונטיות לקלאסטר.
- **תיוג קלאסטרים:**
יצרנו n-grams מהבקשות, דירגנו אותם לפי תדירות, אורך ותבניות POS, והעדפנו מועמדים שיש בהם יותר מילים תוכניות ופחות stop words – כך שהלייבלים יהיו קצרים, נכונים תחבירית ואינפורמטיביים.
- **זמן ריצת התוכנית:**

banking-requests:

```
Total runtime: 24.439527988433838 seconds
```

covid19-requests:

```
Total runtime: 9.073589086532593 seconds
```