

הוראות כלליות: על כל תרגיל עליכם לכתוב מסמך המציג הסברים על מה שעשיתם, את דרך העבודה שלכם, מדוע בחרתם בדרך שבחרתם. את התוצאות ונתוח של התוצאות והשגיאות (אם ישנן).

את כל הפתרונות עליכם להעלות למודל. הפתרון **חייב לכלול** גם את הקוד עצמו בקובץ של פייתון (ולא קישור) ואת המסמך עם הסברים מפורטים עם טבלאות, תרשימים, פלוטים וכדומה.

על כל תרגיל להגיש תוך שלושה שבועות, ההגשה היא במודל ובמעבדות יתבצעו הגנות על התרגילים.

תרגיל 1

הפרלמנט הבריטי (יותר כללי):

<https://data.mysociety.org/datasets/uk-hansard/>

קבצי ה-XML של הפרלמנט:

<https://www.theyworkforyou.com/pwdata/scrapedxml/debates/>

ישנם לא מעט קבצים, אני מעוניין בקבצים מ- debates2023-06-28d.xml עד הסוף, בערך 935 קבצים.

עליכם "לנקות את הקבצים" (כל קובץ צריך לשמור על השם/המספר שלו) משמע שכל מילה תעמוד בפני עצמה ללא שום תווים נוספים, הכוונה היא להפריד את המילה מסימן הפיסוק הסמוך לה (יש פונקציה בפייתון שעושה את זה) לדוגמה (בעברית):

1. **שיהיה לכם שבת שלום טוב ומבורך.** - המילה **ומבורך** שונה מהמילה **ומבורך** הנקודה מציינת סוף משפט. (התוצאה היא שיהיה לכם שבת שלום טוב ומבורך.)
2. ... היום ראש הממשלה אמר **"אני מצפה למשמעת קואליציונית"** בהתייחסו להצבעה על ... - המילה **"אני"** שונה מהמילה **אני** וכן"ל למילה **קואליציונית"** (שימו לב שהמחרוזת **צה"ל** היא מחרוזת תקינה)
3. ועוד תווים מיוחדים נוספים כמו פסיק, נקודתיים : נקודה-פסיק וכדומה.

לכל אחד מקבצי המקור תצרו קובץ עם הלמות (גזע המילים) של המילים שלו. לצורך זה עליכם להשתמש בכלי מתאים, בדו"ח תכתבו באיזה כלי השתמשתם, עדיף להשתמש בכלי הכי טוב שתוכלו.

לכל אחד מהקבצים של עליכם לבנות ווקטור מייצג, כאשר אתם שומרים את שיוך הווקטורים לקבוצה ולמספר/שם הקובץ.

1) את הווקטורים תבנו בעזרת TF-IDF עם BM25/Okapi

הואיל ורוב מוחלט של המסמכים אינם מכילים את כל המילים בשפה, ויותר מכך המסמכים הם יחסית קצרים, אז אתם תקבלו מטריצות דלילות. עליכם לצמצם את המאפיינים על מנת שלא תתקבל מטריצה מאוד גדולה, למשל ניתן להוריד stop-words, מילים המופיעות פחות מ-5 פעמים וכדומה (אלו הצעות אתם יכולים לחשוב על רעיונות אחרים/נוספים). בנוסף עליכם לקחת בחשבון שאולי תצטרכו לייצג את המטריצות במבנה של מטריצות דלילות (חוסך הרבה מקום) על מנת שהזיכרון שלכם לא "יתפוצץ".

לכל אחד מהקבצים של עליכם לבנות ווקטור מייצג, כאשר אתם **שומרים את שיוך הווקטורים לקבוצה ולמספר/שם הקובץ**.

2) בעזרת word2vec או GloVe תבנו ווקטורים המסמכים (**ללא** סימני פיסוק ודברים נוספים כמו מרכאות מרכאה ..., מספרים, ספרות, תאריכים).

בעזרת word2vec או GloVe תבנו ווקטורים לבניית ווקטור המסמכים (**ללא** סימני פיסוק ודברים נוספים כמו מרכאות מרכאה ..., מספרים, ספרות, תאריכים - stop-words).

3) (Unsupervised Contrastive Learning (SimCSE (עיינו באינטרנט), יבוצע על **קבצי המקור** (לא הקבצים הנקיים ולא הלימות)

4) תשתמשו בווקטורים של Sentence-BERT (SBERT) **למסמכי המקור**

5) בסוף התקבלו:

2 קבוצות של מטריצות של TFIDF אחת **למילים** והשניה **ללמות**

(קבוצה אחת TFIDF-Word קבוצה שנייה TFIDF-Lemm)

2 קבוצות של מטריצות של W2V או GloVe אחת **למילים** והשניה **ללמות**

(קבוצה אחת W2V-Word או GloVe קבוצה שנייה W2V-Lemm או GloVe)

קבוצה של מטריצה אחת של SimCSE של **קבצי המקור**

(קבוצה אחת SimCSE-Origen)

קבוצה של מטריצה אחת של Sentence-BERT (SBERT) של **קבצי המקור**

(קבוצה אחת SBERT-Origen)

על כל אחת מהמטריצות של TFIDF (TFIDF-Word & TFIDF-Lemm) עליכם לציין את הערך המוסף של כל אחד מהמאפיינים (מילים או למות) ע"י שתי שיטות, אחת Information Gain ועוד אחת תחפשו באינטרנט, לבחירתכם. בדרכים כגון אלו אנו יכולים לקבל את החשיבות של כל אחד מהמאפיינים של כל אחת מהקבוצות.

אפשרויות בנוסף ל- Information Gain, לא הכרחי, מבחינתי אתם יכולים לבחור מדד אחר ממה שאני מציע כאן. כמובן עליכם לבחור מדד הרלוונטי לכם ולא אחד שאינו רלוונטי (למשל אם יש מדד בין זוג ווקטורים/תוצאות/מטריצות הוא אינו רלוונטי כאן, כי אנו רוצים מדד לכל "עמודה"/מאפיין במטריצה), אין הכרח שכל המופיע ברשימה הבאה רלוונטי.

Gain Ratio, Gini Impurity, Chi-squared statistic, ReliefF, MDL principle, Variance reduction, Correlation, Consistency, AUC-PR, Gini Index, Mutual Information.

התוצאה תהיה: קובץ אקסל שבו לכל מטריצה שתי טבלאות (אחת עם Information Gain והשניה עם המדד הנוסף שבחרתם, כמובן תצינו לכל טבלה מהו הממד) עם רשימת כל המאפיינים והחשיבות של כל אחד מהם.

עליכם לכתוב מסמך readme שיכיל בראשו את שמות הסטודנטים, ת.ז., ומספר קבוצת התרגיל. המסמך יציג הסברים על מה שעשיתם, את דרך העבודה שלכם, מדוע בחרתם בדרך שבחרתם. את התוצאות ונתוח של התוצאות והשגיאות (אם ישנן) ותובנות שהגעתם אליהן. לקובץ זה עליכם לצרף את כל הטבלאות עם שני המדדים, כאן מספיק 20 מאפיינים חשובים **לכל קבוצה** (כמובן גם לשייך את הטבלה למטריצה הרלוונטית). את קובץ ה-readme עליכם להעלות למודל

את כל הפתרונות עליכם להעלות למודל. הפתרון **חייב לכלול** את (1) הקוד בקובץ של פייתון (ולא קישור) את (2) מסמך ה-readme עם הסברים מפורטים עם טבלאות, תרשימים, פלוטים וכדומה, ואת (3) קובץ האקסל.

בנוסף, כל מטריצה עליכם לכווץ בזיפ בפני עצמה.

להכניס את קובץ ה-readme, את קובץ הקוד בפייתון, את קובץ האקסל ואת כל המטריצות לספריה אחת, שם הספריה יהיה שמות הסטודנטים שעשו את התרגיל ולכווץ בזיפ את כל הספריה. שם הספריה יהיה בשם הסטודנטים המגישים.