

הוראות כלליות: על כל תרגיל עלייכם לכתוב מסמך המציג הסברים על מה שעשיתם, את דרך העבודה שלכם, מדוע בחרתם בדרך שבחורתם. את התוצאות ונתוח של התוצאות והשגיאות (אם ישן).
את כל הפתוחנות עלייכם להעלות למודול. הפתרון **חייב לכלול** גם את הקוד עצמו בקובץ של פייתון (ולא קישור) ואת המסמך עם הסברים מפורטים עם טבלאות, תרשימים, פלוטים ועודמה.
על כל תרגיל להגיש תוך שלושה שבועות, ההגשה היא במודול ובמעבדות יתבצעו הגנות על התרגילים.

תרגיל 1

הפרלמנט הבריטי (יותר כללי):

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

קובצי ה- XML של הפרלמנט:

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

ישנם לא מעט קבצים, אני מעוניין בקבצים מ- 2023-06-28d.xml עד הסוף, בערך 935 קבצים.

עליכם "לנקות את הקבצים" (כל קובץ צריך לשומר על השם/המספר שלו) משמעו שכל מילה תעמוד בפני עצמה ללא שם תווים נוספים, הכוונה היא להפריד את המילה מסימן הפיסוק הסמור לה (יש פונקציה בפייתון שעשוה זאת לה) לדוגמה (בעברית):

1. **שיהיה לכם שבת שלום טוב ומברך** - המילה **ומברך** שונה מהמילה **ומברך** הנקודה מצינית סוף משפט. (התואאה היא **שיהיה לכם שבת שלום טוב ומברך**)
2. ... היום ראש הממשלה אמר **"אני** מצפה למשמעת **考艾茨ינית**" בהתייחסו להצעעה על ... – המילה **אני** שונה מהמילה **אני** וכן למילה **考艾茨ינית** (שימוש לב שהמחוזת **זה**"ל היא מחוץ לתקינה)
3. ועוד תווים מיוחדים נוספים כמו פסיק, נקודותים : נקודה-פסיק וכדומה.

לכל אחד מקובצי המקור תצרו קובץ עם הלמות (גזר המילים) של המילים שלו. לצורך זה עליכם להשתמש בכלים מתאימים, בדו"ח תכתבו באיזה כל השתמשתם, עדיף להשתמש בכלים הći טוב שתוכלו.

לכל אחד מהקבצים של עליכם לבנות וקטור מייצג, כאשר אתם שומרים את שיור הווקטורים לקובץ ולמספר/שם הקובץ.

(1) את הווקטורים תבנו באמצעות IDF-TF עם Okapi

הויל ורוב מוחלט של המסמכים אינם מכילים את כל המילים בשפה, וייתר מכך המסמכים הם ייחסית קצרים, אז אתם תקבלו מטריצות דילולות. עליכם לצמצם את המאפיינים על מנת שלא תתקבל מטריצה מאוד גדולה, למשל ניתן להוריד stop-words, מילים המופיעות פחות מ-5 פעמים ועודמה (אלו ההצעות אתכם יכולים לחשב על רעיונות אחרים/נוספים). בנוסף עליכם לקחת בחשבון שאלות תצרכו לייצג את המטריצות במבנה של מטריצות דילולות (חווסף הרבה מקום) על מנת שהזיכרון שלכם לא "יתפוץ".

לכל אחד מהקבצים של עליים לבנות ווקטור מייצג, כאשר אתם שומרים את שיווק הוקטורים לקובץ **ולמספר/שם הקובץ**.

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

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

(3) **הקבצים הנקיים ולא הלימות** (SimCSE) (עינוי באינטרנט), יבוצע על **קבצי המקור** (לא

(4) תשמשו בוקטורים של SBERT (Sentence-BERT) **למסמכים המקוריים**

(5) בסוף התקבלו:

2 קבוצות של מטריצות של TFIDF אחת **למילים** והשנייה **לلمות**

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

2 קבוצות של מטריצות של W2V או GloVe אחת **למילים** והשנייה **לلمות**

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

קובוצה של מטריצה אחת של SimCSE של **קבצי המקור**

(קובוצה אחת (SimCSE-Origin)

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

(קובוצה אחת (SBERT-Origin))

על כל אחת מהמטריצות של TFIDF-Lemm & TFIDF-Word (TFIDF-Word & TFIDF) עליכם לציין את הערך המוסף של כל אחד מהמאפיינים (מילים או למota) ע"י שתי שיטות, אחת Gain Information ועוד אחת תחפושו באינטרנט, לבחירתכם. בדרך כלל אנו יכולים לקבל את החישבות של כל אחד מהמאפיינים של כל אחת מהקבוצות.

אפשרויות נוספות - Gain Ratio, לא הכרחי, מבחינתי אתם יכולים לבחור מدد אחר ממה שאנו מציע כאן. כਮון עליכם לבחור ממד הרלוונטי לכם ולא אחד שאתה יכול להשיג (למשל אם יש ממד בין זוג ווקטורים/מטריצות הוא אינו רלוונטי אז, כי אם רציתם ממד לפחות "עמודה" / מאפיין במטריצה, אין הכרח שכל המופיע ברשימה הבא 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 (אחד עם Gain Information והשנייה עם הממד הנוסף שבחרתם, כਮון תציגו לכל טבלה מהו הממד) עם רשימת **כל המאפיינים** והחישבות של כל אחד מהם.

עליכם לנתח מסמך `sumpme` שיכיל בראשו את שמות הסטודנטים, ת.ז., ומספר קבוצת התרגיל. המסמך יציג הסברים על מה שעשיתם, את דרך העבודה שלכם, מודיע בחירתם בדרך שבחורתם. את התוצאות ניתוח של התוצאות והשגיאות (אם ישן) ותובנות שהגעתם אליהן. לקובץ זה עליכם לצרף את כל התרגילים עם שני חנידים, וכן סכין 20 מאפיינים חשובות **לכל קבוצה** (כמובן גם לשני את הטעלה **למטריצה הרילונטית**). את קובץ ה- `readme` עליכם להעלות למודול.

את כל הפתוחות עליכם להעלות למודול. הפתוח **חייב לכלול** את (1) הקוד בקובץ של פיתון (ולא קישור) את (2) מסמך ה- `sumpme` עם הסברים מפורטים עם טבלאות, תרשימים, קלוטים וכדומה, ואת (3) קובץ **האקסל**.

בנוסף, כל מטריצה עליכם לכונן בדף בפני עצמה.

לחכמי את קובץ ה- `readme`, את קובץ הקוד בכיתון, את קובץ האקסל ואת כל המטריצות לספרה אחת שם הספרה יהיה שמו של הסטודנטים שעשו את התרגיל ולכונן בדף את כל הספרה. שם הספרה יהיה **כשם הסטודנטים המוגשים**.