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.

**תרגיל 2**

בתרגיל 1 היו נתונים לנו 4 עיתונים שונים עם כתבות, לכל עיתון יצרנו מטריצה מייצגת וקיבלנו 4 מטריצות שונות שכל אחת מייצגת עיתון ספציפי. את המטריצות בנינו בשיטות שונות ולכל שיטה יש את 4 המטריצות שלה, **לכל שיטה של ייצוג הכתבות אני קורא "קבוצה"**.

חלוקה לנושאים/מחלקות בצורה מפוקחת ושאינה מפוקחת

1. עליכם לבנות אשכולות ולהציג תוצאות מספרית ע"י precision, recall, F1 & accuracy ותוצאות ויזואליות למשל ע"י שימוש ב- UMAP או ב- t-SNE  (אלו המומלצים, לא חייבים להשתמש בזה, אתם יכולים להשתמש גם בכלים אחרים, ניתן למצוא דוגמאות ב- <https://seaborn.pydata.org>).

את האשכולות תבנו ע"י k-means ע"י DBSCAN וע"י Mixture of Gaussian. (למשל על הווקטורים של שיטה אחת, ז"א 4 מטריצות, כל אחת מייצגת עיתון אחר, תעשו k-means כמובן עם k=4 כי מדובר על 4 עיתונים-שונים, לעיתים אפשר להשתמש בנתון זה, במידת הצורך, גם בשביל Mixture of Gaussian. **התוצאות אינן חייבות להראות שיש הבדל בין העיתונים**. חשוב לציין, שעל מנת ליצור את האשכולות מאחדים אל כל 4 המטריצות למטריצה אחת, שזו למעשה קבוצה)

תחשבו כיצד להעריך את התוצאות, שהרי תהליך האישכול אינו מתייחס כלל לתיוג הכתבות/הווקטורים.

אישכול בעזרת DBSCAN מצריך שני פרמטרים חיצוניים, מרחק (eps) ומספר מינימלי של איברים באשכול (min\_samples). ישנן היוריסטיקות שונות לבחור את שני פרמטרים אלו. למשל לעשות MST ולבחור את הקשת הגדולה ביותר והיא תהיה המרחק, לגבי מספר נקודות מינימלי, תעשו בדיקה. אתם מוזמנים להשתמש בהיוריסטיקות אחרות, כל עוד תציינו את ההיוריסטיקה בה השתמשתם.

בכל אחת משיטות האישכול על כל הייפר-פרמטר עליכם לכתוב מה בחרתם ומדוע בחרתם בזה.

דוגמאות להרצות של קבוצה של אלגוריתמים שונים של אישכול.

<https://machinelearningmastery.com/clustering-algorithms-with-python/>

**חשוב לציין שבתחום שלנו מרחק הוא דמיון-קוסינוס.**

1. בהנתן 5 השיטות:
   1. Artificial Neural Network (ANN)
   2. Naive Bayes (NB)
   3. Support Vector Machine (SVM)
   4. Logistic Regression (LoR)
   5. Random Forest (RF)

**עליכם להשתמש ב- ANN כמפורט בהמשך.**

עליכם ליצור מודל המסווג את כל אחת מהקבוצות.

את המסווג תריצו בתצורה של 10 fold cross validation, הסבר קל:

<https://www.youtube.com/watch?v=fSytzGwwBVw>

דוגמה לקוד:

[**https://scikit-learn.org/stable/modules/cross\_validation.html**](https://scikit-learn.org/stable/modules/cross_validation.html)

**התוצאות אינן חייבות להראות שיש הבדל בין הקבוצות**

בשיטות: NB, RF, SVM, LoR עליכם לציין מהם 20 המאפיינים החשובים לכל קבוצה בהתאם לשיטה, יש שיטות שנותנות משקלים חיוביים לקבוצה אחת ומשקלים שליליים לקבוצה השנייה, ככל שהערך המוחלט של המשקולות גבוהה יותר החשיבות של המאפיין גדולה יותר. ולתת הסברים הגיוניים במסמך ה- readme למאפיינים אלו, אם יש. את רשימת המאפיינים לכל אחת מהקבוצות עליכם להעלות לקובץ אקסל (מיון ופעולות אחרות, כגון תרשים [לפעמים יש נקודה/ברך שבה פתאום המאפיינים מקבלים משקל מאוד נמוך] יכולות לעזור לכם לתובנות נוספות). חשוב לבדוק אלו מאפיינים חשובים שהם משותפים לכל שיטות הלמידה ואלו מאפיינם חשובים לרוב שיטות הלמידה.

לצורך NB אתם יכולים (יכולים, לא חובה) להיעזר ב:

<https://www.datacamp.com/tutorial/naive-bayes-scikit-learn>

<https://scikit-learn.org/stable/modules/naive_bayes.html>

לצורך LoR אתם יכולים (יכולים, לא חובה) להיעזר ב:

<https://www.machinelearningnuggets.com/logistic-regression/>

<https://www.w3schools.com/python/python_ml_logistic_regression.asp>

<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

<https://www.geeksforgeeks.org/ml-logistic-regression-using-python/>

לצורך SVM אתם יכולים (יכולים, לא חובה) להיעזר ב:

<https://randomresearchai.medium.com/svc-model-in-python-2d7b6d9434b4>

<https://machinelearningmastery.com/cost-sensitive-svm-for-imbalanced-classification/>

לצורך RF אתם יכולים (יכולים, לא חובה) להיעזר ב:

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

<https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>

<https://www.datacamp.com/tutorial/random-forests-classifier-python>

לצורך ANN אתם יכולים (יכולים, לא חובה) להיעזר ב:

<https://machinelearningmastery.com/how-to-make-classification-and-regression-predictions-for-deep-learning-models-in-keras/>

<https://machinelearningmastery.com/category/deep-learning/>

**פירוט על ה- ANN (כאן אין צורך להשתמש ב- cross validation):**

עליכם כל קבוצה, כמו בסעיף א', ולהריץ רשת נוירונים ב- keras ולהציג תוצאות כמו בסעיף א'.

בהמשך יופיעו לכם שתי טופולוגיות מעט שונות של ה- ANN, ההבדל בין הטופולוגיות הוא רק בפונקציית האקטיבציה של כל שכבה.

החלוקה של הדוגמאות (בצורה אקראית): למידה 80%, ולידציה (מתוך הלמידה) 10%, בחינה 20%. למשל נניח שישנם 1000 מסמכים, 800 מסמכים ללמידה, 200 לבחינה. מתוך ה- 800 ללמידה 10% לוולידציה, ז"א 720 ללמידה, 80 לוולידציה ו- 200 לבחינה.

Epoch (איטרציות) – מקסימום 15.

Batch Size – 32.

Callbacks –

* EarlyStop – עצירת לימוד המכונה כאשר אין התקדמות במדד ה- Accuracy של הוולידציה בתוך שלוש איטרציות רצופות.
* ModelCheckpoint – שומר את המודל שהשיג את הדיוק (Accuracy) הטוב ביותר על הוולידציה.

טופולוגיית רשת ה- ANN הראשונה:

* שכבה ראשונה Embedding layer מקבלת כקלט ווקטורים
* שכבה שנייה Hidden layer עם פונקציית relu –10 קודקודים
* שכבה שלישית Hidden layer עם פונקציית relu –10 קודקודים
* שכבה רביעית Hidden layer עם פונקציית relu –7 קודקודים
* שכבה אחרונה Activation layer עם פונקציית softmax

טופולוגיית רשת ה- ANN השנייה:

* שכבה ראשונה Embedding layer מקבלת כקלט ווקטורים
* שכבה שנייה Hidden layer עם פונקציית GELU –10 קודקודים
* שכבה שלישית Hidden layer עם פונקציית GELU –10 קודקודים
* שכבה רביעית Hidden layer עם פונקציית GELU –7 קודקודים
* שכבה אחרונה Activation layer עם פונקציית softmax

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 (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) ואת (2) מסמך ה- readme עם הסברים מפורטים עם טבלאות, תרשימים, פלוטים וכדומה, ואת (3) קובץ האקסל עם המשקולות.

בנוסף, עליכם להכניס את קובץ הקוד, קובץ ה- readme ואת קובץ האקסל, שם הספריה יהיה שמות הסטודנטים שעשו את התרגיל ולכווץ בזיפ את כל הספריה. וגם את זה להעלות למודל, סה"כ 4 קבצים, את שלושת הקבצים "הפתוחים" ואת קובץ הזיפ.