תרגיל אמצע בקורס היבטים מעשיים ב- AI ו- NLP

<u>הנחיות כלליות לתרגילי הבית:</u>

- הגשת התרגילים תתבצע באמצעות תא ההגשה במאמא.
- יש להגיש אך ורק מחברת Jupyter (סיומת ipynb), הגשה בצורה אחרת לא תתקבל.
- המחברת חייבת לרוץ ב- Conda פייתון 3.9.1 ומעלה בלבד.
- המחברת חייבת לרוץ בלחיצת כפתור ריצה Run All.
- יש לשים לב לכתוב את הקוד בצורה יפה וכן לשלב הערות במחברת ולהסביר חלקים חשובים.
- הגשה **בזוגות** בלבד.
- תרגיל אשר יוגש באיחור לא ייבדק (למעט אישורים מיוחדים כגון: מילואים, אשפוז וכו').
- לכל שאלה אתם מוזמנים לפנות אלינו במייל.
- הציון של התרגיל יהיה 25% מהציון הסופי.

<u>התרגיל</u>

בתרגיל הבא אנו רוצים שתיישמו את החומר הנלמד בקורס בעזרת תכנות בפייתון של פתרון סיווג טקסטים בעזרת scikit-learn (ו-nltk).
הדגש הוא על טעינה של דוגמאות טקסט, שליפה של פיצ'רים, בניית מודל, ביצוע prediction והערכת הדיוק.

להלן התרגיל:

1. במאמא מקושר לשיעור קובץ דוגמאות של אימיילים המסווגים בחלקם כספאם וחלקם כלא ספאם (ham).
2. כתבו מחברת הטוענת את קובץ הדוגמאות, ומבצעת טוקניזציה.
3. בצעו במחברת EDA הדוגם את המידע ומראה שכיחות של טוקנים לכל קטגוריה ולהציג אותה בטבלה או בגרף\תרשים.
4. בצעו במחברת שליפה של פיצ'רים
5. בנו מודל סיווג עם אלגוריתם מספריית scikit-learn.
6. בצעו במחברת הערכת דיוק למודל ע"י שליפה של מדדים כמו accuracy, precision, recall, f1 מריצה שלו על סט בדיקה מתויג.
7. חזרו לסעיף 4 ובצעו לפחות עוד 3 ריצות נוספות עם שליפה שונה של פיצ'רים ו/או בחירת אלגוריתם אחר.
8. סדרו בטבלה המציגה כל אחת מהריצות שעשיתם ומשווה בין התוצאות שקיבלתם.
9. התייחסו לריצה שנתנה את ה"דיוק" הגבוה ביותר ונסו לנסח סיבות מדוע זו הריצה ה"מדויקת" ביותר.