מטלת בית מס' 3

הנחיות:

הגשה:

- יש להגיש את המטלה עד לתאריך 15.01.2020 בשעה 23:55.
- ההגשה תתבצע ע"י קובץ zip המכיל את הקוד ומסמך PDF שנכתב בהתאם לדרישות. פורמט קובץ ה-zip יהיה ID01_ID02.zip. שם קובץ ההרצה יהיה Lab3.py.
- ההגשה הינה **בזוגות**.
- חבר קבוצה אחד בלבד יעלה את הפתרון לאתר.
- בעיות אישיות בנוגע למועד ההגשה יש להפנות לבודק התרגילים הקורס טרם מועד ההגשה.
- **כל חריגה מנהלים אלו, ללא אישור בכתב מצוות הקורס, מהווה עילה לפסילת המטלה או להפחתת נקודות.**
- **אין להעתיק פתרונות ואין לשתף קוד בין סטודנטים. אין להעתיק קוד מוכן באינטרנט!**
- **במודל שתיצרו אין להסתמך על שום מידע חיצוני לרבות ציוצים מהרשת החברתית של Twitter.**
- **לפתרון המטלה יש להשתמש בגרסאת פייתון 3.x.**
- להבהרות, הכוונות או כל עזרה אחרת, ניתן לשאול שאלות בפורום המתאים למטלה זו באתר הקורס.
- בדיקת המטלה תתייחס בין השאר לפרמטרים הבאים: נכונות הקוד, יעילות הקוד וזמני ריצה. יש לבדוק מקרי קצה.

איחורים:

- איחור ממועד ההגשה המקורי עד 24 שעות – יופחת ניקוד של 10 נקודות מציון העבודה.
- איחור החל מ-24 שעות עד 48 שעות מתאריך ההגשה המקורי – יופחת ניקוד של 20 נקודות מציון העבודה.
- לאחר 48 שעות לא תתאפשר הגשה של המטלה.

תקינות הקבצים:

- באחריות המגישים לוודא כי הקוד רץ במחשבי המעבדה טרם ביצוע ההגשה. העבודות ייבדקו במחשבי המעבדה בסביבת Anaconda 3.
- יש לציין בקובץ Readme.txt את כל החבילות שהשתמשתם בהם אשר דורשות התקנה מיוחדת, דהיינו חבילות שאינן נמצאות ב-Anaconda 3.
- מטלה שתכשל בזמן ריצה על גבי מחשבי המעבדה מכל סיבה שהיא לא תיבדק.

בהצלחה! ☺

מבוא

במטלה זו עליכם לפתור משימת Sentiment Analysis, בהתבסס על מידע מהרשת החברתית Twitter.

מאגר מידע – Dataset :

מאגר המידע מכיל מספר רב של ציוצים, מהתצורה הבאה :

SentimentText	Sentiment
is so sad for my APL friend	0

- Sentiment – האם הציוץ הינו חיובי (1) או שלילי (0).
- Sentiment Text – הטקסט של הציוץ.

במטלה תתמקדו בתהליך ה-Preprocessing, Feature Extraction, Classifier וכן בחלוקה של הנתונים לסט אימון וסט בחינה. עליכם לתאר את חלק א' במסמך PDF ולהגיש אותו כחלק מפתרון מטלה זו.

במטלה זו חלק מהציון יורכב על סמך הערכות של המודל שלכם על סט בחינה שאינו חשוף לכם.

הקובץ מופיע באתר התחרות - Kaggle (ראה הסבר בהמשך הדו"ח).

תהליך יצירת מודל הסיווג:

בחלק זה עליכם לממש את השלבים הבאים :

1. Preprocessing – עיבוד מקדים – עליכם להפעיל תהליך שמבצע עיבוד מקדים של הטקסט. ניתן ואף רצוי להיעזר בחבילות חיצוניות בשלב זה. עליכם להסביר איזה צעדים ביצעתם בשלב זה ומדוע בחרתם דווקא בצעדים אלו. מומלץ ראשית לפתוח את קובץ הציוצים ו"ללכלך את הידיים" במידע.

2. Feature Extraction – בשלב זה עליכם לבנות אוסף של פיצ'רים אשר יוזנו בשלב הבא למסווג. עליכם לפרט את כל הפיצ'רים בהם בחרתם ואת המוטיבציה לבחירה בהם. בצעו בחירה מושכלת וחכמה בשלב זה.

3. Classifier – בשלב זה עליכם להשתמש בטכניקת 10 Cross Validation בשביל להעריך את המדדים השונים. יש לבדוק לפחות שלושה מסווגים שונים (לדוגמה : Naïve Bayes, Decision Tree, Logistic Regression). עליכם לדווח לכל הפחות על מדדי Accuracy, Recall ו- Precision עבור כל אחד מהמסווגים שבחרתם על גבי ה-Train (אשר מצורף בתחרות). בסוף הדו"ח עליכם לדווח על בחירה במסווג אחד נבחר ואת השיקולים בבחירה שלכם. בשלב זה השתמשו ב-scikit-learn .

תחרות – Kaggle:

- עליכם להירשם לפלטפורמה של Kaggle – מספיק חבר צוות אחד.
- עליכם להיכנס לתחרות בלינק הבא.
- עליכם לפרט בדו"ח אותו אתם מגישים למודל, את שם הקבוצה שבחרתם לתחרות (לכל זוג מגישים יש קבוצה אחת בתחרות).
- לכל קבוצה מותר להגיש 5 קבצים שונים ביום.
- שימו לב כי ישנם שני קבצים שונים לתחרות:
  - קובץ ה-Train – קובץ שחשוף לכולם עם טקסט של ציוצים (שדה – SentimentText) והסיווג שלהם (שדה – labels). על קובץ זה יש לאמן מודלים שונים.
  - קובץ ה-Test – לאחר שבחרתם מודל אחד עליכם לבצע חיזוי על טקסטים שונים שנמצאים בקובץ זה. הקובץ אינו מכיל labels. הסיווגים חייבים להיות לפי סדר השורות עליהן חזיתם סיווגים. כאשר התוצאה של החיזוי צריכה להיות 0 או 1 (שדה - Sentiment) על סמך מספר שורה (ID). ראו קובץ sample.csv עבור דוגמה בפורמט בה אתם צריכים להעלות את קובץ הדירוגים שלכם.

בהצלחה! ☺