מערכת: **המכון הטכנולוגי - חולון**
קורס: **למידת מכונה (למידת מכונה (תשפא_ב_63302) - פרידמן)**
אגרון מונחים: **מטלה תחרותית - שאלות ותשובות**

ה

המלצה 1 - שלב התחלתי והנחיות/המלצות כלליות

שלב התחלתי והנחיות/המלצות כלליות:

1. לראות שאתם מצליחים להוריד את כל הקבצים מרכיב המטלה במודל.
2. קראו בעיון את כל ההוראות שבמסמך ההסבר (*מטלה_תחרותית.pdf*)
3. עקבו אחר כל השאלות והתשובות ברכיב השאלות והתשובות בדף הקורס במודל, שנמצא תחת המטלה במודל (שם הרכיב: *מטלה תחרותית - שאלות ותשובות*)
4. נסו לטעון את קבצי הקלט (מי שיש להם בעיות בטעינת קבצי הקלט, יש על כך עוד שאלה ותשובה).
5. ודעו שטעינת קבצי הקלט תשאר כך גם בהגשה (מכיוון שאנחנו מצפים שקבצי ה-excel יהיו בספריה 'input' המקבילה למחברת המטלה, בהרצתה אצלינו).
6. ודאו שאתם מבינים איך אמור להראות קובץ ה-csv אותו אתם מגישים.
7. לגבי שמירת קובץ ה-csv (קובץ הפלט) בסוף המטלה, השתמשו בפקודה מאין זו, כך ששם הקובץ לא ישתנה, והוא ישמר במקביל למטלה (כך גם נניח בבדיקות).
8. לגבי שימוש במודולים/חבילות אחרות ראו תשובה, יש על כך עוד שאלה ותשובה, קראו שם הסבר.
9. מומלץ לקרוא כעת את התשובה *המלצה 2 - מה עושים אם לא מבינים בכלל איך נגשים לבעיה*

המלצה 2 - מה עושים אם לא מבינים בכלל איך נגשים לבעיה

שאלה:

מה עושים אם לא מבינים בכלל איך נגשים לבעיה?

תשובה:

1. קראו קודם את התשובה ל- *המלצה 1 - שלב התחלתי והנחיות/המלצות כלליות*
2. עברו על מצגת ההרצאה בנושא *ניתוח טקסט* וצפו בהקלטת השיעור שוב.
3. מומלץ למי שצריכים עזרה ב-python, להסתכל על החומרים והקישורים שהוספנו (תחת *חומרי עזר ותרגול נוספים (רובם חיצוניים)*, בדף הקורס במודל) וכמובן לעבור על מצגות ה-jupyter בנושא (תחת *python - מחברות מהתרגול וחומרים נוספים*, בדף הקורס במודל), ולצפות בהקלטות התרגולים שוב.
4. עברו על כל החומרים שהעברנו, הנוגעים לפתרון (המחברות החל מתרגול 7 תחת *python - מחברות מהתרגול וחומרים נוספים*), צפו בהקלטות ההסבר ועברו על הקישורים והחומרים הרלוונטים בהמשך (בסוף האזור של *חומרי עזר ותרגול נוספים (רובם חיצוניים)*)
5. שימו לב למעוניינים ישנה אפשרות להשתמש ב-stopwords, וב-tokenizer, שנמצאים תחת *כלים לניתוח טקסט בעברית*, בדף הקורס במודל, הקשיבו גם להקלטות ההסבר. שימו לב - ישנה גם שאלה ותשובה יעודית בנושא כלים בעברית.
6. שימו לב (למעוניינים) גם לתשובה בנוגע לשילוב מודולים וחבילות עליהם לא עברנו.
7. קראו כעת את התשובה *המלצה 3 - כיצד מומלץ לגשת לפתרון המטלה*

המלצה 3 - כיצד מומלץ לגשת לפתרון המטלה

שאלה:

כיצד מומלץ לגשת לפתרון המטלה?

תשובה:

<u>שלב ראשון - הגעה לסיווג בסיסי</u>:

1. ראשית מומלץ לקרוא את התשובה *המלצה 1 - שלב התחלתי והנחיות/המלצות כלליות* ואת התשובה *המלצה 2 - מה עושים אם לא מבינים בכלל איך נגשים לבעיה*
2. למניעת בעיות מומלץ לוודא שיש לכם גרסה מעודכנת של sklearn (לא לשכוח לשמור את מחברת המטלה לפני ולעשות restart kernel אחרי ההתקנה) ולהתקין בסביבה הוירטואלית שלכם, על ידי הרצת הפקודה:
pip install -U scikit-learn
3. מומלץ לראות שאתם מסתדרים עם הכלים המומלצים בעיקר ב-sklearn, ולהצליח להפעיל למשל vectorizer, גם שליטה ב-pipeline יכולה להקל.
4. השאירו את העבודה עם מחרוזות (strings) ו-regular expressions להמשך
5. אם צריך שימו לב שיש מחברות תרגול עם תשובות לחלק מהנושאים וישנם גם קישורים נוספים בדף הקורס.
6. שימו לב שהקטגוריות מופיעות כ-string (האפשרויות: 'f' או 'm'), אך המסווגים מניחים שהם מספריים (ולכן יש לתרגמם למספרים קבועים הן אחרי הטעינה והן לפני הגשת קובץ ה-csv)
7. בחרו אחד מה-vectorizers ואחד מאלגוריתמי הלמידה והריצו אימון פשוט ללא משחק אם שום פרמטר וסיווג הדוגמאות מה-test.

אם הגעתם עד פה עברתם שלב ראשון בהצלחה.

<u>שלב שני - יכולת למדוד את ביצועי המערכת</u>:

1. לאחר סיום מוצלח של השלב הראשון לעיל, נסו למדוד את ביצועיכם על ידי שימוש ב-metrics (תת מודול של skleran) שחלקם נמצאים במחברות התרגול ובנוסף ב-f1_score, accuracy_score וכדו'.
2. תוכלו למדוד את ביצועי הסיווג על דוגמאות ה-*train*.

שימו לב - כמובן שמדידת ביצועים על דוגמאות ה-*train*

3. השתמשו בשיטות שלמדנו הקשורות ל-validation (חלקם מודגמות במחברת על sklearn)

<u>שלב שלישי - יכולת להשתפר</u>

1. השוו אלגוריתמי למידה שונים (בנוסף למה שמופיע במחברות התרגול, למדנו גם אלגוריתמים נוספים בהם המחלקות המתאימות ב-sklearn כוללות בין השאר: LinearSVC (אחת האפשרויות ל-SVM), Perceptron, MLPClassifier (רשתות נוירונים).
2. נסו להשוות פרמטרים שונים בשלבים השונים הבסיסיים.
3. דגש חשוב - כל פרמטר, מאפיין או כלי יכול לשפר ביצועים, אך גם לפגוע בביצועים, כדאי לשים לב בניסויים שלכם.

<u>שלב רביעי - שימוש בכלים לעברית ובכלים חיצוניים</u>

1. אתם מוזמנים לשלב tokenizer, stop words שהעלנו למודל (הסבר נפרד *שילוב כלים בעברית*).
2. אתם מוזמנים להשתמש בכלים חלופיים או לוותר לגמרי על כלים אלו.
3. אתם מוזמנים להשתמש בכלים אחרים, לפי מה שמופיע בתשובה הנפרדת בנושא.

<u>שלב חמישי - מתקדם יותר - מניפולציה של המאפיינים</u>

1. מומלץ לעבור על מחברות התרגול בנושא מחרוזות ו-regular expressions
2. אפשר להשתמש במחברות שמופיעות במודל לתרגול אישי בנושאים אלו.
3. נסו נירמולים ושינויים שונים במאפיינים ושילובים שונים עליהם דברנו
4. נסו מאפיינים מתקדמים עליהם לא למדנו.

<u>הערות כלליות לסיכום</u>:

- לא לשכוח להשתמש בשיטות שלמדנו להערכת ביצועים (או בשיטות אחרות שאתם מכירים).
- לא כל דבר שמוסיפים עוזר, ולפעמים הוא מוריד את הביצועים.
- לא לשכוח לעבור על הדברים הבסיסיים שהוגדרו במטלה (חבל שתעשו הרבה עבודה וירדו נקודות על שטויות).
- לא חייבים פתרון סופר מורכב כדי לקבל ציון טוב

ולבסוף בהצלחה לכולם

מודולים/חבילות/אלגוריתמים שמותר להשתמש בהן בפתרון

שאלה:
באילו חבילות וכלים מותר להשתמש בפתרון?
האם מותר להשתמש במודולים/חבילות/אלגוריתמים שלא הראינו?

תשובה:
כל מה שכתוב פה מתייחס מודולים/חבילות/אלגוריתמים (גם אם רק חבילה למשל מוזכרת):
אין צורך לשאול על כל מודול, חבילה או אלגוריתם, אלא לנקוט לפי הכללים הבאים:
1. הנחיה כללית:
- אפשר להשתמש בכל מודול בתנאי שזה לא פותר את הבעיה ישירות.

2. התקנות דרושות והוראות התקנה:
- צריך לצרף הוראות מדויקות של התקנת כל חבילה מדוברת, כולל כל הגרסאות של מה שהותקן.
- אפשר להשתמש בהתקנות שללא כוללות כוללות קבצי הרצה
- ההתקנה הדרושה צריכה לעבוד בכל מערכת הפעלה (multi platform)
- המודול צריך להיות כלי סטנדרטי, שניתן להתקין ע"י pip install, conda install וכדו' ולא משהו שפשוט צריך להוריד מאיזה אתר.
- יש לציין את הגרסאות המדויקות של החבילה אותה התקנתם.
- את ההוראות יש לצרף בקובץ txt נפרד (נאפשר הגשה של כמה קבצי text לצרכים אלו) ולידע את המתרגל שכך עשיתם.
- כל זאת כדי שאפשר יהיה להריץ את הקוד אצלינו, ולשחזר את התוצאות (חלק מהבדיקה)

3. תיעוד:
- צריך לצרף במחברת המטלה הסבר על כל כלי/אלגוריתם שמשתמשים בו, בעיקר אם לא למדנו אותו.
- צריך להסביר על השימוש בו והמחלקות בהם משתמשים.
- כמו בכל הקוד גם פה יש להראות הבנה בקוד ולהראות שהקוד לא סתם הודבק ממקום אחר, אלא יש הבנה במה שמשתמשים.
- הכוונה להסברים קצרים, ולא לmanuals ארוכים, הסבר תמציתי וקצר.

ק

קבצי הקלט - ישנה שגיאה בקריאת הקבצים מהמחברת

תאור המצב והשאלה:
נכנסים למחברת המטלה ומנסים להריץ את תא הקוד של קריאת קבצי ה-excel, אך ההרצה נכשלת ונזרק exception, מה עושים?

תשובה:
אם הבעיה היא בגלל שה-jupyter לא מוצא את הקבצים (FileNotFoundError):
1. לוודא שהורדתם את קבצי ה-excel בנוסף למחברת המטלה (2 קבצי excel כמתואר במסמך הסבר המטלה).
2. לוודא שיצרתם במקביל למטלה ספריה בשם 'input' ושם שמתם את קבצי ה-excel בספריה.

הטעינה יכולה גם להכשל בעקבות חוסר בחלק מההתקנות:
- לפני ההתקנות הדרושות לשמור את המטלה
- שימו לב לתחתית דף הקורס במודל, תחת הכותרת 'התקנת סביבת עבודה - python, Anaconda, Jupyter' ישנם מסמכי הוראות ההתקנה של סביבת Anaconda הכוללות גם את ההתקנות של ה-modules (כלומר התקנה של החבילות התלויות).
- לאחר כל התקנה כזו, צריך ללחוץ על ה-icon של restart kernel
או תחת תפריט kernel לבחור Restart
- לאחר שעשיתם restart ל-kernel לא לשכוח שיש להריץ את כל תאי הקוד מההתחלה

קבצים שצריכים להגיש במטלה

שאלה:
איזה קבצים צריכים להגיש במטלה?

תשובה:
הקבצים הבסיסיים - כפי שמוסבר במסמך הסבר המטלה (מטלה_תחרותית.pdf):
את מחברת המטלה: competitive_assignment.ipynb
ואת קובץ התוצאות: classification_results.csv

אם הוספתם קובץ עם מילים (כמו stopwords), צריך להניח שמדובר בקובץ טקסט ושהוא מופיע במקביל למחברת המטלה (במחינת הקוד של מחברת המטלה) באותה הספריה.
- בנוסף הקובץ חייב להיות קובץ טקסט עם הסיומת txt או text בלבד
ראו עוד הסברים בתשובת *שילוב כלים בעברית*.
* נא ליידע את המתרגל שלכם אם הגשתם קבצים נוספים

אם השתמשם בכלים חיצוניים, עליכם להוסיף הוראות (בשאיפה פקודות בלבד) של ההתקנה הסטנדרטית שלהם (pip install או conda install) ראו עוד בתשובה על שילוב מודולים/חבילות חיצוניות.
- גם פה, הקובץ חייב להיות קובץ טקסט עם הסיומת txt או text בלבד.
* נא ליידע את המתרגל שלכם אם הגשתם קבצים נוספים

ש

שילוב כלים בעברית

שאלה:
איך ניתן לשלב כלים עליהם למדנו בעברית?

תשובה:
הכלים שלמדנו נמצאים תחת *כלים לניתוח טקסט בעברית*, בדף הקורס במודל.
העלנו למודל שני "כלים" בעברית:
tokenizer לעברית:
- מזכירים ש-tokenizer מפרק text למילים (tokens).
- התקנת הכלי ששלחנו בסביבה הוירטואלית שלכם:
pip install hebrew_tokenizer
(לא לשכוח לעשות restart kernel אחרי ההתקנה)
- הדגמת שימוש תוכלו למצוא בקובץ בעל הכותרת *הדגמת שימוש ב-tokenizer בעברית* (מדובר בקובץ text, ניתן להוריד אותו למחשב, ע"י right-click ואז בחירת save as מהמודל).
* הקשיבו להסברים בתרגולים האחרונים.
* שימו לב, כפי שמופיע ע"מ להחליף את ה-tokenizer של ברירת המחדל ב-vectorizers ב-sklearn, יש לעטוף את הפונקציה, כך שעבור string המייצג טקסט, הפונ' תחזיר רשימה (list) של tokens (כלומר list של מחרוזות).
* אם למשל ממשתם פונקציה כנ"ל שנקראת my_tokenize, יש לקבוע את הפרמטר tokenizer ב-vectorizer, באתחול של אובייקט הvectorizer כך:
tokenizer=my_tokenize

stopwords לעברית:
- העלנו אפשרות אחת של stopwords בעברית תחת
- תוכלו להוריד את הקובץ בעל הכתרת *stop words בעברית* (מדובר בקובץ text, ניתן להוריד אותו למחשב, ע"י right-click ואז בחירת save as מהמודל).
* שימו לב, אין stop words לעברית בברירת המחדל ב-sklearn. אם רוצים להוסיף אותם צריך לקבוע את הפרמטר stop_words של ה-vectorizers ב-sklearn. עליכם לקבוע אותו כlist של stopwords (כלומר list של מחרוזות).
- אפשר לטעון את קובץ ה-stopwords (כקובץ text בלבד) ואז להפוך אותו ל-list. אם כך תחליטו, אז צרפו (בהגשה) את קובץ ה-text והניחו שהוא ממוקם במקביל למטלה
- אם כך עשיתם, שימו לב שסיומת הקובץ היא txt או text בלבד.
- אפשרות נוספת (קצת פחות יפה), היא להוסיף את ה-stopwords לקוד של מחברת המטלה.

מה לגבי "כלים" חלופיים בעברית:
- ניתן להשתמש ב-stop words, או כל רשימה אחרת של מילים או ב-tokenizer אחר ובלבד שהשימוש הוא כפי שכתבנו בתשובה לגבי שימוש בחבילות/מודולים חיצוניים.
- ניתן להשתמש גם בכלים אחרים, לפי ההוראות בתשובה לגבי שימוש בחבילות/מודולים חיצוניים.