**פרויקט מסכם: היבטים מעשיים בכריית תוכן**

**ליזה פרידמן 316871474 ----ליאור פפר 203202122**

**מבט על**

אל הפרויקט ניגשנו עם ידע תיאורטי בסיסי וידע מעשי לוקה בחסר. מרבית משאבי הזמן הוקצו לחקר ולמידת סביבת הפיתוח והספריות הקיימות, משמעותן והבנת API לעומק.

**שאלה 1**

1. **שאלה**

Download the dataset. Perform text pre-processing. Present data exploration: class distribution, terms frequency for the different genders.

1. **פתרון**
2. **קשיים**

* העברת הקוד ל Jupyter לאחר פיתוח ב pycharm וחלוקה נכונה ונוחה לקטעי קוד.
* הצורך בריענון השימוש בשפת פיתון, למידת ביצוע לולאות FOR יעילות במקום FOR לא יעילות.
* למידת הנתונים ומימוש הקונספט לפתרון, הבנת משמעות ערכי NULL והחלפתם, התאמת ביטויים רגולריים.
* מצאנו כי ב panda , בעת ייבוא קובץ לתוך dataframe ניתן להשתמש בקידוד 'latin-1' כאשר משתמשים במידע שהגיע מהאינטרנט [דרך פרוטוקול HTTP] ולקבל ביצועים טובים יותר כיוון שהמידע המקודד קצר יותר.

**סיכום**

* מצאנו כי השימוש ב dataframe.loc מהיר משמעותית מביצוע איטרציות על הDF.
* השאלה רעננה את הידע שלנו בפיתון ובעיקר הכינה אותנו לקראת השאלה השנייה.
* לאחר ההתמודדות עם הקשיים הראשוניים כתבנו את הקוד וביצענו unit testing.
* הורדת הstop words הקטינה את כמות המידע.

**שאלה 2**

1. **שאלה**

Train a machine learning model to predict the gender of the tweet author. Evaluate three models, and tune parameters. One of the models should be based on ‘deep learning’ with Keras. Evaluation metrics: accuracy. Present train and test accuracy for different model and pre-processing combinations.

1. **פתרון**
2. **קשיים**

* נתקלנו בקשיים רבים בעת ביצוע ייבוא לספריית Keras הנחוצה לנו. רק לאחר קריאה מעמיקה באינטרנט ומספר ניסיונות של התקנה מחדש מצאנו כי עלינו לבצע עדכון לגרסת הPIP.
* הבנה ושימוש בחבילות מדף, מצאנו כי השימוש בשורה package?מציג את הAPI של החבילה ולעתים אף דוגמאות שימוש בה.
* הבנת יצירת ושימוש במודל (Sequential), מצאנו [מדריך](https://www.youtube.com/watch?v=VGCHcgmZu24) שעזר לנו מאוד <https://www.youtube.com/watch?v=VGCHcgmZu24>.

1. **סיכום**

* ביצענו ניסויים רבים במטרה למצוא ספים שיניבו תוצאות מקסימליות. עבור למידה עמוקה ביצענו ניסויים עבור: batch\_size , epochs (מספר איטרציות, מצאנו כי המספר שהניב את התוצאות הטובות ביותר היה 3, כנראה שמספר קטן מ3 אינו מספיק עבור למידה ומספר גדול מ 3 יוצר overfitting) ויחס החלוקה בין קבוצות הבדיקה והאימון.
* מצאנו כי השימוש ב Dropout מפחית התאמת יתר במודל.
* השתמשנו בתיעוד של ספריית keras להבנתו בצורה טובה יותר, <https://keras.io/>.

**שאלה 3**

1. **שאלה**

Use Twitter streaming API to collect 15,000 tweets from the country which was most popular in the training data. Optional: you can filter based on an hashtag which is expected to be related for a specific gender. Repeat the same pre-processing you implemented in Question 1 for the collected tweets. Analyze the most popular terms for this test dataset as well. Present terms frequency and discuss the similarity with the train.

1. **פתרון**
2. **קשיים**

* בעת השימוש ב OAuthHandler קבלנו שגיאות רבות. השגיאה שפתרונה דרש זמן עבודה רב ביותר היא שגיאת AttributeError שנבעה מייבוא לא נכון של החבילה.
* איסוף המידע מ Twitter streaming API לקח זמן רב, סה"כ 3 שעות בשל קבלת Time Out ושגיאה המצביעה על עומס על זרם הנתונים.
* המתנה בקבלת אישור מtweeter דחו את זמן הפיתוח ביומיים.

1. **סיכום**

* לאחר פתרון שאלה 2 הגישה לשאלה השלישית והבנתה היו פשוטים יותר והתקדמו לפי תכנון.

**שאלה 4**

1. **שאלה**

Use the best gender classification prediction model which was trained on Question 2 to predict the gender of the authors of collected tweets. Present your conclusions. Present the prediction results and your conclusions.

1. **פתרון**
2. **קשיים**

* בשאלה זו לא נתקלנו בקשיים, למעט ביצוע מספר רב של ניסויים להשגת תוצאה מיטבית.

1. **סיכום**