**פרויקט מסכם בהיבטים מעשיים בכריית תוכן**

מגישים: ליזה פרידמן 316871474 וליאור פפר 203202122

**מבוא**

לאחר מעבר על סוגים שונים של מודלים החלטנו להשתמש ב - **Keras**, **SVM** ו - **KNN** מכיוון שבשלושתם נוכל להשתמש ב - **TF-IDF Vectorizer** כדי להמיר את הטקסט של המידע הנכנס למערך של אינדקסים לפי מילון של המילים הכי נפוצות במאגר שאספנו.

**שאלה 1**

1. **שאלה:**

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

1. **קשיים:**

* שימוש יותר נפוץ בלולאות FOR יעילות לעומת הלולאות שאנחנו מכירים משפות OOP נפוצות.
* 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.
* שימוש ב DataFrame.loc מהיר משמעותית מביצוע איטרציות על הDF ולכן ניסינו להשתמש בו כמה שיותר.

**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. **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. **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.
* בשימוש בVersion Control קיימת סכנה של גניבת API keys של Twitter מכיוון שקיימים בוטים רבים שסורקים את GitHub על מנת לקבל אותם ללא צורך ברישום אפליקציות.

כדי להימנע מבעיות כאלה, ניתן לקרוא את הנתונים מקובץ או להשתמש בPrivate Repository. בחרנו באפשרות השנייה כדי להימנע מקריאות וכתיבות מיותרות לדיסק.

**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. **קשיים:**

* בשאלה זו אנחנו משתמשים במודל שבחרנו על מנת לנתח את הTweets שהורדנו בשאלה הקודמת, אולם במידע שאנחנו מורידים לא כלול המין של כותב הTweet ולכן על מנת לעמוד את הדיוק של המודל שלנו, היינו צריכים להשתמש בסיפרייה חיצונית בשם Genderize שיודעת לנתח את המין של הבן אדם לפי השם הפרטי שלו ולהשוות את הניתוח שלו לניתוח של המודל שלנו.
* בשימוש בGenderize נתקלנו בבעיה של קצת הורדה שאם אנחנו מורידים יותר מדי נתונים זמן קצר מדי, נזרק Exception שמונע הורדה עתידית. כדי להימנע מכך הורדנו מראש 1000 ניתוחים עם דיוק מעל 80% כדי לטעון במהלך הריצה ולהשוות לניתוח של המודל שלנו.

1. **מסקנות:**
   * לאחר הניתוח, נמצא כי הרוב המוחלט של ה - Tweets הם של נשים, בצורה הרבה יותר קיצונית ממה שנצפה במאגר שניתן לאימון המודל.
   * הדיוק של המודל שלנו נמוך מששוערך בהתחלה מכיוון שאנחנו משתמשים במאגר מוגבל יותר של שמות למין בעל השם כדי לקבוע אותו. מאגר זה גם לא כולל את האופציה של brand אך למרות שהורדנו את התוצאות של המודל שלנו שמצביעות על brand, הדיוק עדיין נמוך מששוערך.