**הוראות להרצת הפרויקט**

בכדי להריץ את הפרויקט יש להתקין קודם את כל החבילות המופיעות בקובץ libs.txt המצורף לפרויקט.

אנו ממליצים להתקין את החבילות הנ״ל בסביבת venv **נקייה** באמצעות pip ע״י הרצת הפקודה:

pip install -r libs.txt

בנוסף אנו **לא** ממליצים לנסות להריץ את הפרויקט בסביבת conda בשל בעיות התממשקות של anaconda עם החבילות המכילות את BERT ואת Pytorch.

לפרויקט שני 2 מודי ריצה שניתן למתג ביניהם ע״י שינוי השורות המתאימות בתחתית הקובץ Main.py:

* **מוד מרכזי (ברירת מחדל)** - מריץ את כל שלבי הפרויקט, החל מכריית הנתונים ועד הרצת המסווג ההיררכי. במצב הקוד הנוכחי חלק מהתהליכים הנ״ל (כמו כריית מאגר הנתונים) לא יתבצעו ונתוניהם יטענו ישירות מקבצי ה-JSON המצורפים, אולם ביכולתו של המשתמש לשנות את הדגלים המתאימים בהתאם לתיעוד בקוד בכדי להריץ את כל התהליכים השונים.

ניתן להפעיל/לכבות את המוד ע״י שינוי שורה 194 בקוד.

* **מוד אונליין** - מוד הרצה קטן המאפשר למשתמש להתנסות במסווג ההיררכי. על המשתמש להכניס שם של מאמר (כפי שהוא מופיע בויקיפדיה האנגלית) ואת האורך הרצוי למסלול הסיווג (בין 0-3) ולקבל מהמסווג חיזוי למסלול שבו יש לסווג את המאמר בויקיפדיה. על מנת לקבל חיזוי סביר על המשתמש לוודא שקיים למאמר לפחות מסלול חלקי בהיררכיית הקטגוריות של המאגר שלנו כפי שהיא מופיעה בגרף בעמ׳ 13 בדו״ח המצורף.

ניתן להפעיל/לכבות את המוד ע״י שינוי שורה 204 בקוד.

אנו ממליצים להריץ עד מוד אחד במקביל בכל הרצה.

**תיאור הקבצים המצורפים**

‏Main.py

קובץ ההרצה הראשי של הפרויקט המכיל את שני מודי ההרצה המרכזיים של הפרויקט כפי שהם מתוארים לעיל. בקובץ יש תיעוד לפרמטרים ולדגלים השונים לנוחיות המשתמש.

‏Classifiers.py

מכיל את המסווגים איתם עבדנו, כל מסווג מכיל גם הפרמטרים שעליהם כיילנו אותו.

‏DataClasses.py

שלושת המחלקות בהן שמרנו את מאגר הנתונים ההיררכי, כמתואר בדו״ח.

הקובץ מכיל גם את ה Random Seed לפרויקט כולו (על מנת שיהיה אפשר לשחזר את התוצאות).

‏DataUtilities.py

פונקציות ומחלקה (DataSplitter) לעיבוד וחלוקת מאגר הנתונים לקבוצות הלימוד, האימות והמבחן. כמו כן ניתן לבצע את החלוקה לקבוצות הנ״ל גם עבור קטגוריה ספציפית בלבד בהתאם לחלוקה הראשית של המאגר כולו (לכל קטגוריה לחשב מהם המאמרים שהינם צאצאים שלה ונמצאים בקבוצות הלימוד, האימות והמבחן הכוללות של המאמר). הפיצול הכולל נעשה על פי ה-Random Seed המוזכר לעיל.

‏FilesIO.py

פונקציות לטעינת ושמירת קבצים

‏GraphCreation.py

המחלקה שיוצרת את הגרפים וההיסטוגרמות עבור כוונון הפרמטרים.

‏HierarchicalArticleClassifier.py

מסווג המאמרים ההיררכי הראשי

‏Miner.py

המחלקה אשר מבצעת את כריית הנתונים

‏ParameterSelector.py

המחלקה אשר מממשת את אלגוריתם בחירת הפרמטרים

‏StatisticsGenerator.py

המחלקה אשר מחשבת נתונים סטטיסטיים על הדאטה

‏TextVectorizer.py

המחלקה אשר מבצעת את הוקטוריזציה של המאמרים באמצעות BERT

**תיאור התיקיות המצורפות:**

‏Data

מכילה את הדאטה שנאסף ופלט של הסטטיסטיקות עליו

‏Parameter Selection\Selected Parameters

מכילה את הפרמטרים הנבחרים וטבלאות התהליך מהן ייצרנו את הגרפים

‏Parameter Selection\Graphs

מכילה את הגרפים וההיסטוגרמות

‏Report

מכילה את דו״ח הפרויקט