מס' פרויקט	שם הפרויקט
2021 -01-187	השוואת קלט/פלט במשימות למידה מדיסק

מנחה אקדמי	מנחה שותף
ד"ר יונתן רוזנבלט	

חברי הצוות		
רותם גרבר		
rotemgar@post.bgu.ac.il		

תקציר

תחום נתוני העתק נמצא בעליה מתמדת לאורך השנים לאור ההבנה וההפנה בדבר כוחו של המחשב בביצוע תהליכים. חשיבות העיבוד והלמידה מנפח נתונים כה גדול ברורה וידועה, אך מצריכה התאמות וטכניקות ייחודיות על מנת לעבד נתונים אלו. כאן נכנסים לתמונה אלגוריתמי חוצי-ליבה. שזהו מושג המתאר תהליך עיבוד נתונים הגדולים בנפחם מכדי להיות מאוחסנים על גבי זיכרון ה-RAM. במצב כזה, הנתונים מאוחסנים על גבי זיכרון הדיסק, ומובאים במקבצים קטנים, בזה אחר זה, לטובת עיבודם. במהלך הפרוייקט אבצע השוואת ביצועים במשימות למידה מדיסק לפי שני שלבים המאפיינים את שלבי למידת מכונה.

בשלב הראשון אשווה ביצועי ספריות המכילות שיטות שונות לעיבוד נתוני עתק. וזאת משום שספריית Pandas, אינה מסוגלת לבצע תהליכי חוצי ליבה. הספריות שנבדקו היו Dask, Vaex, Datatable, Koalas. כל אחת מספריות אלו מכילה התאמות כגון אובייקטים מותאמים, עיבוד עצל ומיפוי זיכרון. בשלב השני, שלב למידת המכונה בפועל, הוצג אלגוריתם ממשפחת הלמידה המקוונת. אשר מאפשרת לערוך מודל על חתיכות קטנות מסט הנתונים הרחב ובכך ליצור מודל אינקרמנטאלי. היתרון הוא כפול, גם מודל עדכני באופן תמידי וגם למידה מסט נתוני עתק. ביצוע שלב זה נעשה באמצעות חבילת אלגוריתמי חוצי הליבה של ספריית Scikit-learn.

מתוצאות הפרוייקט עולה כי הספריה המתאימה ביותר להרצת אלגוריתמי חוצי ליבה לטובת תהליכי לימוד מכונה הינה Vaex. וזאת משום ששילוב תוצאות הביצועים, ותכונות המותאמות בדיוק בשביל משימה זו. מעבר לכך, ספרייה זו גם שימשה לשם הצגת יכולת הלמידה האנקרמטאלית. גם ביצועי ספריות Datatable ו- Koalas הראו תוצאות משביעות רצון. אך אלו חסרות תכונות המחייבות בתהליך למידת מכונה. אחרונה זו ספרית Dask שבה התוצאות היו האטיות ביותר. ספריה זו מכילה תכונות נוספות מעבר לספריות האחרות אך בבעיה אותה בדקתי, תכונות אלה לא באו לידי ביטוי.

מסקנות פרוייקט זה מסוגלות להוות בסיס מוצק לפרוייקטי המשך בתחום הלמידה מנתוני עתק בדגש על למידת מכונה. תחום אלגוריתמי חוצי-ליבה צובר תאוצה בשנים האחרונות, וכך הפתרונות המוצעים לשם הרצתם. אי לכך, יש צורך תמידי להתעדכן ולוודא שמסקנות הפרוייקט תקפות.

מילות מפתח: אלגוריתמים חוצי-ליבה, השוואת ביצועי ספריות, למידה מקוונת.