מגישים: מלאק אבו עדייה ובדיע אבו פריח **מנחה:** ד"ר עמי האופטמן

OCR Error Detector

רקע ומטרה

הפרויקט מנסה להקל על תהליך הדיגיטציה של החוקים והחקיקה בישראל ונעשה בשיתוף פעולה עם משרד המשפטים. מטרת הפרויקט לזהות שגיאות שנוצרו עקב דיגיטציה של מסמכים באמצעות OCR. חלק מן המידע המסופק לצורך הפרויקט הם מסמכים שעברו OCR והמקור שלהן בפורמט PDF. לצערנו, סריקת OCR אינה נאמנה למקור וקיימים מקרים בהן ישנן שגיאות עקב הפעלת תהליך זה על המסמכים. מטרת הפרויקט היא לאתר ולהתריע על השגיאות הללו. נעזרנו בשיטות למידת מכונה ולמידה עמוקה על מנת לפתור את הבעיה.

תיאור המערכת:

- קריאה ומיפוי לפי מזהה חוק של קבצי ה XML, הוצאת הטקסט מהקבצים הנ"ל.
- שאיבת מידע נוסף, כ 1990 חוקים נוספים מאתר ויקיטקסט.
- ביצוע עיבוד מקדים רלוונטי על הטקסט בעזרת ביטויים רגולריים.
- מעבר על כ 100 מסמכים באמצעות קריאה צמודה על מנת להכין מילונים של שגיאות נפוצות וכמות הופעתן.
- פגימה מכוונת בטקסט החוקים באמצעות המילונים שהכנו.
- חלוקת המידע לסטים ויצירת טוקניזר הממפה מילים למספרים. הכנסנו את כלל המילים שנמצאות בסט האימון.
- הזנת המידע למודל שאומן מראש וחיזוי למילים שמכילות שגיאות. לבסוף בחינת התוצאות ובחירת המודל הטוב ביותר.

אמצעים טכנולוגים:

- המערכת ממומשת בשפת Python, סביבת המחקר היא Jupiter Notebook וסביבת הפיתוח הינה VS-Code.
- בסיס הנתונים: SQLite. טכנולוגיות נוספות: Keras, Tensorflow, Sklearn.