מגישים:סאמר אבו הדיב, נג'יב מחאג'נה **מנחה:** ד"ר עמי האופטמן

דיגיטציה ועדכוני חוקים בערבית

רקע - מדי פעם נחקקים חוקים חדשים במדינה בנוסף לעדכונים על חוקים אחרים, העדכונים מתבצעים באופן ידני ואיטי לדוגמה אם חל שינוי על חוק כלשהו, הם חייבים לחפש ולאתר את החוק או הסעיף הנדרש ואז רק מתחילים לבצע את השינויים.

משרד המשפטים החליט לייעל את התהליך הבזבזני, ע"י הפיכת מסמכי ה-PDF וה-DOCS ל- XML שבהם תהליך העדכון והחיפוש לחוק מתבצע לפי התגיות.

תיאור המערכת-

למשרד המשפטים קיימת מערכת שהופכת את החוקים הקיימים בעברית לקבצי XML אבל מכיוון שיש הבדל גדול בכללים ובתחביר בין השפות אז הפתרון הקיים לא רלוונטי ולכן אנחנו בונים מערכת חדשה שיודעת להתמודד עם שתי השפות ולקשר בינהם, המערכת מקבלת ספר חוק בשפה הערבית עם XML בעברית ותעשה את השלבים הבאים:

- חילץ המידע מתוך התגיות, תרגום, פירוק מסמך הPDF שבערבית לפי כמה קריטריונים : הגדרות, כותרת החוק, ... וכו'.
- בנוסף נשתמש באלגוריתמים ובשיטות למציאת הביטוי המתאים: תרגום של הטקסט,זיהוי התחלה וסוף, זיהוי חוקיות.

בסוף נחליף את הביטויים בשפה העברית בביטויים בערבית ונקבל מסמך XML בערבית.

אמצעים טכנולוגיים -

ספריות opensource ב Java -

- OpenNLP -לעיבוד טקסט כמו חלוקת משפטים, מילים, נרמול לטקסט וכו..
- XMLDOM – לצורך ביצוע ניווט,שליפת מידע, עדכון תוך עץ ה XML.
- GoogleAPITranslate – ביצוע תרגום על מנת לאתר ביטויים לצורך החלפה.
- PDFBox – לצורך שליפת המידע מספרי החוק.
- Github - לצורך אחסון הקוד ועבודה בצוות בצורה יעילה ביותר.