פרויקט במדעי הרוח הדיגיטליים – כותרות צד בחקיקה

בפרויקט זה נתבקשנו לכתוב יישום המקבל כקלט קבצי xml המכילים מידע על חוקי הכנסת, ומאפשר לבדוק באלו חוקים מופיעים סעיפים המסומנים על ידי כותרת צד מסוימת.

קבצי הxml שמתקבלים כקלט הנם חצי מובנים, זאת מכיוון שלא בכל החוקים קיימים ערכים לתגים השונים, כלומר אין זה בהכרח שכל תג בקובץ ימולא ע"י ערך.
הפעלנו אלגוריתם חיפוש בקבצים שאפשר לנו לקחת את המידע, במידה וקיים, ולשמור אותו במבנה נתונים מובנה מסוג dictionary כך שלכל ערך יש מפתח משלו.
במעבר הנ"ל מקבצי הxml למבנה הנתונים לא שמרנו את המבנה המתויג אלא את תוכן הסעיפים הרלוונטיים לתוכנית, כלומר שמרנו רק את תוכן הסעיף שיש לו כותרת צד בצורה כזאת השומרת על מבנה הסעיף ואותו אנו מעוניינות להציג למשתמש (כותרת, מספר סעיף ותוכן).
מכיוון שלעתים הגדרות זהות או דומות מופיעות בהבדלים קטנים הפעלנו אלגוריתם של מתייג חלקי דיבר (YAP) העושה ניתוח מורפולוגי על כל כותרת צד. מכך קיבלנו את:
segmented text – אשר מכיל את כותרת הצד כך שהתחילית מופרדת מהעיקר

Lemmas – אשר מכילה את הsegmented text כך שכל מילה מוצגת בצורה הבסיסית שלה.

בהינתן כותרת צד אותה המשתמש הכניס לתיבת החיפוש אנו מפעילות אלגוריתם למציאת כותרות צד דומות.
האלגוריתם מפעיל את מתייגת חלקי הדיבר גם על מילת החיפוש, ובודק האם יש התאמה בין מילת החיפוש ותוצאות המתייג שלה ובין כותרת צד נתונה ותוצאות המתייג שלה.

את התוצאות אנו מציגות בצורת טבלה המאפשרת השוואה בין הסעיפים השונים. מבנה הטבלה הוא בצורה כזו בה יש חלוקה לכותרת הצד, שם החוק בה נמצא, שנה בה הוא פורסם ותוכן הסעיף עצמו.
בתוכן הסעיף קיים כפתור המאפשר הצגה מלאה/חלקית של התוכן וכמו כן קיימת אפשרות להציג את התוצאות ממוינות לפי תאריך פרסום החוק.