

יעקב הכהן-קרנר, אילון מלון, יצחק חסון

מערכת הלומדת לסכם כתבים תורניים הלכתיים

כמות המידע האלקטרוני העצומה הנגישה לבני האדם עקב ההתפתחויות בתחומי מסדי הנתונים והתקשורת, מקשה על בני האדם להתמודד עם קריאה של טקסטים, הבנתם וסיכומם. ישנו צורך אמתי ביצירה אוטומטית של סיכומי טקסטים מסוגים שונים. באמצעות קריאת סיכומי טקסטים, יכול אדם לדעת במה עוסק הטקסט, ולעתים אפילו לדעת חלק ניכר מתוכנו, מבלי לקרוא את כולו. בנוסף, הסיכום יכול לעזור לו לקבל החלטה אם לקרוא את כולו או לא. בעיית הסיכום האוטומטי בולטת כאשר מדובר בשפה העברית. בשפה האנגלית, כמו גם בשפות זרות אחרות, ישנן מערכות המבצעות סיכומים אוטומטיים בדמה כזו או אחרת, בעוד שבשפה העברית, ככל הידוע לנו, אין כלל מערכת כזו. הוחלט לבנות מערכת סיכום עבור מאמרים תורניים-הלכתיים. תלמידי חכמים בכלל ופוסקי הלכה בפרט נדרשים בימינו לדעת מספר רב יחסית של כתבים תורניים קודמים, בטרם כותבים כתבים או פסקים משלהם. סיכום אוטומטי של כתבים קודמים כאלו יכול להציג את עיקרם של הכתבים ולאפשר התמקדות בכתבים הרלוונטיים יותר. בכך, ניתן לחסוך זמן ומאמץ ניכרים ואף לעתים לעזור בהבנת הטקסט ובסיכומו. מאמר זה מתאר את בנייתו של אב-טיפוס של מערכת הלומדת לסכם כתבים תורניים-הלכתיים. המערכת פותחה במסגרת פרויקט גמר שבצע על-ידי הכותבים השני והשלישי בהנחיית הכותב הראשון במחלקה למדעי המחשב, בית הספר הגבוה לטכנולוגיה, מכון-לב, ירושלים. המערכת מייצרת מספר סוגים של סיכומים עבור כתבים תורניים מתחום פסיקת ההלכה. סוגי הסיכומים שהוצגו על-ידי המערכת: תחום המאמר, מילות מפתח וסיכום מסקנתי באורך של כ-10% מהמאמר המקורי (האורך נמדד במספר משפטים), הכולל את המשפטים המציגים את עיקרי הדברים המסקנתיים שנכתבו במאמר. ביצועי המערכת שופרו באמצעות אלגוריתם למידה אוטומטי בשם Genetic Algorithm. איכות הסיכום שקולה לאיכות הסיכום של מערכות דומות בשפה האנגלית.

א. מבוא

ברורנו, דור התפוצצות המידע האלקטרוני, קשה לבני אדם להתמודד עם קריאת טקסטים, הבנתם וסיכומם, בפרט כאשר מדובר בכמות רבה של טקסטים אשר חלקם ארוכים. לכן, גדל הצורך

* הננו מורים ללקטור ולעורכת הלשונית על הערותיהם והארותיהם למאמר זה.

בסיכומי טקסטים. באמצעות קריאת סיכום טקסט, יכול אדם לדעת במה עוסק הטקסט, לקבל החלטה אם לקרוא את כולו או לא, ולעתים אפילו לדעת חלק ניכר מתוכנו, מבלי לקרוא את כולו. באופן זה, יכול האדם הנגיש לכמות טקסט גדולה, לדלות את המידע הדרוש לו תוך חסכון ניכר בזמן.

בני-אדם ידועים כבעלי זיכרון עצום ובעלי יכולת סיכום מעולה. אולם המאמצים והזמן הנדרשים לסכם כמות כה גדולה של טקסטים הינם כמעט בלתי-אפשריים. לכן, ישנו צורך אמתי ביצירה אוטומטית של סיכומי טקסטים מסוגים שונים.

בעיית הסיכום האוטומטי בולטת כאשר מדובר בשפה העברית. בשפה האנגלית כמו גם בשפות זרות אחרות ישנן מערכות המבצעות סיכומים אוטומטיים ברמה כזו או אחרת, בעוד שבשפה העברית, ככל הידוע לנו, אין כלל מערכת כזו.

במערכת הנדרונה הוחלט להתמקד בתחום מצומצם יחסית: תחום המאמרים התורניים-הלכתיים. על פוסק ההלכה בימינו לעבור על כמות גדולה של כתבים תורניים, ברצותו לפסוק הלכה. על מנת לחסוך בזמן העיון בטקסטים, יכול פוסק להיעזר בסיכומי הכתבים כדי להתמקד בתוכנו העיקרי של הטקסט, ומגמת פסיקת ההלכה. לכן נדרשת מערכת המסוגלת לסכם טקסטים תורניים באופן אוטומטי. המערכת שנבנתה מאפשרת להציג למבקשי התורה בכלל, ולתלמידי חכמים ופוסקים בפרט, סיכומים סבירים ברמות פירוט שונות, כדלקמן: תחום המאמר, מילות מפתח וסיכום מסקנתי. הסיכומים מבוצעים על פי שיטות שונות מתחום הבינה המלאכותית. סיכומים אלו יכולים לעזור:

- (א) בהצגה ראשונית של עיקרם של הטקסטים כדי שהמשתמש יוכל לדעת אם הם רלוונטיים.
- (ב) בהבנה של הרברים החשובים והעיקריים בטקסט.
- (ג) בסקירה של כתבים תורניים מרובים, בזמן קצר יחסית.
- (ד) בהפניות למקורות וטקסטים אחרים.

ראשיתו של מחקר זה פורסמה על-ידינו במאמר [10]. מאמר זה הציג 9 שיטות בסיסיות לסיכום ללא כל למידה אוטומטית. המחקר המתואר במאמר זה הינו הרחבה במובנים הבאים: (1) נוסחו שתי שיטות סיכום נוספות ייחודיות לשפה העברית בכלל ולתחום התורני-הלכתי בפרט ו-(2) יושמה שיטת למידה אוטומטית בשם Genetic Algorithm שהביאה מחד גיסא לשיפור יכולת הסיכום ומאידך גיסא הביאה גם לתוצאות סיכום סבירות גם יחסית לאלו של מספר מערכות סיכום מקבילות בשפה האנגלית.

ב. סוגי סיכומים

סיכום טקסטים – התהליך של זיקוק המידע החשוב ביותר ממקור (או מקורות) במטרה לייצר גירסה מקוצרת. במדעי המחשב מקובל שהסיכום באורך של 1% עד 30% מאורכו של הטקסט המקורי (אורך נמדד במספר משפטים לעניין זה) עבור משתמש (או משתמשים) ומשימה (או משימות) ספציפיים [17]. לעתים, תמצית באורך של 20% מאורך הטקסט המקורי, אינפורמטיבית

באותה מידה כמו הטקסט המקורי כולו [12]. סיכום אוטומטי של טקסטים מובנו סיכום הנעשה באמצעות תוכנת מחשב באופן אוטומטי לחלוטין.
להלן שלושת סוגי הסיכומים שנבנו במערכת שלנו: תחום המאמר, מילות המפתח שבמאמר וסיכום מסקנתי של המאמר.

תחום המאמר

תחומי המאמרים הינם התחומים שהוגדרו על-ידי עורכי תחומין [28] ותחתם קוטלגו המאמרים השונים. התחומים הרלוונטיים עבור אסופת המאמרים שנבחרו היו שלושה-עשר במספר: חוזים והסכמים, יישוב הארץ, מאכלות אסורות, מניעת פשע ועונשין, מקדש, סדרי דין, עבודה זרה, שבת, מתים, שמיטה, תחוקה ושלטון, ומועד.

אף כי כותרתו של המאמר תעיד במקרים רבים על התחום שאליו הוא שייך, במקרים מסוימים הכותרת אינה אינדוקטיבית מספיק כדי לדעת זאת. לדוגמה: בכותרת "אורכה של שנת עראי" לא ברור אם מדובר על איסור של שנת קבע בתפילין, או שמא על הצורך בנטילת ידיים לאחר שינה כזו. בעיה נוספת בהערכה אינטואיטיבית של התחום שבו עוסק המאמר על-פי הכותרת היא שלא ניתן לגלות תחומים צדדיים שבהם עוסק המאמר. וכל זאת בהנחה שיש כותרת, והרי ישנם טקסטים אשר אין כותרת בראשם כלל וכלל כי אם שאלה (כמו בשו"תים למיניהם), עובדה המאריכה בהרבה את פרק הזמן הדרוש כדי להבין האם המאמר רלוונטי עבורנו או לא. בכל-אופן, בעיה זו לא עמדה לפנינו במחקר זה, הואיל ולכל מאמרי תחומין כותרות.

מילות מפתח

מילת מפתח בנידוננו היא מילה אחת או ביטוי המורכב ממספר מילים, המתארות נושא חשוב מנושאי המאמר הנידון. מילות מפתח יכולות לסייע מאוד לקדאים אנושיים להבין את רלוונטיות המאמר עבורם ולשמש כתקציר כללי ראשוני ובכך לתרום לסינון של מספר רב יחסית של מאמרים. המערכת שנבנתה הגדירה את מילות המפתח של מאמר מסוים מבחינה בסיסית באמצעות בניית טבלת שכיחויות של המילים והביטויים המופיעים במאמר. המערכת מצאה גם ביטויי מפתח המכילים יותר ממילה אחת, כגון: "פסוקי דזמרא". המערכת ניפתה מילים מסוימות אשר להן תפקידים מיוחדים, כגון: לכן, אם, יש, וכאשר (פירוט בנספח א). מילים אלו לא הוצגו כמילות מפתח על אף שחלק גדול מהן הופיעו בשכיחות גבוהה.

סיכום מסקנתי

סיכום מסקנתי הוא סיכום בעל אופי מסקנתי. אורכו (נמדד במספר משפטים) הינו בדרך-כלל כ-10% או 20% מגודל הטקסט המקורי. סיכום זה אמור להציג את עיקרי השיטות המרכזיות, וכמובן, את המסקנה הסופית של הטקסט. סיכום טקסטים הוא מלאכה הדרושת בינה אנושית. שכן, תהליך זה כולל רכיב קוגניטיבי המזהה אלו מחלקי המאמר הם החשובים ביותר. בתחום

הבינה המלאכותית קיימות שתי גישות עיקריות לסיכום טקסט באמצעות מערכת ממוחשבת, נעמור עליהן להלן:

השיטה הראשונה היא שיטת עיבוד שפה טבעית (Natural Language Processing), להלן: (NLP). NLP הוא תחום בבינה מלאכותית אשר מנסה לעשות שימוש במחשבים כרי לעבר מידע המופיע בשפה טבעית כאנגלית. ישנן רמות שונות וגישות שונות ל-NLP הכוללות פונולוגיה (פונטיקה וצליל), מורפולוגיה (עיצוב מילים מתוך יחידות משמעות בסיסיות נוספות, רקרוק (עיצוב משפטים מתוך מילים), סמנטיקה (משמעויות של משפטים אשר נגזרו ממילים), ופרגמטיקה (הבנה של האופן שבו משתמשים במשפט). סיכומי טקסט שעשו שימוש בשיטה זו יושמו במערכות שונות, כגון: [1] Aone, [19] McKeown and Radev ו-[23] Radev.

שיטה זו מבוססת על הבנה של המשפטים המופיעים במאמר (הבנה במובנים של מחשב: קשרים בין חלקי המשפט, וקשרים בין המילים המופיעות בו לתחומים אחרים) ובנייה של משפטים חדשים של המערכת בכוחות עצמה, שיוצרו כמשפטי סיכום. לשיטה זו ישנם מספר מודלים מתוחכמים מאוד אשר דורשים מסרי נתונים גדולים ביותר, וכמורכב גם זמן עיבוד רב. המשפטים נבנים בעזרת תבניות לשוניות וחוקים רקרוקיים שונים. שלב זה של ההצגה למשתמש מצריך אף הוא זמן ריצה ארוך, ומאגרי נתונים בגדלים עצומים. שימוש בשיטה זו הוצגה לראשונה על-ידי Aone ושותפיו [1].

השיטה השנייה, שיטה השכיחה הרבה יותר, היא שיטת חילוץ משפטים — sentence extraction. שיטה זו מבוססת על מחקר שנערך בשנת 95 על-ידי Kupiec [12] אשר גילה כי 79% מהסיכומים אשר נכתבים על-ידי גורם אנושי רומים מאוד למשפטים המופיעים במאמר עצמו. למעשה, חלק מן המשפטים אינם אלא ציטוט משפטים מילה במילה מן הטקסט המקורי. על פי ממצאים אלו, אוסף של משפטים רלוונטיים חשובים שהוצאו באופן ישיר מן הטקסט המקורי ללא כל עריכה לשונית או הגהה, יכול לשמש כתקציר הולם. מערכות סיכום המבוססות על שליפה של משפטים נבחרים, בדרך כלל, נותנות ציון למשפטים על-פי מאפיינים שונים, המעידים על הרלוונטיות של משפטים אלו, ומציגות את המשפטים שקיבלו את הציונים הגבוהים ביותר על-פי סדר הופעתם בטקסט המקורי כסיכום. שיטה זו פשוטה יותר ליישום ויש לה יתרון ברור של זמן ריצה קצר. לכן, היא יושמה במערכות רבות, כגון: [6] Edmondson, [15] Lugin, [3] Barzilay and Elhadad, [12] al. ו-[11] Hovy and Lin.

ג. שיטות סיכום מקובלות לחילוץ משפטים

ישנן מספר שיטות מקובלות בתחום הסיכומים האוטומטיים להערכת מידת הרלוונטיות של משפט להופעה בסיכום שהמערכת מייצרת. להלן שיטות שניסינו במחקר זה:

שיטת סיכום 1: שיטת מירב ביטויי מפתח שכיחים המופיעים במינימום משפטים

TF-ISF (term frequency-inverse sentence frequency)

משפטים הכוללים בתוכם מספר רב של מילות מפתח אשר מופיעות בפחות משפטים הם בעלי סיכויים גבוהים יותר להשתייך לסיכום מאשר משפטים הכוללים בתוכם מילות מפתח המופיעות באותה שכיחות אך במספר רב יותר של משפטים [21]. שיטת TF-ISF מבוססת על שיטת TF [6,15] והתגלתה בניסויים שערכנו כטובה יותר. שיטת TF מעריכה משפט על-פי מספר ביטויי/ מילות המפתח המופיעים בו. כדי להבחין בין מילות מפתח חשובות לבין ביטויים אחרים המערכת בנתה טבלת שכיחויות אשר נתנה לכל ביטוי ניקוד על-פי מספר המופעים שלו ושל הטייתו השונות בטקסט. מילים וביטויים אשר להם תפקיד דקדוקי (כגון: אשר, אם, יש ו-כדי) הוצאו מטבלה זו באמצעות stop list שהוכנה מראש לצורך זה (נספח א). בתום בניית טבלת השכיחויות, כל משפט קיבל ניקוד על פי מספר ביטויי/מילות המפתח המופיעים בו ושכיחותם במאמר. בנוסף לשיטת TF השיטה לקחה בחשבון בבנייתה את טבלת הערכים של מילות המפתח גם את מאפיין ISF המעניק ניקוד גבוה יותר לביטויים שכיחים המופיעים בפחות משפטים. מאחר ו-ISF הוא מאפיין פחות חשוב מאשר TF, הכפלנו את TF ב-ISF במקום ב- $\log_2(ISF)$ עצמו [21].

שיטת סיכום 2: שיטת המילים הרומזות CW (cue words)

שיטה זו אבחנה משפטים חשובים אשר יכולים לסכם את המאמר על-פי מילים הרומזות (cue words) למשפטים מסכמים, כגון: לסיכום, למסקנה והלכה למעשה (נספח ב כולל את רשימת המילים הרומזות). ככל שיש במשפט יותר מילים רומזות כך יקבל המשפט ניקוד גבוה יותר [6].

שיטת סיכום 3: שיטת אורך משפט SL (sentence length)

שיטה זו נתנה ניקוד למשפט על-פי מספר המילים המופיעות בו. שיטה זו הניחה שהמשפטים הארוכים יותר, הם בדרך-כלל החשובים יותר, ועל כן יש להם סיכוי גבוה להיות משפטי סיכום טובים [24]. שיטה זו ניקדה כל משפט על-ידי חלוקה של מספר המילים המופיעות בו במספר המילים של המשפט הארוך ביותר [13].

שיטת סיכום 4: שיטת הניקוד השלילי Negative

מבחינה רעיונית שיטה זו רומה מאוד לשיטת cue words, אלא שבמקום למצוא את המשפטים שיופיעו בסיכום, היא סיננה את אלו שלא סביר לבחורם כמשפטי סיכום. שיטה זו זיהתה משפטים אלו על-ידי מילים רומזות, כגון: "לדוגמה" ו-"ייתכן ש". שיטה זו ניקדה כל משפט על-ידי חלוקה של מספר המילים השליליות המופיעות בו במספר המילים השליליות המקסימלי שמכיל משפט אחר בטקסט [20]. (רשימה של המילות השליליות הרומזות נמצאת בנספח ג).

שיטת סיכום 5: שיטת מיקום המשפט (SP (sentence position)

שיטה זו התבססה על ההנחה שמשפטים מסכמים או חשובים יופיעו תמיד במקום מסוים במאמר, כגון: בסוף או בהתחלה. שיטה זו השתמשה בטבלת הסתברויות אשר קבעה לפי מספר פסקה ומספר משפט בפסקה, את יכולת המשפט לסכם. טבלת הסתברויות זו נבנתה על-ידי ניתוח סטטיסטי של מיקומם של משפטים בסיכומים שנעשו בירי ארם בתוך הטקסט המקורי [6,14,16].

שיטת סיכום 6: שיטת זמיון לכותרת המאמר (TR (title resemblance)

שיטה זו העניקה ניקוד למשפטים על-פי רמת הרמיון שלהם לכותרת המאמר. משפטים אשר דומים יותר לכותרת המאמר קיבלו ציון גבוה יותר, וסיכוי גבוה יותר להיבחר לסיכום [21,6]. גם שיטה זו מבוססת ברובה על פונקציית הרמיון שתירון בהמשך.

שיטת סיכום 7: שיטת זמיון לכותרת הפרק (STR (sub-title resemblance)

שיטה זו אף היא העניקה ניקוד על-פי רמיון לכותרת, אלא שזו השוותה רמיון לכותרת המשנית ולא לכותרת של המאמר כולו. משפטים אשר דומים יותר לכותרת הפרק בו הם מופיעים קיבלו ציון גבוה יותר, וסיכוי גבוה יותר להיבחר לסיכום [16]. למותר לציין שאף שיטה זו נזקקת לפונקציית ההשוואה בין משפטים.

ד. מאפיינים ייחודיים של השפה העברית

כפי שצויין לעיל, מספר שיטות סיכום נזקקות למדר המגדיר את מירת הרמיון בין שני משפטים. אף על פי שקיימים מספר מדרים כאלו בתחום, הם לא התאימו למערכת שלנו כיוון שהמערכת שלנו בשפה העברית, בעור שמררים אלו נבנו עבור השפה האנגלית, עבורה פותחו רוב המערכות בתחום. להלן מספר מאפיינים של השפה העברית אשר מנעו מאיתנו להשתמש בשיטות המקובלות:

א. זמנים

רוב ההטיות בשפה האנגלית שונות מהטיית הבסיס, במספר אותיות בסופה של המילה. לכן, קיצוץ שתי המילים האנגליות המשוות החל מהאות החמישית [9] או השישית [24] מספיק במרבית המקרים בכדי לפתור את שאלת ההשוואה.

בשפה האנגלית יש מספר מועט של הטיות. לרדוגמה, למילה summary קיימות ההטיות הבאות: summarize, summarized, summaries, summarization. כל חמש המילים מתחילות באותיות "summar". בעברית, לעומת זאת, קיצוץ כזה לא יעזור מאחר וההטיות השונות של פועל מסוים משנות את הטיית הבסיס באופנים שונים (ראה: <http://www.morfix.co.il>). במקרים מסוימים אותו שורש יכול להופיע עד כ- 7000 צורות עבור זמנים וגופים שונים. מאפיין זה של השפה העברית גורם לפעולת ההשוואה בין שתי מילים ללא כלים

מורפולוגיים (מורפולוגיה – תורת תצורת המילים) להיות כמעט בלתי אפשרית. לרוגמה, שתיים-עשרה (וזה רק מעט-מזער) המילים העבריות הבאות מקודרן באותו השורש (ס.כ.ס.): "מסוכם", "מסוכמים", "סיכמתי", "סכם", "סיכמו", "תסכמנה", "סיכומינו", "ולכשנסכם", "משסיכמתן", "לסיכומו", "ונסכם" ו-"מסכמים".

ב. אותיות סופיות

בשפה העברית ישנן 5 אותיות אשר נכתבות בצורה שונה כאשר הן מופיעות בסוף המילה (מ-ם, נ-ן, צ-ץ, פ-ף, כ-ך). לרוגמה: במילה 'מסוכם' האות הראשונה והאחרונה נכתבות באופן שונה בזמן שהן, למעשה, מייצגות אותה אות. מאפיין זה של השפה העברית אף הוא מקשה על ההשוואה בין מילים.

ג. אותיות תחיליות (אותיות יחס)

שלא כמו בשפה האנגלית, אשר יש בה מילים המבטאות יחס, כגון: from, to, by, בשפה העברית מילות יחס מבטאות פעמים רבות על-ידי האותיות 'לכבו'ר משה' שמצורפות למילה בתחילתה או בסופה, כאשר כל אות מבטאת יחס שונה. לרוגמה הביטוי: 'from the summary' באנגלית ייכתב בעברית על-ידי מילה אחת: 'מהסיכום', הביטוי: 'when he will summarize' באנגלית ייכתב בעברית על-ידי מילה אחת: 'לכשיסכם'.

ד. אותיות שייכות

באנגלית קיימות מספר מילים המבטאות שייכות, כגון: my, her, his. בשפה העברית, השייכות מבטאת הרבה פעמים על-ידי אות או מספר אותיות אשר מוצמדות לסוף המילה (כגון: י, ינו, ס). לרוגמה, הביטוי: 'my summary' באנגלית, ייכתב בעברית במילה אחת: 'סיכומי'. הביטוי: 'their summary' באנגלית, ייכתב בעברית במילה אחת: 'סיכומם'.

ה. כתיב מלא וחסר

טקסט הנכתב ללא ניקוד נכתב שונה מטקסט הנכתב עם ניקוד, שכן פעמים רבות יש צורך באותיות אהו"י (אמות הקריאה) על מנת לוודא שהמילה תיקרא באופן נכון. עיון בספרים ובעתונים, בחיבורים ובמכתבים שונים יעמידנו על כך, כי אין בידנו למעשה כתיב אחיד בטקסט חסר הניקוד [27, 25]. ישנן מילים רבות בשפה העברית שניתן לכתוב בדרכים שונות, לרוגמה: כיסא – כסא, צרופיהן-צירופיהן ו- לידה-לדה.

ו. ראשי תיבות

ראשי תיבות הם קיצורי מילים באמצעות כתיבת אותיותיהן הראשונות בלבד. לרוגמה: (1) חכמינו זכרונם לברכה בראשי תיבות חז"ל ו-(2) ראשי תיבות בראשי תיבות ר"ת. נראה שראשי

תיבות שכיחים יותר בשפה העברית מאשר בשפה האנגלית. תופעה זו נכונה שבעתיים בכתבים תורניים המשופעים בר"ת. עפ"י עיבודים סטטיסטיים שביצענו על [26] התברר כי השפה העברית מכילה כ-17,000 ר"ת (וזאת ללא הרבה ר"ת הייחודיים לתחומים מקצועיים שונים). מתוכם לכ-6,000 (כ-35%) ר"ת יש יותר מאשר פידוש אחד. דוגמה קיצונית לראשי תיבות שלהם יותר מפידוש אחד הם ראשי התיבות א"א אשר להם הוצעו כ-110 פירושים אפשריים ב-[26], ביניהם: אמד אבדהם, אי אפשר, אשת איש, אבות אבותינו, אבי אבי, אבי אמי, אם אבי, אם אמי, אין אוכלים, אין אומרים ו-אם אומרים.

אמנם המערכת לא התמודדה עם כל מאפייני השפה העברית. אך היא טיפלה בצורה סבירה בחלקם, ביניהם: אותיות סופיות, אותיות תחיליות/יחס, אותיות שייכות, כתיב מלא וחסר וראשי תיבות.

ה. מדדי הישגים

המערכת מדרה את יכולת המשפטים לסכם עפ"י תשעה מדדים (שיטות) שונים(ות) (שבעת הראשונים הוזכרו לעיל ושניים נוספים ייחודיים לשפה העברית שפותחו על-ידינו יידונו בסוף פרק ח). המערכת היתה צריכה להתחשב בכל מרד במידה שונה. בכדי לדעת כמה יש להתחשב בכל מרד מתשעת המדרים, נבדקה יכולת ההצלחה של כל מרד כאשר נעשה שימוש רק בו. מובן, שכדי למדוד את יכולת ההצלחה, יש צורך במדדים מוגדרים מראש. בדיקת יכולת ההצלחה שימושית גם עבור למירה, בעזרתה ניתן לבדוק אם השינויים שנעשו גרעו או הועילו. יש לציין שהמדדים הללו מודדים את יכולת ההצלחה של הסיכום ביחס לסיכום ייחוס (הגדרה בסעיף הבא). ישנם שני מדדים עיקריים הידועים בתחום איחזור המידע: המרר הראשון נקרא precision. מרר זה מוגדר כמספר המשפטים שהופיעו גם בסיכום המערכת וגם בסיכום הייחוס מחולק במספר המשפטים בסיכום המערכת. במילים פשוטות ניתן לומר שמרר זה בודק מהו החלק היחסי של משפטי הסיכום שבנתה המערכת המהווים משפטים מסכמים מתוך כל המשפטים שהוצעו לסיכום על-ידי המערכת.

המרר השני נקרא recall. מרר זה מוגדר כמספר המשפטים שהופיעו גם בסיכום המערכת וגם בסיכום הייחוס מחולק במספר המשפטים בסיכום הייחוס. במילים פשוטות ניתן לומר שמרר זה בודק מהו החלק היחסי של משפטי סיכום הייחוס שהמערכת הצליחה להציג בסיכום שלה מתוך כלל משפטי סיכום הייחוס. לדוגמה, אם סיכום הייחוס הוא בעל 20 משפטים (כמספר משפטי הסיכום של כותב המאמר), וסיכום המערכת הוא בעל 40 משפטים, אשר 10 מתוכם מופיעים בסיכום הייחוס והשאר לא, ערך ה-precision יהיה $10/40$ (25%), וערך ה-recall יהיה $10/20$ (50%).

המרר שהיה לנו חשוב הוא דווקא מרר ה-recall ממספר שיקולים. השיקול העיקרי הוא שהיה לנו שהרבה מידע סיכומי יופיע, על אף שיחד עימו יופיע גם מידע שאינו סיכומי.

שכן, המערכת עדיין לא מומחית אמיתית בתחום הסיכום, ולכן עדיף היה שתסייע בסיכום ולא תתיימר להיות המסכם עצמו. לסיכום נבחרו 10% מהמשפטים במאמר, שזהו מדד מקובל ביותר במערכות סיכום (חלק מהן אף מסכמות בשיעור גבוה יותר, מה ש"תורם" לשיפור תוצאות מדד ה-recall). לאחר עיון בתוצאת המערכת, גודם אנושי מקצועי יסנן בקלות את המשפטים שאינם חשובים לו באם ירצה להשתמש בסיכום שהוכן על-ידי המערכת שנבנתה בסינן ראשוני. שיקול נוסף היה שגם משפטים שאינם סיכומיים מובהקים יוכנסו לסיכום, שהרי הם חשובים משום שהם נבחרו על-פי מאפיינים סיכומיים שהיו בהם. בנוסף, גם מערכות הסיכום שנבנו בידי Kupiec ושותפיו [12], [16] Bloedorn ו-Mani [16] ו-Neto ושותפיו [21] הדגישו דווקא את מדד ה-recall, וניסו לשפר דווקא אותו.

יצירת "סיכום ייחוס"

על מנת שהצלחת המערכת תוכל להיברק היה צורך בסיכומים שהוכנו על-ידי חילוף של המשפטים החשובים במאמר. ברם, לא היו לנו אלא סיכומים של המאמרים אשר הוכנו באמצעות מחברי המאמר עצמם, כאשר המערכת שנבנתה בנתה את הסיכום האוטומטי על-ידי חילוף משפטים מטף המאמר עצמו ללא סיכום המחבר. משפטי המחבר בסיכום לא הופיעו לרוב בצורה זהה בגוף המאמר, ולכן קשה היה להשוות סיכום שנעשה על ידי חילוף משפטים כהווייתם מגוף המאמר לסיכום המקורי של המחבר. כדי לפתור את הבעיה הציעו Mani and Bloedorn [16] תהליך באמצעותו ניתן להכין סיכום ייחוס הבנוי ממשפטים מגוף המאמר הדומה לסיכום הקיים מעשה ידי הכותב. בצורה כזו ניתן להשוות את סיכום המערכת לסיכום הייחוס המכיל רק משפטים מגוף המאמר במקום לסיכום מעשה ידי הכותב.

התהליך מבוסס על האלגוריתם הבא:

1. עבור כל משפט AS_i ממשפטי הסיכום של הכותב מצא משפט RS_j מגוף המאמר הדומה ביותר למשפט AS_i על-פי פונקציית ההשוואה $res(S_i, S_j)$ (המפורטת בפרק ג)
2. קבוצת המשפטים (RS) תהווה את סיכום הייחוס

1. שיפור אוטומטי של המערכת באמצעות שיטת הלמידה

(GA) Genetic Algorithm

למידה אוטומטית היא כלי יעיל למציאת קומבינציות אופטימליות. על מנת שבמהלך הלמידה המערכת תוכל לדעת אם תוצאותיה השתפרו היא השתמשה בשיטת ההשוואה לסיכומי הייחוס אשר יוצרו באופן אוטומטי. כפי שהוסבר לעיל, המשקלים של השיטות השונות נקבעו על-ידי מדרים סטטיסטיים של יכולת ההצלחה של כל שיטה. במהלך המחקר נעשה ניסיון לשפר את תוצאות המערכת על-ידי מתן של משקלות שונים ובכך להגיע לצינוני recall גבוהים יותר. שיטת הלמידה בה השתמשנו מתוארת בפסקה הבאה.

שיטת הלמידה הגנטית מבוססת על המכאניזם של בחירה טבעית וגנטיקה טבעית [8]. הם משלבים את עקרונות ההתפתחות של גנים בטבע, כגון: 'החזק שורר', הורשה, מיוזג הורים, מוטציות ואקראיות. האלגוריתם מפתח דורות של קומבינציות, כאשר בכל דור גורל האוכלוסייה נשאר קבוע. בכל דור נבחרות קומבינציות עם עריפות הסתברותית לקומבינציות המוצלחות יותר. באופן כללי, אלגוריתם בסיסי של GA מורכב מהצעדים הבאים [18]:

1. אתחל אקראית את הרור הראשון של אוכלוסייה ובו n אזרחים.
2. הערך את טיבו של כל אחר מהאזרחים באוכלוסייה.
3. כל זמן שתנאי העצירה (הגרדה בפסקה הבאה) אינו מתקיים צור את הרור החדש באופן הבא:
 - 3.1. בחר חלק מאזרחי הרור הקודם כאזרחים גם ברור החדש (בהסתברות גבוהה יותר ייבחרו האזרחים שהערכתם גבוהה יותר).
 - 3.2. שאר אזרחי הרור החדש יוצרו על-ידי מיוזג (crossover) בין אזרחי הרור הקודם על-פי רגימה הסתברותית.
 - 3.3. בצע שינויים קטנים (mutations) בחלק קטן מן האזרחים שנוצרו בעקבות צעד 3.2.
 - 3.4. הערך את טיבו של כל אחר מהאזרחים באוכלוסייה החדשה.

תנאי העצירה מגוונים ותלויים בסוג פתרון הבעיה הרצוי. דוגמאות לתנאי עצירה נפוצים הן: זמן ריצה מסוים, מספר דורות מסוים, טיב מסוים של אחד האזרחים המספק פתרון מספיק טוב לבעיה ו-אין השתפרות מספיקה מרור לרור (נקודת רוויה).

שני שלבים משמעותיים באלגוריתם הגנטי הינם פעולות המבוצעות על אזרחי הרור הקודם: ומיוזג הזוגות (crossover) ויצירת המוטציות (mutations).

מיוזג הזוגות נעשה בדרך כלל על-ידי הורשה של ערכי האלמנטים מההורים לצאצא, כאשר את ערכי חלק מהאלמנטים יורש הצאצא מאחד ההורים, ואת ערכי שאר האלמנטים יורש הצאצא מהורה השני. ההיגיון הוא שברומה לטבע על-ידי הכלאה של זוג הורים יכול להתפתח דור חדש טוב יותר שיירש מכל הורה את התכונות הטובות יותר. התקווה היא שצאצא יירש מהוריו מאפיינים שונים של הפתרון שיצרו ביחד פתרון טוב יותר לבעיה.

יצירת המוטציות נעשית בדרך כלל על-ידי שינוי אזרח במאפיין אחד. גם עיקרון זה דומה לטבע, שבו לעתים משתנה הגנום בצורה אקראית באופן פתאומי מבלי שירש שינוי זה מאחד מהוריו. שינוי זה מכונה מוטציה, ולעתים דווקא שינוי אקראי זה מסגל תכונות משופרות לגנום שלא יכלו להיווצר בירושה מהוריו. החשיבה היא שלעתים הפתרון לבעיה מגיע דווקא בצורה לא שיטתית של הכלאת הפתרונות הטובים, אלא משינוי אקראי.

ז. מערכות קודמות

להלן סקירה של מספר מערכות קודמות הרלוונטיות במירת מה למחקר זה. ראשית נסקור מערכות סיכום קודמות, רובן בשפה האנגלית. לאחר מכן, נסקור מספר מערכות בשפה העברית

המטפלות במידה זו או אחרת במאפייני השפה העברית לשם איחזור מידע מכתבים תורניים.

א. מערכות סיכום קודמות

1. המערכת הלומרת הראשונה לסיכום אוטומטי באמצעות חילוץ משפטים שנבנתה על-ידי Chen ו-Pederson בשנת 1995 [12]. המערכת השתמשה בחמישה מאפיינים בינאריים לניקוד המשפטים: משפטים בעלי אורך גרול, משפטים המכילים מילים רומזות (כגון למסקנה, המסמך הזה), משפטים מפסקאות חשובות (10 ראשונות ו-5 אחרונות), משפטים המכילים מילים שכיחות ביותר, ומשפטים המכילים מילים רבות המתחילות באותיות גרולות. המערכת השתמשה בלמידה אוטומטית באמצעות חוקי bayes. במערכת זו הוענקה מירב תשומת הלב להצלחת המערכת באחזור (recall). המערכת השיגה 84% אחזור עבור סיכומים באורך 25% מגורל המאמר המקורי.

2. מערכת לסיכום אוטומטי של טקסטים באמצעות חילוץ משפטים שנבנתה על-ידי Mani ו-Bloedorn בשנת 1998 [16]. המערכת השתמשה ב-11 סוגי מאפיינים לניקוד המשפטים. על מנת לבצע למידה הכינה המערכת סיכומי ייחוס המבוססים על סיכומי המחברים באמצעות שימוש בפונקציית רמיון המבוססת בעיקר על כמות המילים המשותפות לשני המשפטים. במסגרת המערכת ניסו המפתחים 3 שיטות למידה שונות. שיטת C4.5 נמצאה כשיטה הטובה ביותר והגיעה ל-67% במדרד אחזור (recall) עבור סיכומים באורך 20% מגורל המאמר המקורי.

3. מערכת בשם SUMMARIST לסיכום אוטומטי של מאמרים בשש שפות שונות שנבנתה על ידי Lin בשנת 1999 [14]. המערכת נעזרה ב-14 שיטות שונות להערכת המשפטים. במסגרת המערכת נבחנה יעילות כל אחת מהשיטות השונות, כאשר את התוצאות הטובות ביותר השיגה שיטה המשלבת בין כל 14 השיטות באמצעות למידה שהתבצעה על-ידי עץ החלטה (decision tree). המערכת הגיעה לביצועים מיטביים בסיכום באורך 30%, שהשיג ציון F-measure (ציון משולב של אחזור וריק) של 46% [21].

4. מערכת לסיכום אוטומטי באמצעות חילוץ משפטים שבחנה יעילות שיטות למידה שונות שנבנתה בשנת 2002 על-ידי Neto ואחרים [21]. ברומה לשיטתם של Mani and Bloedorn [16] המערכת הכינה סיכומי ייחוס, תוך שימוש בפונקציית הרמיון הקוסינוסית. המערכת השתמשה ב-13 שיטות שונות להערכת המשפטים. התוצאות המיטביות הושגו באמצעות למידת Naive Bayesian, 41% אחזור עבור סיכומים באורך 10%, ו-52% אחזור עבור סיכומים באורך 20%. כאשר התבצעה הלמידה על בסיס סיכומי הייחוס אך נבחנו תוצאותיה ביחס לסיכומים ירי ארם (שבחרו משפטי ייחוס הרומים למשפטי סיכום המתבר), השיגה המערכת רק 27% אחזור עבור סיכומים באורך 10%, ו-39% אחזור עבור סיכומים באורך 20%.

ב. מערכות איחזור מידע מכתבים תורניים בשפה העברית

1. פרויקט השו"ת [4] של אוניברסיטת בר אילן <http://www.biu.ac.il/ICJI/Responsa/index.html>. הינו המאגר הנרחב ביותר של טקסטים ומאמרים תורניים. במסגרת שלב האיחזור

של פרוייקט השו"ת ישנו שימוש נרחב במאפייני השפה העברית, כגון: הוספת תחיליות דקרוקיות ו/או סופיות דקרוקיות ו/או תוספות כלליות ו/או יצירת הטיות.

2. מאגר תחומין [28]: (קבצים הלכתיים בנושא תורה חברה ומדינה) מאגר ממוחשב שכלל בתוכו את 22 הכרכים הראשונים של המאמרים שנכתבו, ואיפשר חיפוש לפי נושאים, לפי דב, לפי מילים, וכו'. מאמרים אלו עוסקים בנושאים מגוונים, כגון: כשרות, חברה ומדינה, מתים, שבת ומועדים.

ח. המודל המוצע

סוג הטקסטים שנבחרו לסיכום

המערכת שנבנתה התמקרה בסיכום של מאמרים שנבחרו מתוך המאגר הממוחשב של "תחומין" [28]. מאמרים אלו עוסקים בשאלות הלכתיות שונות שהתעודרו במהלך השנים בעקבות התפתחויות ותמורות שעברה החברה המוררנית בתחומים שונים. שאלות כאלו הן לדוגמה: האם חיות שאינן מופיעות בתורה כשרות? מתי אדם נחשב למת? האם הפרת זכויות יוצרים חשובה כגניבה?

מטרת המאמרים באסופה זו איננה רק לתת תשובות לשאלות הללו. כל תשובה צריכה להיות מבוססת על פוסקים קודמים ועל מקורות שונים. זאת ועוד, גם טענות ומענות אשר סותרות את תשובת המחבר צריכות להיות מובאות על-ידו, כמו גם רחיות ונימוקים מרוע טענות אלו אינן רלוונטיות. כל התשובות והטענות צריכות להיות מיושבות מבוססות ואמינות מבחינה הלכתית.

המערכת סיכמה כל מאמר שהובא לפניה בשלושה אופנים: תחום המאמר, מילות המפתח שבמאמר וסיכום מסקנתי של המאמר.

תחום המאמר

תחומי המאמרים הינם כאמור התחומים שהוגדרו על-ידי עודכי תחומין [28] (ראה פרק ב). המערכת שנבנתה הציעה כתחום המתאים למאמר את התחום שעבורו התגלה המספר הרב ביותר של מילות מפתח. בשלב קדם-עיבוד המערכת בנתה בצורה אוטומטית לכל תחום טבלה המכילה את מילות המפתח האופייניות לתחום הנידון כמתואר בפסקה הבאה.

מילות מפתח

המערכת שנבנתה הגדירה את מילות המפתח של מאמר מסוים מבחינה בסיסית באמצעות בנייה אוטומטית של טבלת שכיחויות של המילים והביטויים המופיעים במאמר. המערכת מצאה גם ביטויי מפתח המכילים יותר ממילה אחת, כגון: "פסוקי רזמרא". המערכת ניפתה מילים מסוימות אשר להן תפקידים מיוחדים, כגון: לכן, אם, יש, וכאשר (פירוט בנספח א). מילים אלו לא הוצגו כמילות מפתח על אף שחלק גדול מהן הופיעו בשכיחות גבוהה.

מילות המפתח שנמצאו לכל מאמר סייעו בבניית טבלאות עבור אחד משלושה עשר תחומי המאמרים, כל מאמר בהתאם לסיווגו על-ידי עורכי תחומין לתחום המתאים. כתוצאה מכך, נבנו שלושה עשרה טבלאות, אחת לכל תחום (רוגמה לטבלה אחת נמצאת בנספח ה). הטבלאות מכילות 660 מילות מפתח ובממוצע קרוב ל-51 מילות מפתח לכל תחום.

בניית סיכומי ייחוס

נבנתה אסופה של 60 מאמרים הלכתיים שנבחרו כאמור מתקליטור תחומין [28]. נבחרו מאמרים שעבור כל אחד מהם נמצא בסופו סיכום מעשה ידי מחבר המאמר עצמו. סיכומים אלו נעשו בידי המחברים של אותם מאמרים. סיכומי ייחוס עבור מאמרים אלה נבנו באמצעות התהליך שתואר בפרק ה. הסיכומים שהמערכת בנתה הושוו לסיכומים הללו כדי למרר את הישגיה. בתהליך זה השתמשו מערכות מספר 2 ו-4, שנידונו לעיל.

פיתוח שיטת השוואה חדשה בין משפטים

תהליך יצירת סיכומי הייחוס נדרש להשתמש בשיטת השוואה בין שני משפטים, על מנת שיוכל למצוא את המשפטים הרומים לסיכום הירני. שיטת ההשוואה המקובלת הפשוטה ביותר בין משפטים היא שיטת ה-cosine measure. שיטה זו מודדת רמיון בין שני משפטים s_1 ו- s_2 על-פי מספר המילים המשותפות המופיעות באופן זהה לחלוטין בשני המשפטים מחולק בממוצע אורך שני המשפטים.

אולם כאשר נעשה שימוש בשיטת השוואה זו לבניית סיכום הייחוס התגלה שסיכומים אלו לא היו טובים דיים. נראה היה שהסיבה לכך היתה שפונקציית ה-cosine measure אינה מתחשבת במספר גורמים משמעותיים, כגון: (1) אי התחשבות בהתאמה חלקית בין מילים (כגון: כתב ונכתב) ו-(2) אי התחשבות במידת החשיבות של המילים לנושא או לתחום המאמר. לכן פותחה שיטת השוואה חדשה אשר לקחה בחשבון בנוסף לפונקציית ה-cosine measure את שלושת הגורמים הבאים:

- מילים המופיעות בשני המשפטים וקשורות לתחום המאמר (נספח ה) קיבלו ציון התאמה גבוה יותר.
 - מילות פסיקה רומזות (נספח ד) המצביעות על פסק הלכתי המופיע במשפט (מותר, אסור וכד'), המופיעות בשני המשפטים קיבלו ציון התאמה גבוה יותר.
 - מילים רומזות מסקנתיות רגילות (נספח ב) אשר מראות על חשיבותו של המשפט (לסיכום, למסקנה, וכד'), המופיעות בשני המשפטים קיבלו ציון התאמה גבוה יותר.
- שיטת השוואה זו יצרה סיכומי ייחוס מוצלחים יותר ואף הביאה לשיפור הישגי המערכת באופן כללי.

בניית סיכום מסקנותי על ידי המערכת

פותחו שתי שיטות נוספות ייחודיות לשפה העברית בכלל ולתחום התורני הלכתי בפרט למציאת המשפטים החשובים ביותר המופיעים בטקסט:

שיטת סיכום 8: שיטת מילות תחום רומזות (DCW (domain cue words)

שיטה זו מצאה בתחילה את התחום שבו עוסק המאמר (ע"פ התהליך המתאים המתואר בתחילת פרק זה). לאחר שהמערכת מצאה את התחום שבו המאמר עוסק היא ספרה כמה ביטויי מפתח השייכים לתחום שנמצא בכל אחד ממשפטי המאמר (גם כן על-פי התהליך המתאים המתואר בתחילת פרק זה). לאחר מכן העניקה המערכת לכל משפט ניקוד באמצעות חלוקה של מספר ביטויי המפתח התחומיים במספר ביטויי המפתח התחומיים המופיעים במשפט המכיל את המספר הרב ביותר של ביטויים כאלו. בנספח ה ניתן לראות את טבלאות הקשר שבהן נעזרה המערכת כדי למצוא את התחום. טבלאות ערכים אלו מכילות 660 ביטויים.

שיטת סיכום 9: שיטת מילות פסיקה רומזות (PCW (psika cue words)

ברומה לשיטת המילות הרומזות הרגילה שהוצגה לעיל, גם שיטה זו ניקדה משפט על-פי מילים מסקנתיות שיש בו. אלא שמכיוון שאסופת המאמרים שנבנתה היא הלכתית, נבחרה שיטה המרגישה דווקא מילות פסיקה מסכמות, כגון: מותר, אסור, פטור, הלכה למעשה, צריך לומר, לכתחילה ו-דיעבד (רשימה מלאה בנספח ד). שיטה זו העניקה לכל משפט ניקוד באמצעות חלוקה של מספר מילות הפסיקה המופיעות בו במספר מילות הפסיקה המופיעות במשפט המכיל את המספר הרב ביותר של מילות הפסיקה.

ט. למידה אוטומטית לשיפור יכולת המודל

במטרה לשפר את תוצאות המודל נעשה שימוש בשיטות למידה שונות. כדי לבחון את שיטות הלמידה השתמשה המערכת בשיטת k-fold cross-validation הפועלת על-פי העיקרון הבא:

1. חלק באופן אקראי את אוסף המאמרים ל-k קבוצות שוות במספר המאמרים שבכל אחת מהן.

2. עבור $k = 1$ to k בצע ניסוי מס' i כרלקמן:

2.1 השתמש בקבוצה ה-i לבחינת הישג הלמידה אשר תבוצע על שאר (k-1) הקבוצות.

3. חשב את ההישג הממוצע של k הניסויים.

השתמשנו בשיטת ה-k-fold cross-validation עם $k=10$ עבור אוסף בן 60 המאמרים, כך

שבכל אחד מעשרת השלבים המערכת למדה באמצעות 54 מאמרים, ואחר כך ברכה את למירתה על 6 המאמרים הנותרים.

שיפור התוצאות על-ידי למידה בשיטת Genetic Algorithm

בפרוייקט מומשה מערכת למידה גנטית בעלת אלגוריתם מורחב במעט מהאלגוריתם שהוצג בפרק ב על-פי העקרונות של האלגוריתם הגנטי המכונה Genetic Steady State המתואר על-ידי Dejong [5]. האוכלוסיות התפתחו למשך 300 דורות. האופרטור מיווג (crossover) יצר 70% מאזרחי הדור החדש. מוטציות בוצעו על 1% של אזרחים. אלגוריתם זה מומש בעזרת פונקציות שנלקחו מספריית Galib [7], המציעה פונקציות המאפשרות הרצה של מגוון נרחב של אלגוריתמים גנטיים.

תהליך מציאת מילות המפתח על-ידי המערכת

1. עבור כל משפט s במאמר.
 - 1.1.1. עבור כל גורל ביטוי אפשרי (ער 3 מילים).
 - 1.1.1.1. עבור כל ביטוי במשפט s שאינו מופיע ב-stop list.
 - 1.1.1.1.1. אם הביטוי אינו מופיע ברשימת מילות המפתח – הכנס אותו לרשימה.
 - 1.1.1.1.2. אחרת הוסף 1 למספר ההופעות של הביטוי.
 2. סנן מרשימת מילות המפתח ביטויים המופיעים פחות מ-3 פעמים במאמר.
 3. סנן מרשימת מילות המפתח תתי-ביטויים המופיעים בתוך ביטויים גדולים יותר.
 4. מיין הרשימה לפי מספר ההופעות של כל ביטוי במאמר, והתאמתה לגורל הרצוי.

קדם-תהליך להגדרת התחומים האפשריים

1. בניית רשימת תחומים אפשריים למאמרים.
2. עבור כל תחום d מרשימת התחומים
 - 2.1. בניית רשימת מילים השייכות לתחום, ומידת חוזק הקשר לתחום (1-4).

תהליך מציאת התחום שאליו קשור המאמר על-ידי המערכת

1. עבור כל משפט s במאמר
 - 1.1. עבור כל ביטוי במשפט המופיע ברשימת מילות התחום, ערכן את ניקוד התחום לפי מידת חוזק הקשר בין הביטוי לתחום.
 2. תחום המאמר הוא התחום שקיבל את הניקוד הגבוה ביותר.

תהליך הכנת קטע המסכם את המאמר על-ידי המערכת

1. ניתוח המסמך וחלוקתו לפרקים, פסקאות ומשפטים.
2. עבור כל משפט s מגוף המאמר
 - 2.1. עבור כל מאפיין F_i של המשפט
 - 2.1.1. חשב את ניקוד המאפיין F_i תוך שימוש בבסיסי המידע.
 - 2.2. חשב את הניקוד המשוקלל של המשפט לפי ערכי המשקולות.

3. בחר את 10% המשפטים בעלי הניקוד המשוקלל הגבוה ביותר לקבוצת משפטי הסיכום.
4. מיין את קבוצת משפטי הסיכום לפי סדר הופעתם בטקסט.

י. תוצאות הניסוי

המערכת יצרה סיכום עבור כל אחד מ-60 המאמרים עבור כל אחת מ-9 שיטות הסיכום בתהליך שתואר לעיל בסוף הפרק הקודם. המערכת חישה את מירת הסיכוי של כל משפט להופיע בסיכום, על-פי כל אחת מ-9 שיטות הסיכום ובחרה את 10% המשפטים שקיבלו את הציון הגבוה ביותר.

כדי למצוא את המשקל של כל אחת מן השיטות בשקלול הסופי נבדקה יכולת ה-recall של כל שיטה באופן עצמאי. הנוסחה הבאה מתארת את קבוצת המשקלות ההתחלתיים עבור כל אחד מהמאפיינים $F_i: i = 1, \dots, n$ כאשר n הוא מספר המאפיינים, ו- $recall(F_i)$ הוא מירת הצלחת ה-recall של מאפיין i .

$$w_{recall}(F_i) = \frac{recall(F_i)}{\sum_{j=1}^n recall(F_j)}$$

לשם מדידת תוצאות הניסויים, כפי שהוסבר באריכות בפרק ב, המדר שהיה לנו חשוב הוא רווקא מדר ה-recall, כאשר אורך הסיכום היה תמיד 10% מאורכו של המאמר.

הישגי המערכת לפני ביצוע הלמידה

תוצאות ה-recall שהשיגה כל אחת מ-9 שיטות הסיכום ללא למידה מוצגות בעמורה המתאימה בטבלה מס' 1. על-פי תוצאות אלו ניתנו המשקולות למאפיינים השונים בשיטת הסיכום המשלבת בין השיטות ללא למידה.

נוסחה משוקללת לסיכום הורכבה בעזרת ערכי ה-recall המצויים בטבלה מס' 1 באופן הבא: כל תוצאה של מדר/שיטה הוכפלה במקרם שהיה ערך ה-recall שלה. כלומר, הנוסחה המשוקללת למדידת ערכו של משפט כלשהו S לצורך בחירתו האפשרית כמשפט סיכום היתה:

$$0.19 * TF-ISF(s) + 0.12 * SP(s) + 0.10 * PCW(s) + 0.12 * SP(s) + \dots$$

10% ממשפטי המאמר, בעלי הערך המשוקלל הגבוה ביותר נבחרו כמשפטי סיכום. המערכת הגיעה באמצעות נוסחה זו ליכולת recall של 0.39, גבוה בהרבה מעל התוצאה של השיטה העצמאית המוצלחת ביותר על-פי טבלה מס' 1 – שיטת Domain עם recall של 0.23.

הישגי המערכת לאחר למידת Genetic Algorithm

הרצת המערכת לאורך 300 דורות גנטיים הניכה תוצאת recall של 0.46. דהיינו, הלמידה הו שיפרה את התוצאות הראשוניות ב-0.07.

העמודה האחרונה בטבלה מס' 1 מכילה את המשקלות לאחר שעודכנו על-ידי למידת Genetic Algorithm. ניתן לראות שלאחר הלמידה נותר מספר מצומצם יותר של מאפיינים הרלוונטיים לסיכום. ניתן לראות שמאפיין מילות התחום הרומזות (DCW) מילות הגריל את כוחו כמעט בחצי והוא המאפיין החשוב ביותר עבור ההחלטה האם לכלול את המשפט בסיכום המאמר. מאפיין מיקום המשפט (SP) התחזק ב-שני שלישי, וכך גם מאפיין מילות הפסיקה הרומזות (PCW). מאפיין הדמיון לכותרת ראשית (TR) התחזק אף הוא והגריל כוחו פי 3. לעומתם, המאפיינים אורך משפט (SL) ודמיון לכותרת משנה (STR) איבדו את כוחם כמעט כליל. כמו כן חלה נסיגה בחשיבותו של המאפיין TF-ISF, למרות שהוא מוסיף להיות גורם חשוב יחסית (מקום שלישי).

טבלה 1: משקלות המאפיינים לפני ואחרי למידת GA

מס' מאפיין	מאפיין	לפני למידת GA	אחר למידת GA
1	TF-ISF	0.19	0.17
2	מילות רומזות רגילות (CW)	0.08	0.08
3	אורך המשפט (SL)	0.11	0.00
4	ניקוד שלילי (Negative)	0.03	0.03
5	מיקום המשפט (SP)	0.12	0.19
6	דמיון לכותרת משנית (STR)	0.11	0.01
7	דמיון לכותרת ראשית (TR)	0.03	0.10
8	מילות Domain רומזות (DCW)	0.23	0.32
9	מילות פסיקה רומזות (PCW)	0.10	0.15

יא. דוגמה לסיכומים שיצרה המערכת לאחר למידת GA

תחילת המאמר המקורי

הג'ראף — כשרותו לאכילה/ הרב אברהם המאמי [29].

"בפסוק 'זאת הבהמה אשר תאכלו... איל וצבי ויחמור ואקו ודישן ותאו וזמר' (דברים יד, ד-ה) תרגם רב סעדיה גאון: זמר — אלזראפה. זוהי הג'יראפה, שכן הוא שמה גם בערבית, וכפי שתיאר הרב עמרם קורה (בספרו נוה-שלום): 'מין חיה מבעלי הטלפים. יריו ארוכות ורגליו קצרות, ועורו מנומר כעור הנמר, וצוארו כצואר הסוס, אלא שהוא ארוך וקונף יותר. ויש לו שני קרניים קטנים'. רבים מהראשונים זיהו כך את הזמר. ר"ק מביא כך בשם ר' יונה אבן ג'נאח בספר השורשים, בשורש זמר...".

סיכומו של מחבר המאמר

- א. הג'יראף הוא, לפי מסורת וקבלת הגאונים, הזמר האמור בתורה.
- ב. כיון שיש לג'יראף סימני הטהרה הנחוצים — מעלה גרה, מפריס פרסה ושסע שסע שתי פרסות — הרי שרי בכך כרי להתירו באכילה, ואין צורך לרעת הרמב"ם, השו"ע והפריי-מגרים, במסורת כרי לאוכלו.
- ג. לרעת הש"ך ורבים מן האחרונים גם חיה אינה נאכלת אלא במסורת, ולכן, אין לקבל שום מין חרש להביאו על שולחן ישראל, "מפני גרר למאכלות אסורות אשר פרצה טהרה בישראל, ואין לנו לפרוץ גדר בזה" (לשון חזון-איש).

תחום המאמר ומילות המפתח

תחום המאמר זוהה נכונה על-ידי המערכת כ"מאכלות אסורים". המערכת למעשה זיהתה את התחום הנכון לגבי 59 מאמרים מתוך 60, סה"כ הצלחה של מעל ל-98%. מילות המפתח שזוהו מתוך המאמר הזה עפ"י האופן שפורט בפרק ג הרכיבו חלק גדול מן הטבלה של מילות מפתח שנבנתה עבור התחום "מאכלות אסורים".

סיכום הייחוס שיצרה המערכת על-פי סיכום המחבר

- (המשפטים המודגשים הם אלו שנבחרו גם בסיכום שיצרה המערכת בכוחות עצמה, שיוצג בהמשך)
1. רומה שמעיני האחרון נתעלמו כל אותם מקורות שמנינו לעיל, ולפיהם הזמר האמור בתורה ומנוי בין החיות הטהורות הוא הג'יראף.
 2. מתוך סימני הטהרה שנמנו בתורה לחיות, זכה הג'יראף והוא גם מעלה גרה וגם מפריס פרסה ושסע שסע שתי פרסות.
 3. הבאת הג'יראף בין החיות כאשר לדיני ניקור החלב גם היא אינה מוכיחה שהיתה לו מסורת שהג'יראף נאכל — כל שבא לומר הוא איזה מין טעון ניקור חלב, ואיזה אינו צריך, ותו לא מדי.

ניתוח סיכום הייחוס

ניתן לראות שעבור שני משפטי הסיכום הראשונים של המחבר הצליחה המערכת למצוא שני משפטים מגוף המאמר המכילים את תוכנם המרכזי. עבור המשפט הראשון מצאה המערכת משפט שאומר אף הוא שישנם מקורות לכך שהזמר האמור בתורה הוא הג'יראף. עבור המשפט השני מצאה המערכת משפט שמכיל את הנושא העיקרי במשפט המחבר של הג'יראף יש את סימני הטהרה הנחוצים: מעלה גרה ומפריס פרסה. אמנם עבור המשפט השלישי לא הצליחה המערכת למצוא משפט שמכיל את התוכן העיקרי, שצריך מסורת על מנת לאכול חיה (כגון הג'יראף), אך לפחות זיהתה משפט המדבר על כך שאין הוכחה שהיתה מסורת שהג'יראף נאכל. לפיכך ניתן לומר שהמערכת בנתה סיכום ייחוס שמכיל מעל שני שלישים מהתוכן העיקרי של סיכום המחבר.

סיכום באורך 10% שיצרה המערכת בעצמה

(המשפטים המורגשים הם אלו שנבחרו גם עבור סיכום הייחוס. המשפטים המוטטים הם משפטים המכילים תוכן מרכזי שהופיע בסיכום המחבר).

1. רר"ק מביא כך בשם ר' יונה אבן ג'נאח בספר השורשים, בשורש זמר.
2. וז"ל הרשב"ץ: "הרבה העירו לי על חיה גדולה, שהוא בלשון התורה זמר.
3. כך גם בהלכות ניקור (עמ' קע): "דע שבהמות הבר — היינו: הצבי והאיל והיחמור והיעל ואלארוא, ויש אומרים הכרכן והתותל והוראפה — אין בהם כל חלב אסור, ולא עורקים רמיים.
4. הבאת הגיראף בין החיות כאשר לדיני ניקור החלב גם היא אינה מוכיחה שהיתה לו מסורת שהגיראף נאכל — כל שבא לומר הוא איזה מין טעון ניקור חלב, ואיזה אינו צריך, ותו לא מידי.
5. ומצאתי וראיתי שאמת וצרק, שחיה טהורה מעלת גרה ומפרסת פרסה היא, וכל סימני טהרה בה.
6. המסורת המזהה את הזמר עם הגיראף לכאורה אינה עולה בקנה אחד עם מחלוקת האמוראים בחולין פ, א כאשר לזוהיים של עיזי דבאלא (עיזי יער).
7. והיא חיה טהורה המעלת גרה, ומפרסת פרסה שתי פרסות, השוכנות באפריקה.
8. אבל כיון שהמין המוכר לנו נוצר מהזורקקות התחש שהיה בימי משה למין גמל טמא, "נשתנה הבריה מתחש למין גמל".
9. מתוך סימני טהרה שנמנו בתורה לחיות, זכה הגיראף והוא גם מעלה גרה וגם מפריס פרסה ושוסע שסע שתי פרסות.
10. רהיטת הרברים, שמנהגנו שלא לאכול מין חדש על-ידי סימנים, כמו שכתב הרמ"א בעוף.
11. החזו"א ציין שרעה זו של הש"ך ובעל חכמת-אדם היא דלא כפרי-מגדים (שפ"ד פ, א) הסובר ש"רק עופות אין לאוכלם אלא במסורת אבל בהמה חזיה — כל שמכיר שמעלת גרה ומפרסת פרסה, סגי בהכי ואין צריך בקיאות לזה".
12. ראייה לתפיסה זו, שאין צורך במסורת על מיני החיה, יש להביא מרברי הרמב"ם (הל' מאכ"א א, ח), שכתב: "אין לך בכל בהמה חיה ועוף שבעולם שמתור באכילה, חוץ מעשרת המינין המנויין בתורה — שלשה מיני בהמה, והן: שור, שה ועז ושבעה מיני חיה: איל וצבי ויחמור ואקו ורישון ותאו וזמר".
13. הוויכוח על כשרות הגיראף הולך ונמשך עד ימינו.

ניתוח הסיכום האוטומטי של המערכת

מבחינת מרדי איחזור ורינק הרי שמאחר שהסיכום האוטומטי הכיל 2 מ-3 המשפטים של סיכום הייחוס הרי שבמדר ה-recall זכה הסיכום ל-67% הצלחה. לעומת זאת מאחר ורק 2 מתוך 13

משפטי הסיכום האוטומטי הופיעו גם בסיכום הייחוס הרי שבמדר ה-precision זכה הסיכום האוטומטי ל-15% הצלחה. אולם כפי שכבר צוין, מדרד האיחזור הוא החשוב יותר עבורנו ולכן הסיכום האוטומטי נחשב מוצלח.

ניתן לראות שהסיכום האוטומטי מכיל אכן את רוב התוכן המרכזי שמכיל סיכום המחבר. עיקר התוכן של המשפט הראשון של סיכום המחבר, שהג'ראף הוא הזמר לפי המסורת, במשפט מס' 6 בתחילתו. התוכן של המשפט השני של המחבר מוכל במשפט מס' 9 כמעט מילה במילה (וגם נבחר על-ידי סיכום הייחוס). התוכן העיקרי של המשפט השלישי של המחבר, שחיה נאכלת רק במסורת לפי הש"ך (תוכן שלא הובא כמעט בסיכום הייחוס), מוצג במשפטים מס' 10 ו-11. יתירה מכך, חלקם של משפטי סיכום המערכת (כגון: משפטים 6, 10 ו-13) מציגים אף פרטי מסקנות שאינם מצויים בסיכום המחבר.

בנוסף, המשפט הראשון של סיכום הייחוס נמצא באופן חלקי במשפטים 2 ו-6, כך שעל אף שערכו של מדרד האחזור הוא 67%, ערכו המעשי גבוה יותר. בסיכומו של רבר, בעזרת דוגמה זו ניתן לראות שבאמצעות סיכום אוטומטי שאורכו הוא רק כ-10% מאורך המאמר המקורי ניתן היה לקבל את עיקרי מסקנות המאמר. יוצא מכך, לעניות רעתנו שלפחות ברוגמה זו המערכת הצליחה יפה מאוד בסיכומה.

ניתן גם לראות מרוגמה זו שכפי שציינו גם Kupiec ושותפיו [12] שקשה למצוא בגוף המאמר משפט יחיד שיכיל בדיוק את התוכן של משפט סיכום המחבר, ולעתים קרובות סיכום מיטבי מגוף המאמר מצריך מספר רב יותר של משפטים כדי להכיל את התוכן של סיכום המחבר. לכן בדרך כלל יש צורך בסיכום אוטומטי בהיקף גדול יותר מסיכום המחבר על מנת למצוא אותה כמות המירע שמכיל סיכום המחבר, כשהמדרד העיקרי צריך להיות האיחזור ולא הדיוק. רוגמה זו מבליטה בעייתיות מסוימת של השימוש בסיכום הייחוס. למרות שכפי שהוסבר לעיל הסיכום האוטומטי מכיל את כל התוכן של סיכום המחבר, הרי שבהשוואה לסיכום הייחוס השיג הסיכום האוטומטי 67% אחזור בלבד. ולכן יש צורך למצוא פונקציית השוואה טובה יותר ליצירת סיכום הייחוס, במטרה שסיכום הייחוס יכיל יותר משני שליש מהמירע שמכיל סיכום המחבר, וניתן יהיה לקבל מדרדים מרוייקים יותר של הצלחת הסיכום האוטומטי.

כפי שהודגם, סיכומי הייחוס שהוכנו על-ידי המערכת אינם מכילים בהכרח את התוכן העיקרי של הסיכום שנכתב על-ידי המחבר. לכן, יש לבחון אפשרות של יצירת סיכומי ייחוס שייבנו ירנית על-ידי בני-אדם שיקראו את המאמרים, ויבחרו בעצמם את המשפטים הרלוונטיים ביותר, ובכך לקבל סיכומי ייחוס אמינים יותר. ראוי לציין כי רוב מערכות הסיכום האוטומטי אכן נברקות בצורה כזו.

ב. סיכום, מסקנות ומחקר עתידי

מאמר זה הציג מערכת סיכום אוטומטית ראשונה מסוגה בעברית בכלל ובתחום התורני-הלכתי בפרט. המערכת סיפקה מספר סוגי סיכומים למאמרים תורניים הלכתיים בעברית: תחום המאמר,

מילות מפתח וסיכום מסקנתי. הסיכום המסקנתי בוצע על-פי חילוץ המשפטים אשר נמצאו מתאימים ביותר לסיכום מסקנתי. המערכת גילתה יכולת למירה אוטומטית מוצלחת באמצעות אלגוריתם גנטי. יכולת הסיכום של המערכת התגלתה כשקולה לזו של מספר מערכות סיכום מקבילות בשפה האנגלית. אולם, צריך להודות כי יכולת זו עריין פחותה מיכולת הסיכום של מומחים אנושיים.

לכן, אב הטיפוס שנבנה איננו מצוי עריין ברמה הנדרשת עבור תלמידי-החכמים, אם כי בהחלט יכול לסייע לאלו שאינם מומחים בכתבים תורניים ואין זמן פנוי רב בידם. בכל אופן, לעניינות-רענתנו הפוטנציאל קיים ואף נראה מבטיח. מחקר עתידי רב במספר כיוונים בסיוע תלמידי-חכמים נחוץ לשם הבאת המערכת ליכולת מומחית.

מוצעים כיווני מחקר כלליים, כגון: ניסוח וניסוי שיטות סיכום נוספות, הן כלליות והן ייחודיות לשפה העברית ולתחום התורני, פיתוח שיטות למירה הן לבחירת משפטי סיכום והן לשקלול המאפיינים של פונקציית הרמיון בין משפטים, וכן שיפורים נוספים של פונקציית השקלול, הרחבה של המודל עבור כתבים בעברית במגוון תחומים אחרים, שיפור יכולת הניתוח המורפולוגית (=הרקדוקית) של המערכת, סוגי סיכום נוספים, כגון: סיכום איזכורים, אפשרות להשתמש בסיכומי ייחוס אשר נבנו על-ידי בן-ארם, סיכום של מספר טקסטים לטקסט אחד וסיכום טקסטים רב-לשוניים.

כיווני מחקר נוספים המוצעים בתחום הספציפי של המערכת (התחום התורני-הלכתי) הם: מתן חשיבות גבוהה יותר למשפטים אשר מצוטט בהם פוסק שהוא יותר בר-סמכא מפוסקים אחרים (כגון מתן עדיפות לגמרא על פני שו"ת, תלמוד בבלי על פני ירושלמי וכו'). בנוסף, ירוע כי מחברים מסוימים מתבססים בפסיקתם על פוסקים מסוימים (לדוגמה, פוסקים ספרדים בדרך-כלל מתבססים על פסקיו של מרן בעל השו"ע בעוד הפוסקים האשכנזים בדרך-כלל מתבססים על פסקיו של הרמ"א). ניתן לתת חשיבות גבוהה לפוסקים אשר מועדפים על מחבר המאמר עצמו. ניתן להציג את רשימת האיזכורים והציטוטים המופיעים במאמר, תוך התגברות על בעיות שונות באיתור ממוחשב של האיזכורים (לדוגמה, שינוי ספרותי של שם הספר: "הרמב"ן במלחמותיו כותב" במקום "במלחמות הרמב"ן כתוב"). ניתן להציג סטטיסטיקות של מחברים שונים בכיווני פסיקת ההלכה. הפוסקים המצוטטים ביותר, הפוסקים שלפיהם פסק המחבר הכי הרבה פעמים, וכן מיון הפוסקים המצוטטים לפי עדות, תקופות, וכו'.

1. Aone, C., Okurowski, M. E., Gorlinsky, J. and Larsen, B., "A Scalable Summarization System Using Robust NLP". In Mani, I. and Maybury, M. (Eds.), *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 66-73, 1997.
2. Bar Ilan, J. and Gutman, T., "How Do Search Engines Handle Non English Queries? – A Case Study", *Proc. of the Alternate Papers Track of the Twelfth International World Wide Web Conference*, 2003.
3. Barzilay, R. and Elhadad, M., "Using Lexical Chains for Text Summarization". In Mani, I. and Maybury, M. (Eds.), *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10-17, 1997.
4. Choueka, Y., "Full-text Systems and Research in the Humanities", *Computers and the Humanities*, 14: 153-169, 1980.
5. De Jong, K. A., "An Analysis of the Behavior of a Class of Genetic Adaptive Systems", Ph.D. Dissertation, Univ. Michigan, Ann Arbor, 1975.
6. Edmundson, H. P., "New Methods in Automatic Extraction". *Journal of the ACM* 16(2), pp. 264-285, 1969.
7. GALib: A C++ Library of Genetic Algorithm Components, <http://lancet.mit.edu/ga>.
8. Goldberg, D. E., *Genetic Algorithms in Search Optimization & Machine Learning*, Addison-Wesley, 1989.
9. HaCohen-Kerner, Y., "Automatic Extraction of Keywords from Abstracts". In Palade, V., Howlett, R. J. and Jain, L. C. (Eds.), *Proc. of the Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, Vol. 1, Lecture Notes in Artificial Intelligence 2773, pp. 843-849, Berlin: Springer-Verlag, 2003.
10. HaCohen-Kerner, Y., Malin, E. and Chasson, I., "Summarization of Jewish Law Articles in Hebrew". In Lyguard, K. (Ed.), *Proc. of the 16th International Conference on Computer Applications in Industry and Engineering*, pp. 172-177, 2003.
11. Hovy, E.H. and Lin, C.-Y., "Automated Text Summarization in SUMMARIST". In Mani, I. and Maybury, M. (Eds.), *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Massachusetts, 1999.
12. Kupiec, J., Pederson, J. and Chen, F., "A Trainable Document Summarizer". In Fox E.A., Ingwersen P. and Fidel R. (Eds), *Proc. of the 18th Annual International ACM SIGIR*, pp. 68-73, 1995.
13. Lin, C.-Y., "Training a Selection Function for Extraction". *Proc. of the 8th International Conference on Information and Knowledge Management (CIKM 99)*, pp. 1-8, 1999.
14. Lin, C.-Y. and Hovy, E.H., "Identifying Topics by Position". *Proc. of the Applied Natural Language Processing Conference (ANLP-97)*, 283-290. Washington, 1997.

15. Luhn, H. P., "The Automatic Creation of Literature Abstracts". *IBM Journal of Research and Development*, 2(2):159-165, 1958.
16. Mani, I. and Bloedorn, E., "Machine Learning of Generic and User-Focused Summarization". *Proc. of AAAI-98*, pp. 821-826, 1998.
17. Mani, I. and Maybury, M. T., "Introduction", In Mani I. and Maybury M. (Eds.), *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Massachusetts, pp. 821-826, 1999.
18. Mitchell, T. M., *Machine Learning*, McGraw-Hill, New-York, Second Edition, 1997.
19. McKeown, K. and Radev, D. R., "Generating Summaries of Multiple News Articles". In Fox E.A., Ingwersen P. and Fidel R. (Eds.), *Proc. of the 18th Annual International ACM SIGIR*, pp. 74-82, 1995.
20. Myaeng, S. and Jang, D., "Development and Evaluation of a Statistically Based Document Summarization System". In Mani and Maybury (Eds.), *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Massachusetts, 1999.
21. Neto, J. L., Freitas, A. A. and Kaestner, C. A. A., "Automatic Text Summarization Using a Machine Learning Approach". In Bittencourt G. and Ramalho G.L. (Eds.), *Proc. of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence*, pp. 205-215, 2002.
22. Neto, J. L., Santos, A. D., Kaestner, C. A. A., Freitas, A. A. and Nievola, J. C., "A Trainable Algorithm for Summarizing News Stories". In Zaragoza H., Gallinari P. and Rajman M. (Eds.), *Proc. of the PKDD-2000 Workshop on Machine-Learning and Textual Information Access*, Lyon, France, 2000.
23. Radev, D. R., Language Reuse and Regeneration: Generating Natural Language Summaries from Multiple On-Line Sources, PhD thesis, Department of Computer Science, Columbia University, New York, 1999.
24. Zechner, K. A., "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences". *Proc. of the 16th international Conference on Computational Linguistics*, pp. 986-989, 1996.
25. א' אבן שושן, המילון החדש, המהרורה המשולבת, הוצאת קרית ספר בע"מ, ירושלים, 1983.
26. ש' אשכנזי ור' ירון, אוצר ראשי תבות, הוצאת קרית ספר בע"מ, ירושלים, 1994.
27. י' שוויקה, מילון רב מילים, מילון מקיף לעברית מודרנית, 1997.
28. עורכים: הרב י' שביב, הרב י' רוזן, הרב א' דסברג, הרב א' ורהפטיג, תחומין כרכים א-כ, הוצאת "צומת" – צוותי מדע ותורה, אלון שבות, גוש עציון.
29. א' תמאמי, הג'ראף – כשרותו לאכילה. עורך: י' שביב, תחומין, הוצאת "צומת" – צוותי מדע ותורה, אלון שבות, גוש עציון, כרך כ, עמ' 89-93, 2000.

נספח א: רשימת stop list

רשימת ה-stop list מכילה 243 מילים אשר להן תפקיד דקרוקי בשפה העברית. מילים אלו כפי שהוזכר בגוף המאמר אינן יכולות לשמש כמילות מפתח.

א'	בהם	הטי"ז	וכתב	כלל	מפני	רק
אבל	בו	הי'	ולא	כמו	מצד	שאין
אבל לא	בזה	היא	ולכן	כן	משום	שאינה
או	בין	היה	ועי'	כפי	סייס	שאינו
אומרים	בכל	היו	ועייע	כשהוא	סייק	שאינם
אותה	בלא	הכי	ועייש	כשהתירו	סגי	שאם
אותו	בנייד	הלי'	ועוד	כתב	סיי	שאמר
אותם	בסייד	הלשון	ועיי	לא	סימן	שאמרו
אז	בפני	הם	ועיין	לאחר	עייב	שגם
אח"כ	בשביל	הן	ושכן	לה	עיי	שהוא
אחא	בשו"ת	הנייל	זי'	להם	עייכ	שהיא
אחד	בתשוי	הני	זייל	לו	עיימ	שהם
אחר	גי'	הרי'	זאת	לומר	עייש	שהרי
אחרים	גי"כ	הרי	זה	לי	עד	שהתירו
אחת	גם	השלחן	זו	ליי	עוד	שום
איך	ד'	ו'	זכיי	לידי	עייש	שיהיה
אין	דבר	ואי"כ	ח'	ליה	על	שייך
אינה	דברי	ואין	חאו"ח	לך	על ידי	שיש
אינו	דהא	ואינו	טי'	לכל	עליו	שכי
אך	דהכי	ואם	טוב	לכתחלה	עם	שכל
אלא	דוקא	ואף	ידי	לנו	עמהם	שכן
אלו	דידן	ואפילו	יהיה	לענין	עפיי	שכתב
אם	דלא	ובשו"ת	יוכל	לפי	עצמו	של
אמנם	דלדעת	וגם	יותר	לפני	פייד	שלא
אמר	דמתני'	ודאי	יש	מ	פירוש	שם
אני	דעת	והנה	כי	מ"מ	פסוק	שמה
אף	דף	והרב	כאן	מדוע	פסוקים	תנא
אפי'	ה'	וזה	כבר	מה	פסק	
אפילו	הא	ויש	כדי	מהני	פעם	
אפשר	האם	וכי"כ	כוי	מותר	צריך	
אשר	הגאון	וכבר	כי	מטעם	קשה	
את	הדבר	וכי'	כיון	מי	ראוי	
ב'	הוא	וכמ"ש	כך	מיד	ראיתי	
בגמ'	הואיל	וכן	כל	מידו	רבתינו	
בגמרא	הוי	וכן כתב	כל שכן	מכיוון	רבי	
בה	הזה	וכן פסק	כלום	מן	רבינו	

נספח ב: רשימת מילים רומזות מסקנתיות (cue words)

נאספו 61 מילים אשר נוכחותן במשפט מסוים מעידה על חשיבותו למסקנה.

אין כך פני הדברים	יש לקבוע	מכאן יש ללמוד
אין לו על מה שיסמוך	יש מקום להקל	מסיבה זו
אין ספק	יש שהקלו	מסקנת הדברים
אכן	יתר על כן	משמע
אם כן	כאמור	נוכחתי פעמים רבות
בוודאי	כיום	ניתן ללמוד
במאמר זה	לא נתבארו לי	נמצא איפוא
ברור	לאור האמור	נעמוד
הדעה העיקרית	לאור כל הנ"ל	נפסק
הדרך הטובה ביותר	לכאורה	נפסקה
ההיתר מבוסס	לכן	נראה
העולה לענייננו	למיטב ידיעתי	סברא גדולה
העולה מדבריו	לפי דבריו	סניף להקל
וכן הכריע	לפי זה	סניף להתיר
ועדיין	לפייז	על כרחך
זאת ועוד	לפיכך	עליו הראיה
יוצא א"כ	מאמרנו עוסק	עם כל זאת
יש לדון	מבחן המציאות	פתח היתר
יש להסיק	מדבריו	צ"ל
יש להעיר	מכאן	צריך לומר
יש להתיר		

נספח ג: רשימת ביטויים/מילים "שליליים"

נאספו 31 ביטויים/מילים אשר נוכחותם במשפט מעידה באופן מסוים על חוסר החשיבות שלו ואי-הרלוונטיות שלו למסקנה.

א"כ	עולה מדבריו	לפי זה
אולם	והלא	לפייו
אין כאן מקום להאריך	וכן כתב	מ"מ
אכמ"ל	ידוע	מכאן
אם כן	יצויין	מכל מקום
אמנם	יש לציין	מצאתי
אפשר	כמי"ש לעיל	עכ"פ
אריכות	כמו שכתב לעיל	על כל פנים
גם כתב	למשל	ראיה לכך
דוגמא	לעתים	ראיתי
דוגמה		

נספח ד: רשימת מילות פסיקה רומזות

נאספו 102 ביטויים אשר נוכחותם במשפט מסוים מעידה עליו שיש לו משמעות מבחינה הלכתית.

א"א	התר	להלכה	נפסקה
אזלין בתר	היתר	להתיר	נקבע
אין לאסור	הלכה למעשה	לכאורה	נראה לאסור
אין להתיר	הקל	לכתחילה	נראה להתיר
אין לחשוש	התיר	למיטב ידיעתי	ניתן להכשיר
איסור	התירו	לנהוג	ניתן ללמוד
איסור מוחלט	וכן הכריע	לעני"ד	סברא גדולה
אנוס	זכינו לקיימו	לפי"ז	סברו
אסור	חובה	לפי זה	סוב
אסורה	חייב	לפי"ז	סוברים
אסורים	חייבות	לרוב הדעות	סניף להקל
אסר	חייבים	מבטל	סניף להתיר
אסרו	יכשר	מותר	על כרחך
בזה"ז	ינהג	מותרים	עליו הראיה
בזמננו	ינהגו	מותרת	פוסלים
גזר	יש לאסור	מכשירים	פטור
גזרו	יש לדון	מנהגנו	פטורים
דיעבד	יש להתיר	מעיקר הדין	פטורות
הדעה העיקרית	יש לחשוש	נאסר	פתח היתר
הדרך הטובה ביותר	יש לעיין	נאסרה	צי"ל
ההיתר מבוסס	יש לקבוע	נאסרו	צריך לומר
הותר	יש מקום להקל	ניתן ללמוד	ראוי לגדור
היתר	יש צד	נמצא איפוא	רוב דעות
הותרה	יש שהקלו	נעמוד	רצוי
הותרו	כהלכה	נפסק	רשאי
החמיר	לאסור		

נספח ה: טבלאות הקשר לפי נושא המאמר

נספח זה מכיל דוגמה לטבלת הקשר של מילים הקשורות לתחום המאמר (domain cue words), במקרה דנן "תחוקה ושלטון" עבורו נאספו 30 ביטויים/מילות מפתח. עבור כל אחד משלושה עשר התחומים השונים בהם עוסקים המאמרים ישנה טבלה המכילה מילים הקשורות לתחום, וכן את מידת חוזק הקשר של המילים לתחום (1-4, הערך 4 מוענק למילים המוערכות כקשורות ביותר לתחום). כל הטבלאות עבור כל התחומים בהם טיפלה המערכת מכילות כסה"כ 660 ביטויים/מילות מפתח.

מילות מפתח עבור התחום "תחוקה ושלטון"

מילה	ערך	מילה	ערך	מילה	ערך
אחרי רבים להטות	2	מלוכה	3	סמכות	2
אחריות	2	מלך	3	סנהדרין	2
דמוקרטי	4	מלך ישראל	4	עולם	1
דמוקרטיה	4	מלכות	3	עקרונות	2
דמוקרטיה	4	מלכי ישראל	4	ערכים	2
יהודית	2	מלכים	3	רוב	2
ירושלים	1	מנהיג	3	רשות	3
לבחור	2	משטר	4	שוטר	2
ליברלית	4	משפט	2	שוטרים	2
מדינה	3	נשיא	3	שלטון	4