זיהוי ציטוטים מהתלמוד הבבלי במקורות היהודיים

הקדמה:

משימת זיהוי ציטוטים מהתלמוד הבבלי במקורות היהודיים היא משימת ניתוח טקסט לא פשוטה אך חשובה. אם נוכל לבצע אותה בצורה טובה ויעילה, מכאן שנוכל לקשר טקסטים מקוונים של המקורות היהודיים עם המקורות בתלמוד ובגמרא מהם הם מצטטים. באופן זה יוכל הקורא להבין בנקל במה הטקסט עוסק, יוכל לעבור בקלות בין ציטוט למקור, יוכל לראות כיצד מקושרים חלקים שונים של המקורות היהודיים אחד לשני, יוכל לראות סטטיסטיקות מתקדמות העוסקות בשכיחות ציטוטים, ציטוטים הנפוצים ביותר אצל רבנים או מחברים מסוימים, ועוד.

במסמך זה נציג תחילה את האלגוריתם המקורי, כפי שתכננו אותו בהתחלה, ונסביר במפורט על כל אחד משלבי האלגוריתם. לאחר הסברים אלה נציג בכל פעם בעיה בה נתקלנו, או ציטוט אותו האלגוריתם בצורתו זו לא יכול היה לתפוס, או ציטוט לא נכון אותו האלגוריתם תפס, ואת הפתרון שהצענו ומימשנו למקרה זה. לבסוף נציג בעיות פתוחות ומקרים איתם לאלגוריתם עדיין קשה להתמודד.

דרך פעולה כללית:

אלגוריתם זיהוי הציטוטים עובד, בבסיסו, כך:

1. בנה אינדקס (מבוסס Lucene) של התלמוד הבבלי
2. פרמט את הטקסט. שנה את הטקסט מתוכו עלינו למצוא ציטוט כך שאינו יכיל תווים ייחודיים כמו אחוזים, מקפים, קווים נטויים וכו'.
3. תייג כל NGram באורך MINIMAL_NGRAM_LENGTH אם קיים לו מקור מתאים באינדקס.
4. הסר תגיות המצביעות לתוספות לתלמוד שאינן חלק מהתלמוד המקורי (לדוגמא התוספות של רש"י).
5. מזג כל NGrams חופפים בעלי אותו תיוג כך שהתיוג המשותף יופיע במסגרת הגדולה ביותר האפשרית (כלומר ב-NGram הרחב ביותר).
6. הסר תגיות ב-NGrams פנימיים שאינם מקסימלים.
7. הסר תיוגים מהתאמות פאזי שההבדל ביניהן אינו באות אהו"י.
8. מזג כל שני NGrams בעלי תיוג זהה אם המרחק ביניהם הוא עד MAXIMAL_HOLE_SIZE מילים, כך של-NGram המתחיל במילה הראשונה של זה ומסתיים במילה האחרונה של השני יהיה את התיוג המשותף.
9. הסר תיוגים מכל NGram באורך פחות QUOTE_PROBABLY_TOO_SHORT_VALUE ומטה.
10. החזר כקלט את כל התגיות שנותרו.

הסבר מפורט על כל שלב בנפרד:

בניית אינדקס JbsBavliIndex:

כאן השתמשנו באינדקס של Lucene, כלומר באותו כלי איתו עבדנו בזיהוי ציטוטים מהתנ"ך. האינדקס של Lucene (ממנו המחלקה JbsBavliIndex יורשת) מאגד תחתיו את כל הטקסט של התלמוד הבבלי, הנמצא בפורמט Json כחלק מהפרוייקט. האינדקס מאפשר חיפוש טקסט מדויק או פאזי במהירות וביעילות, ולכן בחרנו להשתמש בכלי זה. כל חיפוש טקסט בהמשך האלגוריתם נעשה מתוך אינדקס זה אותו יצרנו.

פרמוט טקסט BavliNgramFormatter:

בטקסט ממנו אנו רוצים למצוא ציטוטים עלולים להופיע סימני פיסוק רבים שאינם מופיעים בטקסט המקורי בתלמוד הבבלי ועלולים למנוע מאיתנו לזהות התאמות. לדוגמא, ייתכן והמילה האחרונה בטקסט מתוכו אנו רוצים למצוא ציטוטים היא "ברכות.", כשבתלמוד הבבלי מופיעה המילה המתאימה "ברכות" מבלי הנקודה בסוף. איך שהמילים מופיעות בצורתן הגולמית לא נמצא התאמה מדויקת, ולכן נרצה להיפטר מהנקודה.

סיבה נוספת שלא נרצה סימני פיסוק בטקסט בו אנו מחפשים ציטוטים היא לצרכי בדיקת האלגוריתם. כשאנו מכניסים טקסט ובו אנו יודעים היכן נמצאים הציטוטים המדויקים, לצרכי בדיקה, אנו מסמנים כל ציטוט בתחילתו ובסופו ב-%. גם סימנים אלה נרצה למחוק מהטקסט לפני שנתחיל במלאכת החיפוש.

צורת פרמוט שלישית שנרצה לעשות היא שינוי שם אלוהים כפי שהוא מופיע בהרבה מהמקורות היהודיים לשם אלוהים כפי שהוא מופיע בתלמוד הבבלי, מהסיבה שכך נוכל למצוא את הטקסט בחיפוש מדויק ולא נאבד מילים בגלל חוסר התאמה.

באופן זה ומשיקולים דומים אנו נפטרים מסימני פיסוק רבים, מספר דוגמאות: @ \\ : \ % ; . ,.

תיוג BavliTagger:

בשלב זה אנו מתאימים לכל 3-gram (MINIMAL_NGRAM_LENGTH) את התגיות המתאימות לו. עבור כל שלישיית מילים צמודות בטקסט אנו מחפשים התאמה לשלישיית מילים סמוכות בתלמוד הבבלי, כאשר אנו מרשים מרחק לווינשטיין 1 או 2 (כלומר פאזי 1 או 2) כתלות באורך המילה.

באופן זה אנו מוצאים התאמות רבות לכל 3-gram, רובן הגדול רועשות, אך חלק קטן מהן נכונות. שיטת העבודה של האלגוריתם, שהתגלתה כמובילה לתוצאות הטובות ביותר מבין השיטות שנוסו, היא למצוא תוצאות רבות מאוד בהתחלה, מלאות ברעש אבל גם בתקווה במקורות הנכונים, ואז לסנן את הרעש בהדרגה. שיטת עבודה זו הובילה לתוצאות טובות יותר משיטת עבודה שהגבילה בצורה מחמירה יותר מההתחלה, שכן שיטות כאלה פספסו תוצאות נכונות רבות.

הסרת תגיות ממקורות אחרים EliminateRashiTosafotRashbam:

שלב טכני בלבד.

בזמן כתיבת מסמך זה הפרוייקט מכיל, בנוסף לתלמוד הבבלי עצמו, תוספות לתלמוד של רש"י וחכמים אחרים. כברירת מחדל האינדקס נבנה כך שחיפוש הטקסט יהיה גם מתוך התלמוד הבבלי וגם מתוך תוספות אלה. שלב זה מסיר את התגיות הלא רצויות, כלומר אלה שאינן מהתלמוד הבבלי המקורי.

שלב זה אמנם לא לוקח זמן רב, והוא אינו מאיט את פעולת האלגוריתם, אך הוא נכון פחות תכנותית מהאפשרות לברור מההתחלה ולבחור לא לבנות את האינדקס כך שיכלול גם את המקורות הנוספים האלה.

שלב זה יוכל להיות מוסר אם האינדקס יוכל להיבנות מההתחלה ללא התוספות לתלמוד הבבלי.

מיזוג NGrams חופפים MergeToMaximalNgrams:

רבים מן הציטוטים אותם אנו מחפשים מתוך הטקסט אינם באורך 3, אלא ארוכים יותר. התיוג הראשוני מתבצע על 3-grams אך זהו תיוג ראשוני בלבד. כעת, נרצה לאחד תיוגים של NGrams חופפים בעלי תיוג זהה, כך שה-NGram המאוחד יכיל את אותו התיוג.

דוגמא: נאמר שמצאנו כי 3-gram 22-23-24 (המילים ה-22,23,24 בטקסט) מכיל את התגיות:

1. bavli-1-1-1
2. bavli-13-5-2

וה-3-gram ה-23-24-25 מכיל את התגיות:

1. bavli-1-1-1
2. bavli-23-2-1

במקרה זה אנו מבין שה-4-gram 22-23-24-25 הוא ציטוט מתוך bavli-1-1-1, ונוסיף לו את תגית זו. נשים לב כי 4-gram זה, כמו גם כל NGram באורך גדול מ-3 לא הכיל שום תגית עד עכשיו, שכן תייגנו בשלב הראשוני רק 3-grams.

שלב זה ממזג באופן מקסימלי, כלומר לא רק מ-3-grams ל-4-grams, אלא עד לגודל המקסימלי האפשרי של מילים להם יש תגית משותפת.

לאחר מזוג זה מצאנו כבר את הציטוטים המקסימלים המופיעים במלואם בתלמוד הבבלי, והשלבים הבאים יהיו או סינון רעשים או מיזוגים של ציטוטים סמוכים שאינם מתחברים לציטוט אחד ארוך, אך יש מן ההגיון למזג אותם בכל מקרה.

הסרת תגיות מ-NGrams פנימיים RemoveTagsInContainedNgrams:

לאחר שמיזגנו את כל ה-3-grams כך שה-NGram המקסימלי יכיל את התגיות המתאימות, אין לנו כבר צורך שכל 3-gram, או כל n-gram (n<N) המוכל בתוך ציטוט גדול יותר יכיל תגיות כלשהן. לכן אנו מסירים את כל התגיות מכל ה-NGrams המכילים תגיות אשר מופיעות גם ב-NGrams אשר מכילים אותם.

לדוגמא אם התגית bavli-1-1-1 מופיעה גם ב-3-gram 5-6-7, וגם ב-4-gram 4-5-6-7, נסיר את התגית מ-3-gram 5-6-7, כיוון שתפסנו כבר את הציטוט הרחב יותר, ואין לנו צורך שה-3-gram יהיה מתויג, זה יהיה הרי רק רעש עבורנו.

הסרת תיוגים פאזי שאינם נבדלים מהמקור באות אהו"י BavliRemoveNonEhevy:

סינון רעשים אחרון בו השתמשנו במקרה זה הוא סינון רעשים בו גם השתמשנו כאשר חיפשנו ציטוטים מתוך התנ"ך.

כאשר אנו מחפשים התאמות בין מילים אנו מרשים פאזי 1 בדרך כלל, כלומר הבדל של אות אחת בין מילת המקור למילת היעד. קריטריון זה הוא הגיוני, אך כמעט בכל המקרים בהם אנו מוצאים התאמה הגיונית ההבדל הוא באות אהו"י.

לדוגמא, הגיוני מאוד שבטקסט המקורי יהיה כתוב "ויאמר" ובטקסט המצטט יהיה כתוב "ואמר", אך אין הגיון כלל שבטקסט המקורי יהיה כתוב "מיד" אך בטקסט המצטט יהיה כתוב "מיץ", למרות שבשני המקרים מדובר במרחק של אות אחת בין שתי המילים.

לכן, כשאנו מרשים חיפוש פאזי בלא שום תנאים נוספים אנו מקבלים על עצמנו הרבה רעש. אנו מקבלים התאמות רבות שאין ביניהן קשר רב לטקסט המקורי.

מסיבה זו הגיוני לא לקבל התאמות פאזי שחוסר ההתאמה אינו באות אהו"י, כי במקרה זה סיכוי גדול שמדובר בהתאמה שאינה נכונה.

מיזוג NGrams סמוכים FinalMergeTags:

בשלב זה טמון הבדל עיקרי בין מציאת ציטוטים מתוך התנ"ך לבין מציאת ציטוטים מתוך מקורות יהודיים נוספים כגון התלמוד הבבלי. כאשר חכמים מצטטים פסוק מהתנ"ך הם מצטטים אותו בדרך כלל כמשפט רצוף ולא מקוטע. אולי הם מחסירים או מוסיפים אותיות אהו"י, אך הציטוט בדרך כלל מופיע כולו כמו שהוא, ללא הפסקות וחידושים.

כאשר חכמים מצטטים מהתלמוד הבבלי לעומת זאת, לעתים קרובות הם מצטטים כמה מילים, רושמים כמה מילים אחרות, ולאחר מכן שוב מצטטים מאותו מקום, ואז שוב רושמים כמה מילים שאינן ציטוט וחוזר חלילה.

דוגמא מתוך טקסט של אחד מפירושי הרמב"ם:

שם (דף י"א ע"ב) אמר רב יהודה אמר שמואל אהבה רבה וכו' וכו' ורבנן אמרי אהבת עולם וכן הוא אומר וכו' תניא נמי הכי אין אומרים אהבה רבה אלא אהבת עולם ע"כ כך היא גירסת הרי"ף ז"ל בהלכות.

המילים המסומנות בירוק הן ציטוטים מדויקים, המילים הצהובות הן התאמות פאזי, ומילים שאינן מסומנות אינן התאמות. כאן ניתן לראות שפירוש הרמב"ם מצטט מתוך התלמוד הבבלי, אך הציטוט אינו רצוף. הוא מצטט מספר מילים, ואז מסיים את הציטוט עם "וכו'", שלאחריו הוא מצטט עוד כמה מילים שלאחריהן הוא שום רושם "וכו'", ושוב עוד כמה מילים שהן ציטוט.

היינו רוצים לתפוס את כל הטקסט המסומן מתחילתו ועד סופו כציטוט אחד, אך עד עכשיו לא נוכל לעשות את זה, בגלל הציטוטים הקטועים.

לכן אנו צריכים מנגנון מסוים של מיזוג ציטוטים סמוכים.

בשלב זה האלגוריתם ממזג כל זוג NGrams בעלי תגית משותפת אם המרחק בין המילה הראשונה של זה והמילה האחרונה של זה אינו עולה על MAXIMAL_HOLE_SIZE. כך, אם יש שני ציטוטים סמוכים מאותו מקום בתלמוד, המופרדים על ידי לא יותר מ-MAXIMAL_HOLE_SIZE מילים, הציטוט המאוחד, שמתחיל בתחילת זה ומסתיים בסיום זה, יקבל את התגית המשותפת הזו, וכך נתפוס גם מקרים כאלה.

גישה זו מתאימה הרבה יותר לזיהוי ציטוטים מהמקורות היהודיים שאינם התנ"ך, כיוון שכאמור ציטוטים מהתנ"ך הם רציפים במקרים רבים.

הסרת תיוגים של NGrams קצרים מדי RemoveLowLengthMatches:

בשלב זה תפסנו בתקווה את הציטוט המורחב אותו היינו צריכים לתפוס בעזרת שיטות העיבוד והמיזוג בהן השתמשנו, אך יש בידנו עדיין הרבה רעש. עדיין קיימים תיוגים רבים של 3-grams או 4-grams שלא נוכל לקחת ברצינות. לציטוטים באורך קטן או שווה ל-QUOTE_PROBABLY_TOO_SHORT_VALUE (שערכו בזמן כתיבת מסמך זה הוא 4) אין כל משמעות, ולא נוכל להרשות לעצמנו להציג כל ציטוט באורך כזה כתוצאה הגיונית.

לציטוטים באורך 5 יש כבר יותר משקל, וסיכוי נמוך יותר שנמצא ציטוט באורך 5 שאין לו באמת קשר לטקסט. לכן החלטנו על סף. נתעלם מציטוטים באורכים עד 5 לא כולל, כלומר אורכים 3 או 4, ונזרוק אותם כתוצאות לא מהימנות. ציטוטים באורכי 5 ומעלה נוכל לקחת אותם כציטוטים נכונים.

סף זה הושג על ידי ניסוי וטעיה כמספר אופטימלי בגבולות הדוגמאות בהן התנסינו. אמנם תופסים מדי פעם ציטוט לא נכון באורך 5 או אפילו 6, אך ההקרבה של ה-Precision היא קטנה לעומת ההקרבה של ה-Recall שתתקבל אם לא נסכים לקבל ציטוטים אפילו באורכים אלה.

<u>תוספות ותיקונים אחרי בניית האלגוריתם:</u>

כעת נציג בעיות בהן נתקלנו וכיצד פתרנו אותן. בעיות שכותרתן מסומנת שאדום הן בעיות שהפתרונות המוצעים להן מושחרים נכון לזמן כתיבת מסמך זה בהערה, כלומר מימושם קיים אינו רץ כחלק מן האלגוריתם מסיבות אחרות.

<u>בעיה ופתרון:</u>

הרבה פעמים במקורות היהודיים מחליפים את המילה מחליפים את המילה **רבי** בקיצור **ר'**. לכן אם הציטוט מתחיל ב:"**ר' יהושע אומר**" אבל בגמרא המקור מתחיל ב:"**רבי יהושע אומר**", האלגוריתם לא יתפוס את המילה הראשונה.

כדי לטפל בכך נוסיף לפורמטר של האלגוריתם את ההחלפה: **ר' ← רבי.**

<u>בעיה ופתרון:</u>

בהרבה ציטוטים מחליפים את צמד המילים **"בית שמאי"** בגמרא בראשי התיבות **ב"ש**, וכך ציטוט שמתחיל, נניח ב-**"בית שמאי אומרים..."** לא ייתפס אם המצטט מתחיל ב-**"ב"ש אומרים..."**.

לדוגמא בציטוט הבא מתוך משנה תורה, המצטט מתוך 2-10-1: "ב"ש אומרים כל אדם קורין כדרכן שנאמר ובלכתך בדרך א"כ למה נאמר בשכבך ובקומך בשעה שדרך בני אדם שוכבים ובשעה שבני אדם עומדים".

כדי לטפל בכך נוסיף לפורמטר את ההחלפה **ב"ש ← בית שמאי**

בעיה ופתרון:

לעיתים ההגבלה של להסיר התאמות פאזי עם הבדל שאינו באות אהו"י מחמיר מדי, והאלגוריתם מסיים מבלי למצוא אפילו התאמה אחת למרות שהיו בהחלט צריכות להיות התאמות, לדוגמא אם מורידים התאמות פאזי בטקסט הבא מתוך משנה תורה:

```
.וכו לקרות ומקדימין :בסמוך שאכתוב המשנה מתוך מתבאר כך .'וכו קורא הוא ומה  משנה כסף
מלכות עול עליו שיקבל כדי לוהיה שמע פרשת קדמה למה ריב"ק אמר% (.ג"\י דף) דברכות פ"ב
ביום בין נוהג שמוע אם שוהיה לויאמר שמוע אם והיה מצות עול עליו יקבל כך ואחר תחלה שמים
%ביום אלא נוהג אינו ויאמר בלילה בין
```

האלגוריתם נשאר בלי תגיות ברגע שמורידים התאמות פאזי שאינן באות אהו"י.

כדי לפתור בעיה זו נגדיר שני מצבים בהם ניתן לחפש התאמות. מצב אחד בו מורידים התאמות פאזי שלא באות אהו"י כפי שהיה עכשיו, ומצב שני בו לא עושים זאת. אלגוריתם חיפוש הציטוטים מתחיל בלבצע את התהליך כאשר מריצים אותו כמו שעשינו עד עכשיו. אם מצאנו התאמה לפחות אחת, יופי, הכול בסדר. אם האלגוריתם כפי שרץ עד עכשיו לא מצא אפילו התאמה אחת, נעבור להריץ אותו בצורה השנייה. כלומר ניצור מחפש את האינדקס, את הnGramDocument ואת הכול, ונריץ שוב את האלגוריתם, רק הפעם עם שני שינויים:

- לא מסירים התאמות פאזי שאינן באות אהו"י.
- מגדילים פי שלוש את הסף המינימלי של אורך ציטוט לפיו אנחנו מחליטים שהציטוט רועש.

הסיבה לשינוי השני היא שאם לא מורידים התאמות פאזי שאינן באות אהו"י אז מקבלים אמנם יותר תוצאות אך הרבה יותר רעש, לכן נבחר בגישה זו על מנת לצמצם את התוצאות.

בעיה ופתרון RemoveMatchBlankMatchesTags:

איחוד הציטוטים שביניהם יש חור באמצע יוצר לנו את הבעיה הבאה:

שתי התוצאות הבאות אינן נכונות:

בה יציאת מצרים הרי יש פרשה אחרת בתורה שהוזכר בה יציאת מצרים

tags: jbr:text-bavli-1-13-2,

והיה אם שמוע לויאמר שזו אינה קושיא כלל שהרי בתורה כתובה פרשת והיה אם שמוע

tags: jbr:text-bavli-1-14-2,

כיצד האלגוריתם מצא את הטעויות האלה? בכל אחד מהמקרים שלוש המילים הממורקרות אכן נמצאות בפרק הנ"ל, אך האלגוריתם מאחד אותם עם הרווחים ביניהם לציטוט אחד גדול ולא נכון.

על מנת לפתור בעיה זו, נבצע בסוף האלגוריתם את המניפולציה הבאה: כל תוצאה שהאלגוריתם מוצא שמכילה אותן שלוש עד חמש מילים (MINIMAL_NGRAM_LENGTH : MINIMAL_NGRAM_LENGTH+2) בהתחלה ובסוף נמחקת. כך ניפטר מתוצאות שגויות כאלה. המניפולציה האחראית לכך היא removeMatchBlankMatchTags.

בעיה ופתרון:

באופן כללי עד עכשיו הגדרנו NGramDocument בו אורך ה-ngrams הוא בין 3 ל-40. אורך מקסימלי זה של 40 עובד מצוין בטקסטים קצרים עד בינוניים, אך בטקסטים ארוכים מאוד, של אלפי מילים, נקבל כי יכולים להיות ציטוטים באורך הרבה יותר מ-40 מילים, ובמקרה כזה האלגוריתם ימצא במקום את הציטוט האחד הארוך המון ציטוטים באורך 40, שהוא כאמור האורך המקסימלי שהגדרנו לו

פתרון: נגדיר ליצור ngrams בין אורך 3 ל40 או ל-10% מאורך הטקסט, הארוך מביניהם. כך, עבור טקסטים ארוכים מאוד, נקבל שניצור גם ngrams באורכים של כמה מאות, וכך נפתור את הבעיה.

בעיה ופתרון RemoveMarginalLengthTagsIfManyMatches:

לפעמים האלגוריתם מוצא ציטוטים באורך סביר, לדוגמא באורך 5, עם עשרות רבות של התאמות, כלומר רעש. לדוגמא חמש המילים:

"אמר רבי שמעון בן לקיש" מניבות יותר מ-30 התאמות, שכמעט כולן רועשות כמובן.

פתרון: עד אורך QUOTE_PROBABLY_TOO_SHORT_VALUE אנחנו לא מקבלים כציטוט בכל מקרה. נוסיף את המקרה הבא- עבור אורכים עד פי 2 מהגבול התחתון, שהם גבוליים, אם בסוף התהליך יש לציטוט כזה יותר מ-NUM_MATCHES_THRESHOLD_FOR_MARGINAL_LENGTH_QUOTES - התאמות, נסיר את כל ההתאמות שלו. כנראה מדובר בביטוי נפוץ בגמרא שמופיע בהרבה מקומות. במקום הנכון אותו רצינו לתפוס הציטוט כבר אוחד בעזרת שאר האלגוריתם לציטוט ארוך יותר, ומה שנשאר הוא הרעש משאר המקומות שתפסנו מילים לו.

בעיות פתוחות:

למרות שהאלגוריתם מבצע עבודה בהחלט טובה על חלק מן המקורות היהודיים ועל חלק מן הטקסטים, עדיין קיימות בעיות פתוחות ומקרים איתם האלגוריתם מתקשה להתמודד.

בעיה פתוחה כללית:

במקורות יהודיים רבים בהם מצטטים את הגמרא משתמשים בשמות מלאים של רבנים או מוסדות, לדוגמא "בית שמאי", "רבי שמעון בן לקיש" ועוד. לעומת זאת בגמרא, שמתוכה הם מצטטים, מופיעים ראשי תיבות על מנת לקצר, לדוגמא עבור שתי דוגמאות אלו מופיעים בדרך כלל הקיצורים ב"ש ו-רשב"ל.

למרות שהאלגוריתם עדיין מוצא את הציטוט ללא ראשי התיבות אותן לא ציטט המקור ממנו לקוח הטקסט, היינו רוצים שהציטוט שהאלגוריתם מוצא יהיה מדויק יותר, ועבור כך דרוש מנגנון הממיר ראשי תיבות לשמות המלאים או להיפך.

בעיה פתוחה כללית:

טכניקת מיזוג ציטוטים סמוכים בה אנו משתמשים עדיין מביאה איתה רעש. למרות שאנו מגבילים את המרווח המקסימלי האפשרי בין מילים כך שעדיין נמזג אותם לציטוט אחד ארוך, עדיין שיטה זו מביאה איתה רעש. לדוגמא הטקסט "זירא מאי קרא דמצוה להתפלל עם הנץ החמה שנאמר ויראוך עם שמש" נחשב על ידי האלגוריתם על ידי ציטוט אחד ארוך מתוך 2-9-1. קשה להאשים את האלגוריתם, שכן שלושות המילים "זירא מאי קראה", "עם הנץ החמה", "ויראוך עם שמש" מופיעות שלושתן אכן ב-2-9-1, אך במקומות נפרדים. לכן היה יכול להיות נהדר אם הייתה דרך לוודא אם ציטוט מסוים מופיע לפני ציטוט אחר בתוך דף נתון, ולא רק האם שניהם מופיעים באותו הדף. בעזרת שיטה כזו היינו יכולים לפתור את הבעיה, אך כרגע התשתית אינה תומכת במנגנון זה. אולי, כמובן, ייתכנו גם פתרונות אחרים.

בעיה פתוחה כללית:

הרבה פעמים במקורות היהודיים חכמים מצטטים פסוקים מן התנ"ך, ואז האלגוריתם מוצא ציטוטים אלה כציטוטים במקומות רבים בתלמוד הבבלי כיוון שגם בתלמוד הבבלי התנ"ך מצוטט הרבה. אולי יהיה חכם להריץ את אלגוריתם מציאת הציטוטים מהתנ"ך על מנת לתפוס מקרים כאלה ולא לתייג אותם על מנת שלא נתאר מקרים כאלה כציטוטים מהגמרא, שכן במקור הם מהתנ"ך.

בעיה פתוחה כללית:

כרגע אנו מרשים ציטוטים באורך עד MAXIMAL_NGRAM_LENGTH מילים. זה מספיק לרוב הגדול של המקרים, אך במקרים קיצוניים (לדוגמא קטע מספר 13 במסילת ישרים) גודל זה יכול להיות לא קרוב אפילו למספיק. אנו לא יכולים באופן כללי להגדיל מספר זה ככל שנרצה משיקולי זמן ריצה.

ציטוטים ספציפיים אותם האלגוריתם אינו תופס:

להלן רשימת ציטוטים אותם האלגוריתם לא תופס כעת, כמו גם הערות מדוע ייתכן והאלגוריתם נכשל בזיהוי. מדובר לא בבאגים של האלגוריתם, אלא במקרים בהם על מנת לזהות את הציטוט יידרש שינוי דרסטי באלגוריתם, שינוי שלאו דווקא יהיה לטובה בראיה כללית. עדיין, מדובר במקרים קיימים להם יש לתת את הדעת. להלן הרשימה:

"מה הוא רחום אף אתה רחום" – ציטוט מקורי "מה הוא רחום וחנון אף אתה היה רחום וחנון".

"היום לעשותם ומחר לקבל שכר" – ציטוט מקורי "היום לעשותם **ל**מחר לקבל שכרם", הבדל שאינו באות אהו"י בציטוט קצר.

"היום לעשותם ומחר לקבל שכרם" – ציטוט מקורי "היום לעשותם **ל**מחר לקבל שכרם", הבדל שאינו באות אהו"י בציטוט קצר.

"אמר רבי יצחק: מלמד שנתקבצו כולן למקום אחד והיתה כל אחת אומרת, עלי יניח צדיק ראשו" – רווח של יותר מ-3 מילים בין ציטוטים קיימים.

"כל השם ארחותיו בעולם הזה, זוכה ורואה בישועתו של הקדוש ברוך הוא" – ציטוט אשר מופיע בשני מקומות בתלמוד. במקום אחד הוא מופיע כך כמו שהוא, ובמקום השני מוחלף הביטוי "הקדוש ברוך הוא" בראשי תיבות. לאחר שאנו מוצאים את המקום הראשון במלואו ואת השני באופן חלקי, אנו מעיפים את התגיות מהמציאה המצומצמת יותר כחלק מהמניפולציה שמסירה תגיות מציטוט מוכל בתוך ציטוט. בעייתי.

"אמר רשב"ל ויקרא יעקב לבניו" – ציטוט מקורי " דאמר רשב"ל {בראשית מט-א} ויקרא יעקב אל בניו". הבדל שאינו באות אהו"י.

"מאי לא הפסיד לא הפסיד ברכות" – ציטוט מקורי "מאי לא הפסיד שלא הפסיד הברכות".

דקרי ליה ליליא משום דאיכא אינשי דגנו בההיא שעתא – ציטוט מקורי "דקרי ליה יום דאיכא אינשי דקיימי בההיא שעתא".

"ההוא זוגא דרבנן דאשתכור בהילולא" – ציטוט מקורי "דההוא זוגא דרבנן דאשתכור בהילולא", ציטוט קצר עם אות שונה שאינה אות אהו"י.

" וכי מטא זמן וכו' ומאי ברכות אינם מעכבות [זו את זו] סדר ברכות" – ציטוט מקורי "וכי מטא זמן יוצר אור הוו אמרי ליה ומאי ברכות אין מעכבות זו את זו סדר ברכות". שלוש מילים ראשונות תואמות, אחר כך חור של יותר מהכמות המינימלית המתאפשרת של מילים, והבדל באות שאינה אהו"י (אינם לעומת אין) מונע ממנו להשיג תוצאות טובות יותר.