נושאים במדעי הרוח הדיגיטליים

מנחה: ד"ר יעל נצר

אוניברסיטת בן גוריון

סמסטר אביב תשע"ז

מגישים:

גלעד הושמנד

יואב אמיר

דניאל גל

שחר

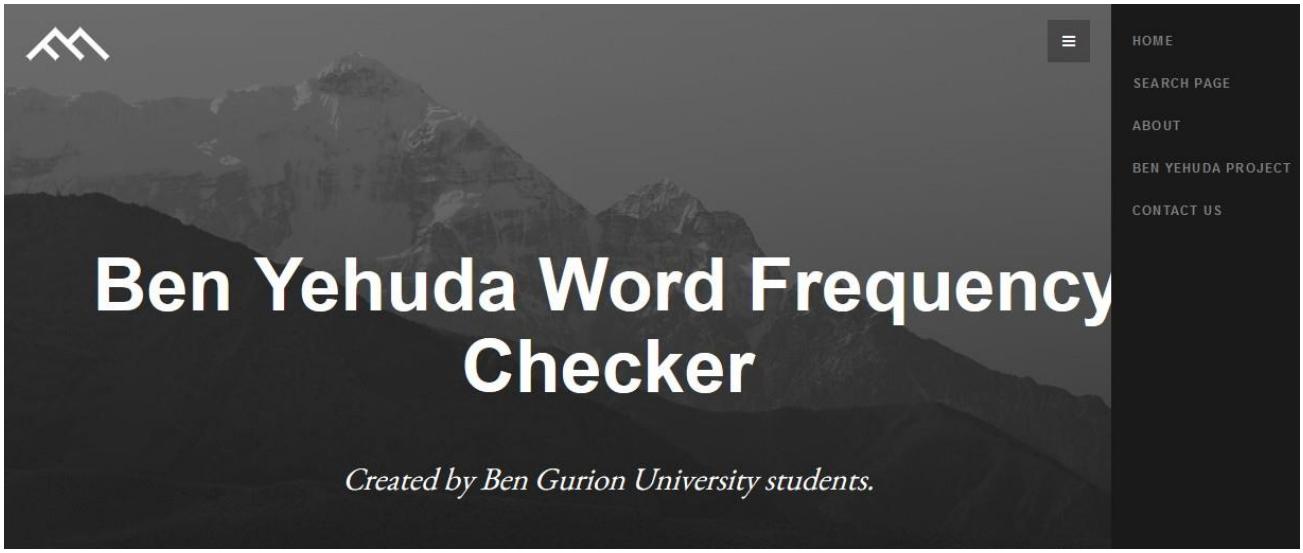
הקדמה

המשימה:

מיני פרוייקט בסגנון "האקתון", בנושא "מאגר בן יהודה". בחרנו, תחת הנחיה, ליצור ויזואליזציה של המילים שבמאגר. לשם כך יצרנו תצוגה במספר גרפים שמתאימים לחיפושים שונים ומבליטים היבטים מגוונים בתוצאות, המאפשרים לזהות דפוסים שונים במידע. ההשראה לפרוייקט הגיעה מ- Google ngrams.

בהנחיית ד"ר נצר הוחלט ליצור אתר שיציג נתונים הנמצאים במאגר. תחילה, הוחלט לייצג מידע עבור שתי מילות חיפוש בו זמנית ובהמשך החלטנו באופן עצמאי להרחיב את התמיכה עבור מספר לא מוגבל (תיאורטית) של מילים.

האתר:



דף נחיתה, מצד ימין תפריט נפתח ובו ניוד אינטרקטיבי לחלקים השונים בדף ולדפים חיצוניים.

Home: גולל לראש הדף.

Search Page: מבצע מעבר לכתובת דף החיפוש עליו נרחיב בהמשך.

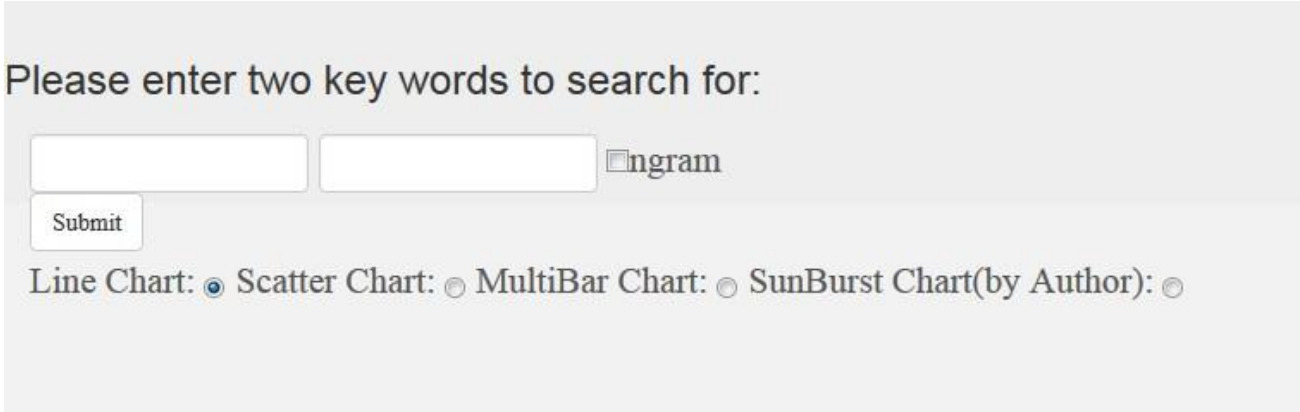
About: גולל להצגת הפרוייקט.

Ben Yehuda Project: קישור חיצוני לאתר של מאגר בן יהודה.

Contact Us: גולל לכתובות להתקשרות עם חברי הקבוצה.

דף החיפוש:

החיפוש נעשה מול מסד נתונים והתוצאות מאוגדות לפי שנת השימוש במילה ומוצגות לאורך ציר זמן כרונולוגי. ציר ה-y מייצג את מספר הפעמים שהמילה היתה בשימוש בשנה מסוימת.

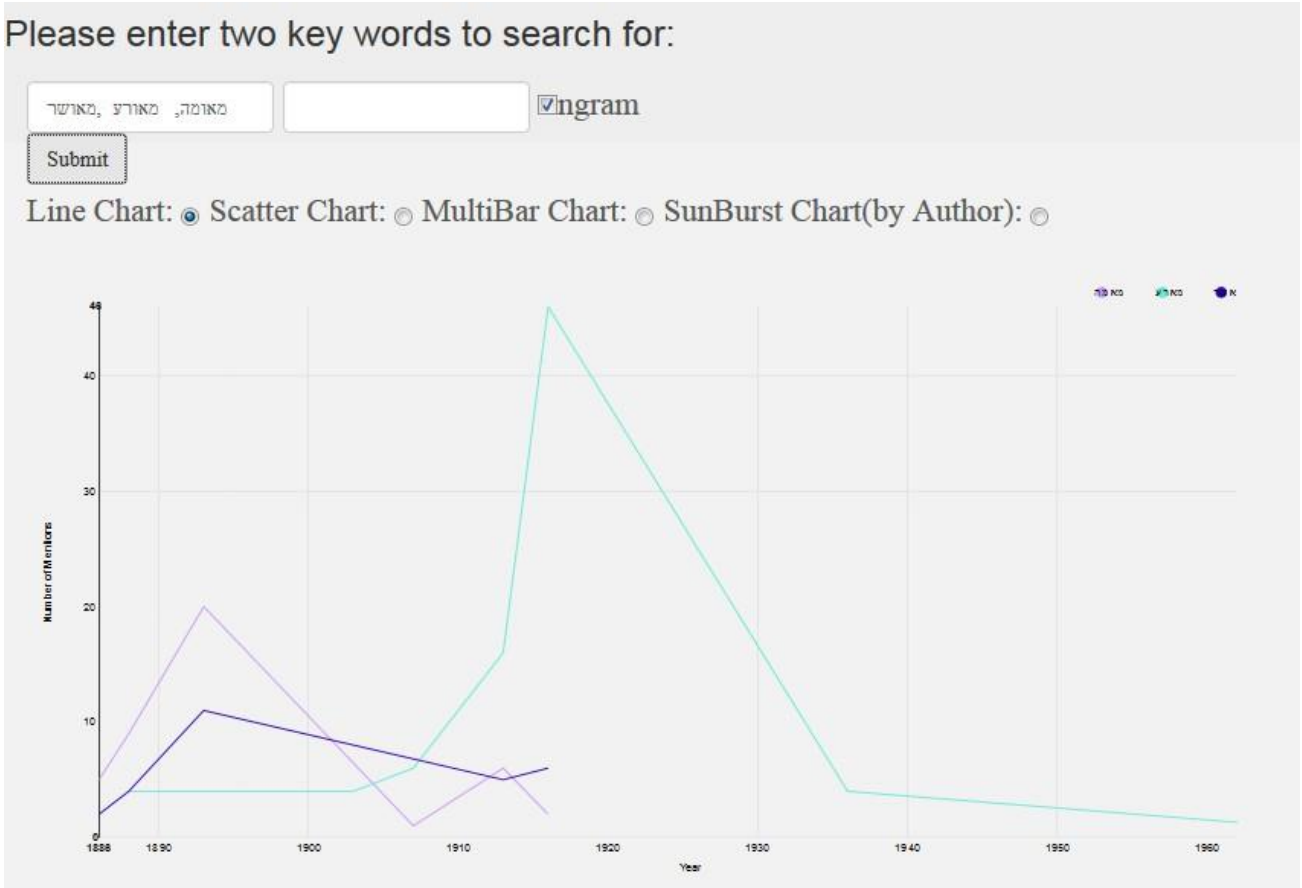


Line Chart, Multi-Bar Chart, Scatter Chart: אלו שלוש אפשרויות לחיפוש מילים במאגר. לכל גרף צורת הצגה שונה המבליטה היבט אחר של התוצאות.

Sun Burst Chart: באפשרות זו יש להשתמש בשדה החיפוש השמאלי. יש להזין שם סופר אחד.

ngram: סימון האפשרות הזו מאפשרת לנו לחפש מספר אינסופי של מילים ולהציגם יחד בגרף אחד. את מילות החיפוש יש להזין לשדה החיפוש השמאלי עם מפרידים ',' (פסיק). הערה: רווחים לא מפריעים לחיפוש, האלגוריתם יודע להתעלם מהם.

דוגמא לשימוש בngram:

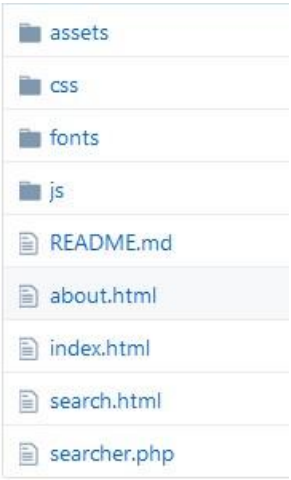


נתונים:

מאגר הנתונים הוא SQL Database. התקנו שרת web מקומי והרצנו את השאילתות מקומית.

פירוט נוסף בפירוט על קובץ searcher.php

מבנה וקבצי מפתח:



בכל הקבצים יש הערות רבות וניתן להבין לפיהם בלבד אך נוסיף כאן הערות חשובות על קבצי הjavascript וה php של דף החיפוש שהם, כאמור, ליבת הפרוייקט שלנו.

קבצי האתר הויזאלי בכניסה: index.html, about.html.

דף החיפוש - search.html.

קובץ javascript שמפעיל את דף החיפוש – js/scripts.js

המתודה runSearch:

1. מעבדת את הנתונים מהדף search.html.

2. בדיקת קלט.

3. מבצעת קריאת GET בעזרת דף הsearcher.php.

4. בודקת את הנתונים שחזרו וקוראת ל buildGraph אם הם תקינים.

המתודות הבאות ניתנות להבנה לבד בעזרת הרבה הערות שנכתבו לאורך הקוד.

```
/**
* Recieves an array from query result
* Returns an array in chart data format: [ {x:__ , y: ___ extra_field:___} ... {x:__ y: __ extra_field:___} ]
*/
function getMyDataReady(array) {

    /**
    * Method to build the desired graph
    * Receives js array (gets graphData from it)
    * gets max/min values of graph data into global data_info
    * switch on type of graph and build
    */
    function buildGraph(resultsArr) {
```

```
/**
 * Section Graph data population
 */
function createSunburstChart(data){
function createLineChart(data){
function createScatterChart(data){
function createMultiBarChart(data){

/**
* Receives data
* Returns [minX, maxX, minY, maxY , groupcolor[] ]
*/
function getMaxs(data){

/**
* helpers
*/
function IsJsonString(str) {

/**
* used for debugging
*/
function sinAndCos() {
```

דף ה-searcher.php backend

נתוני התחברות לשרת הSQL:

```
$db_host        = 'localhost';
$db_user        = 'root';
$db_pass        = '1234';
$db_database    = 'mini';
$db_port        = '3306';
$db_table       = 'mini.wordsby';
$max_results    = '31';
```

* max_results נועד להגביל את מספר התוצאות שחוזרות לבניית sun burst chart.

הגדרת השאילתות עבור כל מילת חיפוש לתוך מערך:

```
for($a = 0; $a<$number_of_search_terms; $a++)
{
    $sql[$a] = "";
    if($chart_type == "sunburstChart")
    {
        $sql[$a] = "SELECT * FROM ".$db_table." WHERE author = '".$searchWords[0]."' ORDER BY (count) DESC LIMIT ".$max_results.";" ;
    }
    else{
        $sql[$a]="SELECT * FROM ".$db_table." WHERE word = '".$searchWords[$a]."';" ; //  sql query
    }
}
```

אחר כך מתבצע:

1. חיבור לשרת
2. בלולאה: A. אתחול מערך לתוצאות.
   B. הרצת שאילתא.
   C. אכלוס תוצאות במערך הסופי.
3. סגירת החיבור.
4. שליחת תוצאות בעזרת echo json_encode() חזרה לדף הjavacsript.

סיכום:

פרוייקט זה אפשר לנו ליצור אתר העוסק בתחום ויזואליזצית מידע. מלבד התחלה איטית, העבודה התנהלה בקצב אחיד שהושפע מלמידה של סביבה ושפות תכנות חדשות. חילקנו בין חברי הצוות את העבודה כמטלות ליחיד ולשלישייה. הדבר נתן לנו אפשרות לעבוד לבד וביחד לסירוגין.

פרוייקט זה הציב בפנינו מספר אתגרים. תחילה היינו צריכים לעבד את המידע במספר שלבים, שכל אחד מהם דרש למידה עצמאית ונסיון וטעייה שלקחו זמן רב. היה עלינו ליצור מאגר של אומנים ובו שנת לידתם ועיר\ארץ הולדתם, כל אלו על מנת לקבוע את תאריך השימוש במילים. למדנו להשתמש במתייג של מני אדלר (קישורים) לצורך תיוג הטיות של מילה. לאורך השלבים ביצענו סינון של מידע פסול (כמו מילה שמתחילה באות סופית). היה עלינו ללמוד באופן עצמאי על הקמת ואחסון אתר ותקשורת צד לקוח ושרת,

דבר שלקח זמן רב. התמיכה באפשרות של הצגה ויזואלית של מספר מילים בלתי מוגבל, בשונה משתי מילים בלבד כפי שסוכם מראש, הצריכה מאמץ גדול מאיתנו והייתה לא פשוטה למימוש.

מבחינת הלמידה האישית שלנו, כל דבר שעשינו בפרוייקט היה חדש (לנו) ולכן למדנו הרבה דברים. מתוך ההתמודדות עם האתגר התאפשר לנו ללמוד שפות תכנות וסביבות חדשות, טכניקות עיבוד מידע והצגתו.

ההצלחה שלנו נמדדת בעינינו ביצירת פרוייקט שניתן ואף פשוט יחסית להמשיכו. הצלחנו לממש יכולת עיבוד של מספר בלתי מוגבל (תיאורטית) של מונחים. כל הכתיבה של הקוד משתמשת במשתנים לא מוגבלים (מערכים) ובנייה דינמית של כל תהליך החיפוש וההצגה של הנתונים ולכן ניתן בקלות לתחזק את המוצר ואף להרחיבו ולשפרו.

המשך הפרוייקט:

הפרוייקט נכתב בצורה מסודרת כמה שניתן על מנת לאפשר שימוש מעשי בו ואת המשך הבניה שלו.

דברים שניתן להוסיף\לשפר:

1. ניתן למחוק את תיבת החיפוש הימנית ולהשאיר רק תיבה אחת, במקרה שרוצים לחפש מספר מונחים משתמשים בסימני פיסוק ',' בינהם.
2. מספר מונחי החיפוש מוגבל כרגע ע"י גודל מערך הצבעים והצורות(אם משתמשים ב Scatter Chart). ניתן לשנות למספר צבעים וצורות אינסופיים במספר שעות עבודה.

```
function getMyDataReady(array) {
    var series = [];
    var shapes = ['circle', 'triangle-up', 'cross', 'triangle-down', 'diamond', 'thin-x', 'square'];
    var colors = ["#cc99ff" , "#63edd6"  , "#3300CC" , "#FF9933" , "#990033", "#FFFF99"];
    var ans = [];
```

3. אופציית sunburst chart – יש לסנן מילים מובילות מכיוון והן לרוב מילות קישור. ניתן להוסיף עוד רמה לפירמידה, המשתנים מוכנים לכך, מדובר במספר שינויים קטן בפונקצית הבניה של הגרף(ראה\י הערות בקובץ scripts.js מתודה buildSunBurstChart.
4. יש לצמצם את טווח המידע (מבחינת שנים) ובכך ליצור מידע יותר "צפוף" ותוצאות אייכותיות יותר.

הוראות הפעלה:

1. יש להוריד את התקייה מאתר גיטהאב (ראה\י קישורים).

בשרת הSQL:

2. יש להוריד את קובץ המאגר מאתר github (ראי\ה קישורים).

3. יש לייבא את מאגר הנתונים לתוך דטהבייס.

בשרת הWEB:

2. להעתיק את כל הקבצים במבנה הנוכחי שלהם לתקיית האתר.
3. יש לעדכן את פרטי ההתחברות בקובץ searcher.php (ראי\ה פירוט לגבי הקובץ).

<u>קישורים:</u>

1. אתר פרוייקט בן יהודה: http://benyehuda.org/
2. קבצי הפרוייקט: https://github.com/giladhosh/Ben-Yehuda-Ngrams
3. המתייג של מני אדלר:

https://www.cs.bgu.ac.il/~elhadad/nlp12/hebrew/TagHebrew.html