

# שפה אנלוגית בעולם דיגיטלי

ד"ר זף סגל ופרופ' אורן סופר שקדו יחדיו בשנים האחרונות על שורה של פרויקטים סביב מאגרי העיתונות הישנה מהעולם היהודי - לרבות ביתוח אלגוריתמים מילוליים שיכולים לעקוב ביעילות אחרי שיח תקופתי, וחקר של העיתונות העברית בעידן הטרומ-ציוני. "המשותף לכל המחקרים הללו הוא הניסיון ללמוד מתוך המדיום העיתונאי על תופעות חברתיות, תרבותיות ופוליטיות רחבות יותר, המושפעות מתנועה בזמן ובמרחב", מסביר ד"ר סגל לאסכולה. על זמן, מרחב ותקשורת בהיסטוריה של העם היהודי

## זף סגל

ד"ר זף סגל הוא חבר סגל המחלקה להיסטוריה, פילוסופיה ומדעי היהדות באוניברסיטה הפתוחה

מיוחד באשכנז :  
14 רייכס מאלק למנה .  
טעמתיך 8 סל' נפסה .  
ופן לרבע מנה .  
שכר מרשתת :  
מכל מנה קנה 5 קפ' .  
(מסורה רחבה 10 קפ' .)

# הצפירה

HAZEFIRAH.

ברוסא ופולין .  
סעלה ממוקדם נמנה 5 הי"ב .  
גלו מנה 2.75 כ"י .  
לרבע מנה 1.50 הי"ב .  
נמנה הולמנה 20 סלמק .  
מחמתיקה 5 דלולר .

## מכתב עתי משמיע חדשות

בקרב עם ישחון, מכל הדברים הנוגעים בעניני המדינות.

דברי חכמה ומרע, ידיענת העולם והמבע.

יצא לאור אחת בשבוע, מאת

היים זעליג סלאנימסקי.

№ 18.	ווארשויא יח אייר תרמ"א .	Варшава, 5 (17) Мая 1881.	שנה שמינית
-------	--------------------------	---------------------------	------------

## חדשות שונות

המכ"ע II פראג. Вѣст. מודיע, כי ביום (15) (27) מארץ העבר בשעה 12 בצהרים, התנפלו הנוצרים על היהודים בעיר יעליסאוועטגראד, ויכו בהם ככה רבה, אחר מהם המיתו ונפל הלל, וזעזורה עד עשרים נפצעו, ואך בשעה 2 אחר הצהרים בבוא אנשי הצבא לעזרה, שקטה שאון ההמון, והמכ"ע גאלאס מודיע כי גם בעיר אליוואפאל הקרובה, היתה מכת הנוצרים ביהודים שמה.

ממעם המיניסטער שעל הדרכים ומסלות הברזל, באו הפקודה להרשות להיהודים לעשות שולחן ערוך (ביופעט) עם מאכלים כשרים על מסלת הברזל מברעסט לקיעוו, ויש תקוה כי בקרוב יעשו כן בכל יתר המסלות, דבר נכבד מאד לכל אתב"י העוברים ושבים במלא רוחב ארצנו.

הבאראן ה' האראין גינצבורג נ"י, הקריש את אחד מבתי ארמונו בפארוין, לבית הועד ומקום מצבה לחכמי תחרשים חוצבי פסילים ואומנים ציירים הבאים טרוססיא לפארוין להשתלם בתכמה, להיות למזכרת נצה לבנו האהוב שהיה משכיל במלאכת מחשבת, ואשר נקטף בעורו באבו לפני שתי שנים על האי מאדעריא.

בימים האלה יטרו האנטיסעמיטים מכתב עתי הנכון לחפצים בשם, "דייטשעס טאנעכלאמט", ויהי לפלא כי המו"ל (רעראקטאר) שמו, "המן" I (Haman), והנומר הראשון אשר יצא בעיר ווירצבורג גאסר מאת הפאליציא. (Is. Woehs.)

פני כשלוש וחצי שנים הוזמנתי לשמש כעמית מחקר במעבדת המדיה והמידע של האוניברסיטה הפתוחה בתחום העיתונות ההיסטורית. החוקר הראשי של פרויקט זה היה פרופסור אורן סופר ז"ל. אף על פי שלא יכולתי לשער זאת, המפגש בינינו הוביל לשלוש שנים של עבודה מאומצת, מרתקת וחדשנית ששילבה בין טכנולוגיות מתקדמות, רעיונות תיאורטיים ולא מעט ניסוי וטעייה. כדי לתאר את המסע שעברנו, עליכם להכיר את נקודת המפגש של העולם הדיגיטלי עם חקר העיתונות ההיסטורית, ובפרט עם חקר העיתונות העברית. המפגש בין שני העולמות איננו ייחודי לחקר העיתונות העברית, והוא מקביל לתהליכים דומים שאירעו במדינות אחרות. ראשיתו של מפגש זה הוא הקמת מאגר מקוון של עיתונים וכתבי עת היסטוריים מהעולם היהודי, מאגר שהוקם ביוזמת אוניברסיטת תל אביב והספרייה הלאומית. האתר **עיתונות יהודית היסטורית** עלה לאוויר בשנת 2008 והנגיש לציבור הרחב מעל שלושה מיליון עמודים שפורסמו בין השנים 1783 ו-2014 ב-588 כתבי עת שונים. מאגר זה הציע אפשרות לבצע חיפוש מלא בכל המלל שפורסם בכל אחד מהעיתונים שהועלו לאתר. חיפוש מסוג זה ייעל באופן משמעותי את יכולתם של חוקרים לאתר תכנים רלוונטיים עבור מחקריהם, אך אף יותר מכך – הוא פתח צוהר לשימוש באלגוריתמים של "כריית מידע", המאפשרים לגלות באופן אוטומטי מגמות ודפוסים בבסיסי נתונים. כתוצאה מכך, באופן תיאורטי ניתן היה לשאול שאלות חדשות, שלפני כן כלל לא היו אפשריות. במקום לראות את כתבי העת כמקבצים של יחידות טקסט נבדלות, ניתן היה לראות כל אחד מכתבי העת, ובמידה מסוימת את כלל העיתונות היהודית, כמושא מחקר העומד בפני עצמו, ומורכב מכלל יחידות הטקסט ומיחסי הגומלין ביניהם. תפקיד החוקרים, על כן, היה להבין ולהסביר את יחסי הגומלין בין יחידות הטקסט, ואת התופעות ההיסטוריות והחברתיות הבאות לידי ביטוי ביחסים אלו.

### דיגיטציה וכריית מידע

שימוש בכלים חישוביים לחקר מסמכים היסטוריים דורש המרה של אובייקטים חומריים, כמו העיתונים המודפסים, לקבצים דיגיטליים הניתנים לקריאה. המרה כזו כוללת שלושה שלבים: ראשית, יש לצלם או לסרוק את החומר המודפס ולהפכו לקובץ תמונה פשוט; שנית, יש לחלק את העמוד הסרוק למקטעים המרכיבים אותו – כשמדובר בעמוד עיתון, הכוונה היא לחלוקת הדף

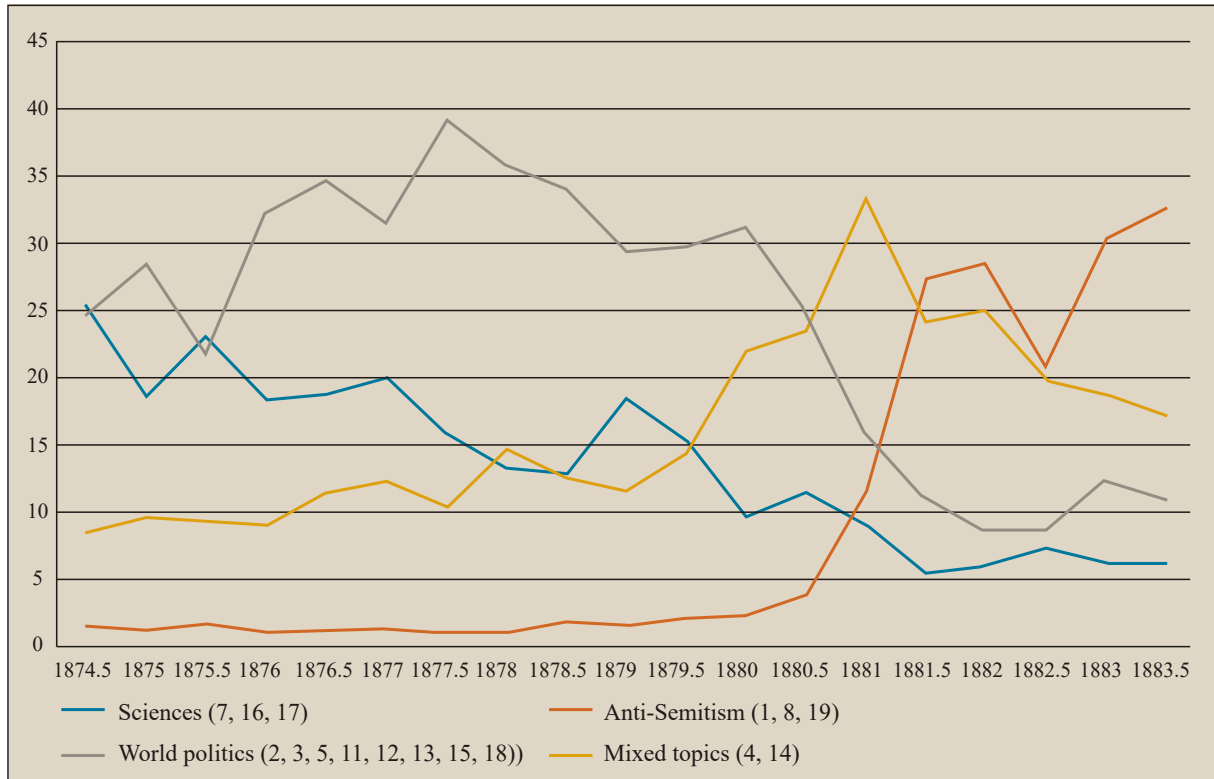
לכתבות השונות הכלולות בו; ושלישית, יש להמיר את קובץ התמונה לקובץ טקסט קריא באמצעות טכנולוגיית זיהוי כתב (Optical Character Recognition).

אף על פי שהמאגר של הספרייה הלאומית פתר את שתי הבעיות הראשונות, איכות זיהוי הכתב נותרה נמוכה מאוד (כ-60% דיוק בזיהוי התווים). כתוצאה מכך, תחילת העבודה של מעבדת המדיה והמידע באוניברסיטה הפתוחה התמקדה בפיתוח תהליך חישובי ששיפר את איכות הזיהוי לכדי 98% דיוק, ובמקביל שמר על החלוקה הפנימית שזוהתה על ידי הספרייה הלאומית. באמצעות תוכנת טרנסקריבוס (Transkribus) – תוכנה ללמידת מכונה של כתבי יד, פיתח צוות המעבדה חבילת תוכנה שמאפשרת לכל מתכנת ומתכנתת להריץ את התהליך במספר פקודות פשוטות.

### על נושאים משתנים ו"סופות בנגב"

לאחר שהיו בידינו קבצים דיגיטליים "קריאים", עברנו לשלב הניסויי של הפרויקט. התמקדנו בעיתון אחד, ה**צפירה**, שיצא לאור בוורשה ובברלין, ובעשור בודד, 1874-1883. זאת במטרה לבחון את יעילותם של הכלים החישוביים. תוכנות לחקר טקסט, כגון Antconc, או לחקר מרחב, כגון QGIS, שימשו במקביל לצורך אפיון העשור. מחקרנו הראה, לדוגמה, שבעקבות שנת 1881 שבה פרצו פוגרומים בדרום רוסיה, המכונים "סופות בנגב", חל שינוי דרמטי בשיח. בעוד שלפני הפוגרומים הוזכרו בעיקר המעצמות האירופיות השונות: פרוסיה, רוסיה, צרפת ואנגליה, הרי שבעקבות הפרעות החל העיתון לעסוק בארץ ישראל ובאמריקה, יעדיהם של מהגרים אפשריים ממזרח אירופה.

מחקר זה העמיק עוד יותר כאשר הפעלנו על העשור שבידינו אלגוריתם LDA לזיהוי "נושאים". אלגוריתם זה מתייחס לכל מסמך כאוסף מילים, ולקורפוס כולו כמספר רב של אוספי מילים. באמצעות ניתוח סטטיסטי של הופעות משותפות באותו אוסף, מאפיין האלגוריתם מקבצי מילים הנוטים להופיע יחד, ואלו מכונים "נושאים". מילה אחת יכולה להופיע במספר נושאים בהסתברויות משתנות, כשם שהמילה "ספר" תהיה בעלת משמעות אחת כאשר היא מופיעה יחד עם המילים "שמפו" ו"שיער", בעלת משמעות אחרת כאשר היא מופיעה יחד עם המילים "עיון" ו"פרוזה", ובעלת משמעות שונה לחלוטין כאשר היא מופיעה יחד עם המילים "בית" ו"מורה". יתרונו של ה-LDA הוא שהחלוקה לנושאים נעשית באופן לא מפוּקח, בהתאם לשכיחות המילים הקיימת, ולא בעקבות התערבות



איור 1 חלקם היחסי של תוכני העיתון מסך כל הטקסט בשנים 1874-1883. הנושאים מקובצים לארבע קבוצות: תכנים מדעיים (כחול), תכנים בנושאי פוליטיקה עולמית (אפור), תכנים מעורבים (צהוב) ותכנים בנושאי יהודים וארץ ישראל (כתום).

Zef Segal & Oren Soffer, "One Journal, One Decade, 3,797,592 Words: Computational Analysis of HaTzifira's Discourse (1874-1883)," *Journal of Jewish Studies*, 73 (2021)

בדרום רוסיה שינו לחלוטין את נושאי השיח של העיתון היהודי מוורשה.

### על זמן ותאוצה – משבועון ליומון

העיסוק שלנו באקטואליה ובתוכני העיתון הביא אותנו לתהייה לגבי סוגיות של זמן בעיתון הצפירה ובעתונות בכלל. לכתבי עת יש תפקיד חשוב ביצירת משמעות לזמן בחברה מודרנית, בין אם מדובר ביומון המונח בבקרים מחוץ לדלת הבית, בשבועון הנקרא בסופי השבוע, או במגזינים החודשיים הנקראים כספר לאורך זמן ממושך. בפרט, יש למדיום העיתונאי תפקיד בהגדרת ה"הווה", האירועים המתקיימים "עכשיו". אי לכך, ניתן לצפות להבדלים בתפיסת ה"הווה" בין חדשות המתעדכנות בלי הרף (כמו באתרי חדשות וירטואליים), לבין אלה המתעדכנות מדי יום (כמו ביומונים) ולבין אלה המתעדכנות מדי שבוע (כמו בשבועונים).

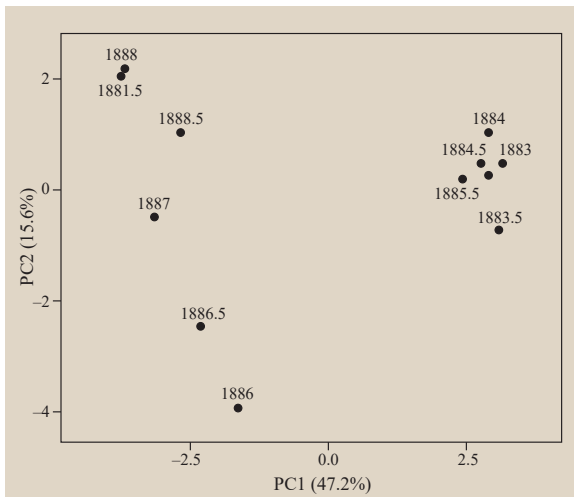
אנושית באלגוריתם. יחד עם זאת, "יצירתיות" זו מקשה לא פעם על אפיון הנושאים עצמם, המוגדרים על ידי המילים השכיחות שבהם ועל ידי המסמכים שבהם הם נפוצים. על כן, אפיון הנושאים מתבסס על ההיכרות של החוקרים עם חומרי המקור ועם הנסיבות ההיסטוריות. הפעלת האלגוריתם על עשור נבחר אפשרה לנו לאפיין את השינויים בשיח של אותו עשור. באמצע שנות השבעים של המאה ה-19, עיתון הצפירה התמקד בהנחלת ידיעה מדעית לקוראיו, כפי שבא לידי ביטוי באיור 1 (הקו הכחול). אך שיח זה, שתאם למטרותו המקורית של מייסד ועורך העיתון חיים זליג סלונימסקי, דעך בהדרגה. התחרות מצידם של עיתונים אחרים והדרישה מצד הקהל הביאו לעליית חשיבותן של החדשות העולמיות (הקו האפור). אולם, אירועי "סופות בנגב" הביאו להחלפת נושאי השיח האקטואליים ה"כלליים" בשיח אקטואלי יהודי (הקו הכתום), ובפרט לעיסוק ער בסוגיות האנטישמיות. המאורעות החריגים

כדי לבחון את הקשר בין השיח לזמן, שבנו לעיתון **הצפירה**, אך בנקודת זמן שונה. בשנת 1886, החליטו שני עורכי העיתון, סלונימסקי ונחום סוקולוב, לשנות את אופיו של העיתון משבועון ליומון. שינוי זה עלה בקנה אחד עם ייסודם של מספר ימונים עבריים נוספים ברחבי אירופה. למעט השינוי בתדירות ההופעה, לא השתנה דבר. העורכים נותרו אותם עורכים, והנסיבות ההיסטוריות נותרו כשהיו. מדובר במקרה מבחן המאפשר לבדוד את השפעת תדירות ההופעה מהשפעות אחרות, כגון צוות העורכים או הנסיבות ההיסטוריות.

השימוש באלגוריתם LDA על תקופה של שלוש שנים לפני השינוי משבועון ליומון ושלוש שנים לאחר מכן (1883-1888) גילה תוצאות מרתקות. ראשית, כפי שניתן לראות באיור 2, הצגת הגיליונות השונים על פי התפלגות הנושאים בתוכם משקפת חלוקה ברורה בין כלל הגיליונות שפורסמו לפני עידן היומון (צד שמאל של האיור) לבין כלל הגיליונות שפורסמו בזמן עידן היומון (צד ימין של האיור). שנית, תוכני השיח השתנו לתמקד השיח בנושאים יהודיים שונים, ובפרט בסוגיית האנטישמיות. אולם, עם המעבר ליומון, התמקד השיח בנושאים כלכליים, הן בשל עליית חלקן היחסי של פרסומות בדפי העיתון, והן בשל הצורך למלא את דפי העיתון במידע זמין, אקטואלי ומתחדש על בסיס יומי. שלישית, ובניגוד למצופה, השיח של השבועון היה נתון להרבה יותר שינויים באופן יחסי מאשר השיח של היומון. בשבועון, עלו לכתורות נושאי שיח אקטואליים, ולאחר תקופת זמן הם הוחלפו בנושאי שיח אקטואליים חדשים, באופן שהזכיר תנועה של מטוטלת. לעומת זאת, ביומון, נושאי השיח המשתנים תפסו מקום קטן בהרבה, ומרבית התוכן היה שגרת. הצורך המערכתי לייצר תחושה של רציפות והמשכיות במציאות של עדכון יומי הביא לקביעת מתכונת אחידה של החדשות וליצירה של שגרה עיתונאית.

עיקריים: הראשון – זיהוי שימושים חוזרים של קטעי טקסט בעיתונים שונים וניתוחם, והשני – זיהוי חתימות סגנון המשקפות כותבים משותפים או סוגות משותפות. השילוב בין שני הכלים מאפשר לחשוף יוצרים ושחקנים בולטים בשדה העיתונאי, ובמקביל לחשוף גם אסכולות וזרמים אידיאולוגיים שחצו את הגבולות הגיאוגרפיים והמוסדיים ופעלו ברחבי התפוצה היהודית.

המשותף לכל המחקרים הללו הוא הניסיון ללמוד מתוך המדיום העיתונאי על תופעות חברתיות, תרבותיות ופוליטיות רחבות יותר, המושפעות מתנועה בזמן ובמרחב. אף על פי שכל אחד מהכלים החישוביים שבהם אנו משתמשים מגיעים מתחומי דעת שונים – בלשנות, גיאוגרפיה, ספרות, מדעי המחשב ומתמטיקה – הם מאתרים עבורנו מגמות ותהליכים הסמויים מהעין בתוך הקורפוס המלא. למגמות אלו, בפני עצמן, אין משמעות בלעדי ההקשר ההיסטורי. אנו מספקים את ההקשר ההיסטורי באמצעות חזרה אל הטקסטים עצמם ואל הדמויות והמוסדות שהפיקו את אותם טקסטים. המחשב והתוכנות אינם מחליפים את עבודתנו כהיסטוריונים וחוקרי תקשורת, אלא מעניקים לנו עוד שכבת מידע שממנה אנו מתפתחים. ■



**איור 2** ניתוח גורמים ראשיים (PCA) של התפלגות הנושאים, 1888-1883. כל ציר משקף ממד אחד של הווקטורים שזוהו על ידי האלגוריתם. האחוז בכל ציר משקף את החלק היחסי של השונות המתגלה על ידי ציר זה. אין משמעות מוחלטת להבחנה בין שמאל לימין, בין למעלה ולמטה ולערכים המופיעים על גבי הצירים. ערכים אלה מייצגים את השוני בין הווקטורים. ההקצנה (הוויזואליזציה) נעשתה בתוכנת ClustVIs.

Zef Segal & Oren Soffer, "From Weekly to Daily: Computational Analysis of Periodical Time Cycles," *Journalism Studies* 21.14 (2020), 1952-1972. <https://doi.org/10.1080/1461670X.2020.1807394>

## על רשתות וקשרים – שימוש חוזר בטקסטים

מחקרנו האחרון, שנמצא בעיצומו, מבקש לרתום את הכלים החישוביים לקורפוס רחב בהרבה, לכלל העיתונות העברית במאה ה-19, ולחקור באמצעותם את תהליכי ההיווצרות של רשת עיתונאית עברית כלל עולמית בתקופה שקדמה להקמת התנועה הציונית. כדי לעמוד על זיקות בתוך הרשת אנו מפעילים שני כלים חישוביים