

בית ספר החורף למדעי הרוח הדיגיטליים #3

האוניברסיטה העברית בשיתוף אוניברסיטת חיפה



Digital Humanities @ The Hebrew University



בית ספר ג'ק, ג'וזף ומורטון מנדל
ללימודים מתקדמים במדעי הרוח



ברוכים הבאים

- דברי ברכה פרופסור ניסים אוטמזגין, דיקן הפקולטה למדעי הרוח, פרופסור דניאל שוורץ, ראש אקדמי של בית ספר מנדל ללימודים מתקדמים במדעי הרוח
- מטרות, ציפיות, הצגת תכנית בית הספר
- הכרות עם המרצים והסטודנטים

יום שלישי 15.2.22

פתיחת בית ספר החורף		09:00-10:30
הפסקה		10:30-10:45
מושגי יסוד במדעי הרוח הדיגיטליים – רננה קידר	סדנה 1	10:45-12:30
צהרים		12:30-13:30
גישות מדעיות וחשוביות במדעי הרוח – ברק סובר	סדנה 2.1	13:30-15:00
הפסקה		15:00-15:15
גישות מדעיות וחשוביות במדעי הרוח – ברק סובר	סדנה 2.2	15:15-16:45
הפסקה		16:45-17:00
הצגת פרויקטים במדעי הרוח הדיגיטליים		17:00-18:30

יום רביעי 16.2.22

09:15-10:45	סדנה 3.1	E-scriptorium – משה לביא וסימון גבאי
10:45-11:00	הפסקה	
11:00-12:30	סדנה 3.2	E-scriptorium – משה לביא וסימון גבאי
12:30-13:30	סמינר צהרים	איתי מרינברג-מיליקובסקי: הערפל, המריצה והתחבולן: גבולות ופריצתם במעבדה
		למחקר חישובי במדעי הרוח
		מגיבים: משה לביא, אופיר מינץ-מנור ורננה קידר
13:30-15:00	סדנה 4	טרנסקריבוס - סיני רוסינק
15:00-15:15	הפסקה	
15:15-16:45	סדנה 5.1	CATMA – איתי מרינברג-מיליקובסקי
16:45-17:00	הפסקה	
17:00-18:30	סדנה 5.2	vis-À-vis – אופיר מינץ מנור

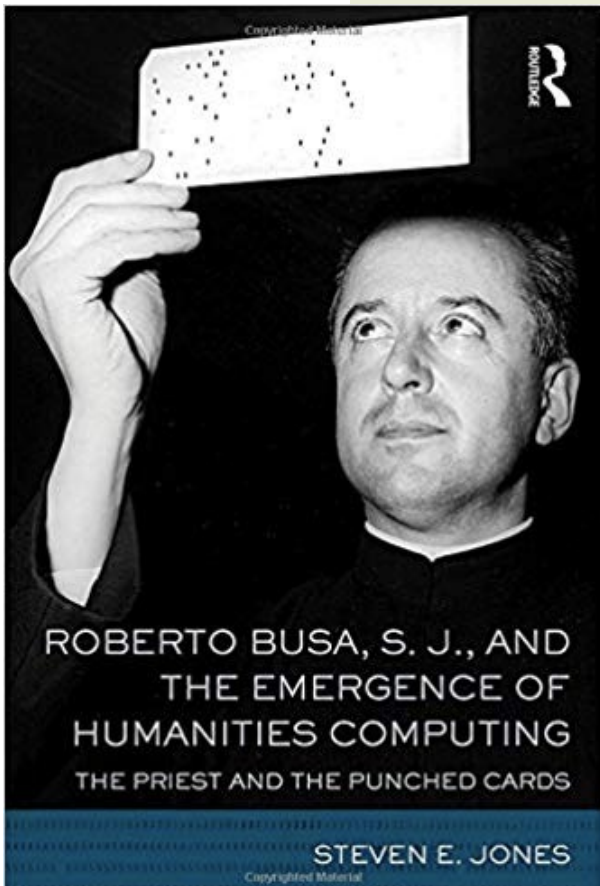
יום חמישי 17.2.22

פתיחת יום מרחבי – משה לביא מטקסט לזיהוי מקום – מורן זגה, סיני רוסינק יצירת מפה – מורן זגה	סדנה 6	09:15-11:30
	הפסקה	11:30-12:00
ארגון דאטה ו Open Refine – יעל נצר	סדנה 7.1	12:00-13:30
	צהרים	13:30-14:30
ארגון דאטה ו Open Refine – יעל נצר	סדנה 7.2	14:30-16:00
	הפסקה	16:00-16:15
סדנת תכנון פרויקט אישי	סדנה 8	16:15-17:15
	הפסקה	17:15-17:30
סיכום בית ספר החורף: משובים, הערות, הצעות לשנה הבאה		17:30-18:30

רננה קידר, מבוא למדעי הרוח הדיגיטליים

- מה זה מדעי הרוח הדיגיטליים?
- למה לעשות מדעי הרוח הדיגיטליים?
- איך לעשות מדעי הרוח הדיגיטליים?

מה זה מדעי הרוח הדיגיטליים?



CORPUS THOMISTICUM
INDEX THOMISTICUS

by Roberto Busa SJ and associates
web edition by Eduardo Bernot and Enrique Alarcón
English version

Search:

[concordances](#) [terms](#) [works](#) [options](#) [new search](#)

STATISTICS SHOWN ARE FOR ALL 220 WORKS
FREQUENCY BASE IS FOR A TOTAL OF 10631974 WORDS IN 87151 PLACES

Lemma 39225

Ierusalem: Ierusalem

Hierosolyma: Hierosolyma Hierosolym-orum

Hierusalem: Hierusalem

Ierosolyma: Ierosolyma Ierosolym-orum

Ierosolima: Ierosolima Ierosolim-orum

Jerosolyma: Jerosolyma Jerosolym-ae

Jerusalem: Jerusalem

Type of lemma: (B) proper name

Type of meaning: (NP) proper noun

[all](#) [none](#) [remove](#)

cases	c. freq.	places	pl. freq.	form	type	infl.	num.	gen.	case	deg.	voice	tense	mood	pers.	comp.
568	0,05‰	295	0,34%	☐ Ierusalem	n.	indecl.									
560	0,05‰	246	0,28%	☒ Jerusalem	n.	indecl.									
32	0,00‰	29	0,03%	☐ Hierusalem	n.	indecl.									
9	0,00‰	8	0,01%	☐ Ierosolyma	n.	1 st decl.	sing.	f.	nom.	pos.					
3	0,00‰	3	0,00%	☐ Jerosolyma	n.	1 st decl.	sing.	f.	nom.	pos.					
14	0,00‰	11	0,01%	☐ Hierosolyma	n.	1 st decl.	sing.	f.	nom.	pos.					

Terras and Nyhan, Father Busa's Female Punch Card Operatives





ניהול שאילתות | שמירת שאילתה | שמור | הדפסה

[>>הקודם | הבא<<]

ירושלים

חיפוש

נמצאו 20195 מסמכים [1 עד 20]

חיפוש בתוצאות | מסמכים בעמוד: 20

עבור לעמוד: 1 2 3 ... 1010

1. משנה מסכת אבות פרק א

... את הרב שלא על מנת לקבל פרס ויהי מורא שמים עליכם:משנה ד[*] יוסי בן יעזר איש צרדה ויוסי בן יוחנן איש **ירושלים** קבלו מהם יוסי בן יעזר אומר יהי ביתך בית ועד לחכמים והוי מתאבק בעפר רגליהם והוי שותה בצמא את דבריהם:משנה ה[*] יוסי בן יוחנן איש **ירושלים** ...

2. משנה מסכת יומא פרק ז

... קורא על פה ומברך עליה שמנה ברכות על התורה ועל העבודה ועל ההדאה ועל מחילת העון ועל המקדש בפני עצמו ועל ישראל בפני עצמן (ועל **ירושלים** בפני עצמה) ועל הכהנים בפני עצמן ועל שאר התפלה:משנה ב[*] הרואה כהן גדול כשהוא קורא אינו רואה פר ושעיר הנשרפים והרואה פר ושעיר ...

3. משנה מסכת מגילה פרק א

... אוכלים קדשים קלים ומעשר שני בכל הרואה ובירושלים לפנים מן החומה וכאן וכאן קדשי קדשים נאכלים לפנים מן הקלעים קדושת שילה יש אחריה היתר וקדושת **ירושלים** אין אחריה היתר:...

4. משנה מסכת מגילה פרק ד

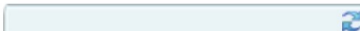
... נקרא ולא מיתרגם ברכת כהנים מעשה דוד ואמנון לא נקראין ולא מיתרגמין אין מפטירין במרכבה ורבי יהודה מתיר רבי אליעזר אומר אין מפטירין בהדע את **ירושלים**:נשלמה מסכת מגילה:...

5. משנה מסכת מכשירין פרק א

... בכי יותן והאוכל שומשמן באצבעו משקין שעל ידו רש"א אינו בכי יותן וחכ"א בכי יותן [*] הטומן פירותיו במים מפני הגנבים אינו בכי יותן מעשה באנשי **ירושלים** שטמנו דבילתן במים מפני הסיקריין וטהרו להן חכמים הנותן פירותיו בשבולת הנהר להביאן עמו אינו בכי יותן:...

6. משנה מסכת מנחות פרק י

... בא לו לעשרון ונתן שמנו ולבונתו יצק ובלל הניף והגיש וקמץ והקטיר והשאר נאכל לכהנים:משנה ה[*] משקרב העומר יוצאין ומוצאין שוק **ירושלים** שהוא מלא קמח וקלי שלא ברצון חכמים דברי רבי מאיר רבי יהודה אומר ברצון חכמים היו עושים משקרב העומר הותר החדש מיד והרחוקים מותרים מחצות היום ולהלן ...



- ☐ תנ"ך
- ☐ ספרות חז"ל
- ☐ משנה
- ☐ תוספתא
- ☐ מסכתות קטנות
- ☐ תלמוד בבלי
- ☐ תלמוד ירושלמי (יולגא)
- ☐ תלמוד ירושלמי (נוציה)
- ☐ מדרשי הלכה
- ☐ מדרשי אגדה
- ☐ זוהר ומפרשיו
- ☐ גאונים
- ☐ מפרשי תנ"ך
- ☐ מפרשי המשנה ומדרשי הלכה
- ☐ מפרשים ופוסקים על הבבלי והירושלמי
- ☐ ספרי הלכה ומנהג
- ☐ ספרי מצוות ומפרשיהם
- ☐ ספרי מחשבה ומוסר
- ☐ רמב"ם ומפרשיו
- ☐ טור שולחן ערוך מפרשים וחיבורים
- ☐ ספרי מערכות ועניינים
- ☐ דרשות ודרושים
- ☐ ספרי חסידות
- ☐ ספרי כללים וסדר הדורות
- ☐ ספרי שאלות ותשובות (שו"ת)
- ☐ אנציקלופדיה תלמודית

חיבורים **מילים** **חיפוש**

מילה

הקלדת מילה לחיפוש

סוג: הכול

חיפוש מילה בטקסטים שבמאגרים על פי הערך המילוני ועל פי צורת המילה. לאחר קבלת תוצאות הביניים לחצו על איקון הצגת המובאות בשורה המתאימה. באזור הסינון מימין אפשר לצמצם את קורפוס הטקסטים לפני החיפוש.

ברוכים הבאים

אתר "מאגרים" נפתח לקהל הרחב בשנים 2014-2016 בתמיכת תוכנית מורשת במשרד ראש הממשלה.

סינון

הוגדר סינון: חיפוש ב-134 חיבורים

תקופה

משנת -300 עד שנת 2000

רק הסוגה ☐ ללא הסוגה

ספרות יפה

רק המתברים ☐ ללא המתברים

הקלדת המחבר

רק החיבורים ☐ ללא החיבורים

הקלדת שם חיבור

האקדמיה ללשון מכרטסת מלים כבר משנות ה-60

חדשות | חינוך וחברה

1 | שמרו | קריאת זן

המילון ההיסטורי אינו רק מפעל רוחני ומחקרי, אלא גם מיזם טכנולוגי מרתק שניצל את המחשב מימיו הראשונים לאחסון ולעיבוד מידע רב. ב-1957 סייר מייסד המילון ועורכו הראשון, פרופסור זאב בן חיים, באירופה בחיפוש אחר השיטה הטובה ביותר לניהול וכתובת המילון.

בן חיים ביקר אז במקומות שנכתבים בהם המילונים ההיסטוריים החשובים ביבשת - האנגלי, הגרמני והצרפתי - וכתב על כך דו"ח לחברי האקדמיה. הוא בחן את השיטות השונות ומצא בהן בעיות. השיטה האנגלית הסתמכה על מאות רבות של מתנדבים, שרובם אינם בלשנים מקצועיים. "רוח זו, מסירות זו שקולות כנגד ליקויים ניכרים

חדש ב"מאגרים"

- קטעי הגניזה של התלמוד הירושלמי משולבים מעתה כגרסה בתוך חיבורי
- שישה ספרים נוספים ממשנה תורה לרמב"ם: [ספר המדע](#), [ספר אהבה](#), [ס](#)
- [הרה](#), [ספר קניין](#)
- [שערי תשובה](#) לרבנו יונה גירונדי
- [ספר פתרון תורה](#) לרב האיי גאון
- [ספר המסע](#) לארץ ישראל של משולם מוולטרה, בן המאה החמש-עשרה, שהגיע לידינו



ניר חסון

התראות במייל

פורסם ב-06.04.12

אופיר מינץ-מנור, פיוט אנלוגי בעולם דיגיטלי

פרספקטיבות
כמותיות או
חישוביות, מבוססות
סטטיסטיקה ולמידת
מכונה, למחקר
במדעי הרוח

פער גדול בין מאגר
ממוחשב שניתן
לאחזר ממנו מידע
במהירות לבין
מערכת ממוחשבת
שמסוגלת לבצע
משימות שדורשות
ניתוח ועיבוד של
תוצרים אנושיים.

גישות חישוביות
לחקר מדעי הרוח
המאפשרות ביצוע
סוגי מחקר חדשים
שמבוססים בחלקם
על למידת מכונה

מערכות ממוחשבות
שאינן משנות באופן
מהותי את דרכי
המחקר אלא מייעלות
אותו ובמקרים
מסוימים הופכות
אותו למקיף ומדויק
יותר

אופיר מינץ-מנור, פיוט אנלוגי בעולם דיגיטלי

- השימוש בגישה כמותנית לחקר מדעי הרוח הוא חדש?
- האם זאת דיסציפלינה עצמאית במדעי הרוח? האם זאת הרחבה של דיסציפלינות קיימות?
- האם אפשר לדבר על **מדעי הרוח** הדיגיטליים?
- האם אפשר לדבר על מדעי הרוח **החשובים**?

סקאלה של חישוביות



הנגשת קורפוסים,
סריקה, דיגיטציה
OCR



קורפוסים מתוייגים



מאגרי מידע, חיפוש,
אחזור טקסט,
ניתוחים



כלי עזר לניתוח



כלים לניתוח שפה
טבעית NLP

סוגי פרוייקטים במדעי הרוח הדיגיטליים

ייצור: יצירת טקסטים דיגיטליים ומקודדים



Producing

Digital texts, text encoding.

גילוי: הנגשת מקורות, דיגיטציה של כתבי יד



Discovering

Digitizing and making materials available

ניתוח: כריית מידע, עיבוד וניתוח דאטה



Analyzing

Text mining, network analysis

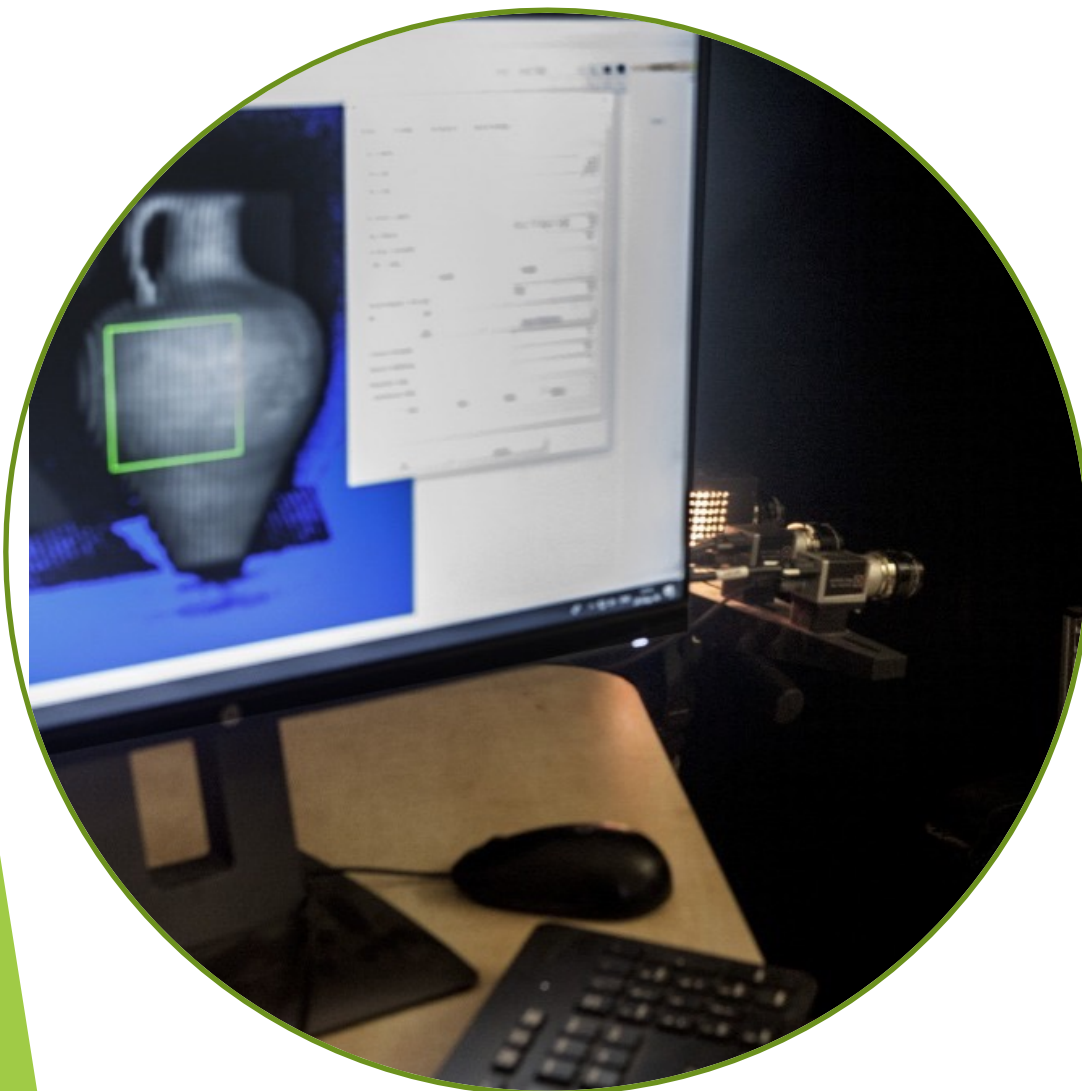
שימור: ארכוב ויצירת מאגרים וארכיונים דיגיטליים



Preserving

Archiving data and repositories

סוגי אובייקטים במדעי הרוח הדיגיטליים



קורפוס עמך"

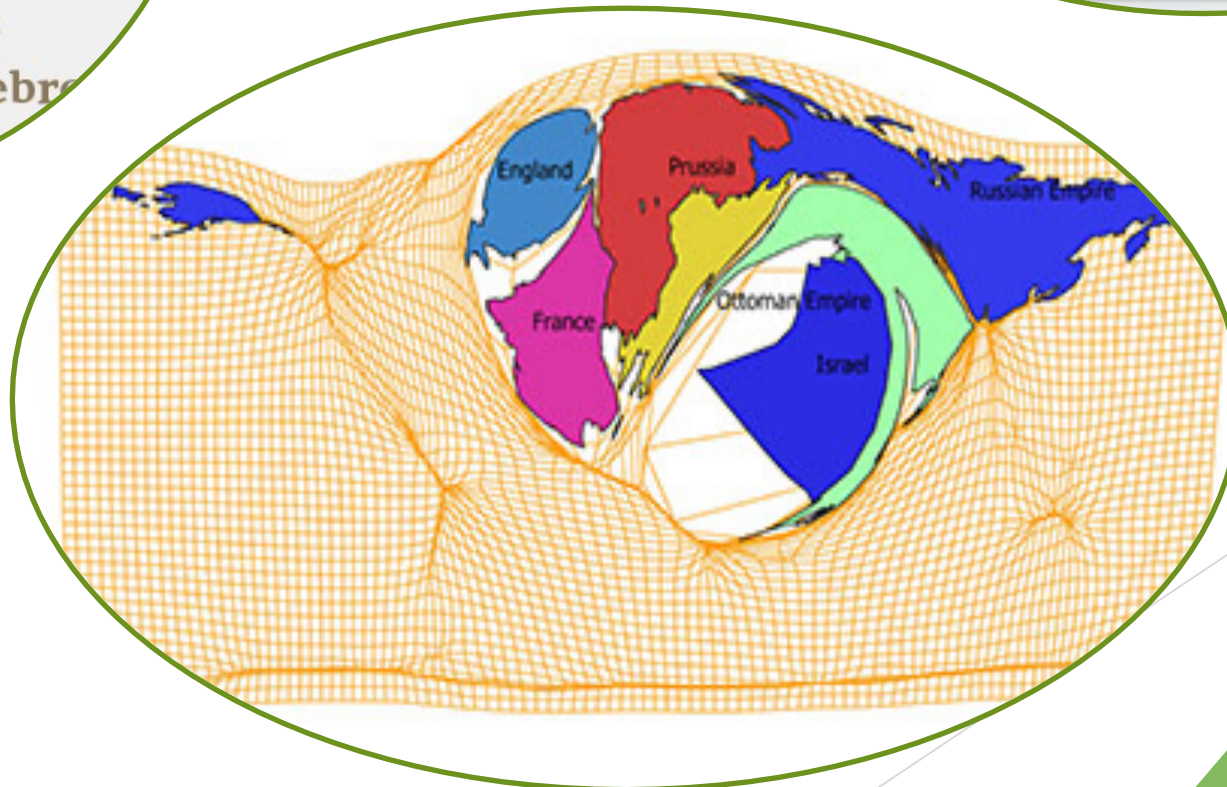
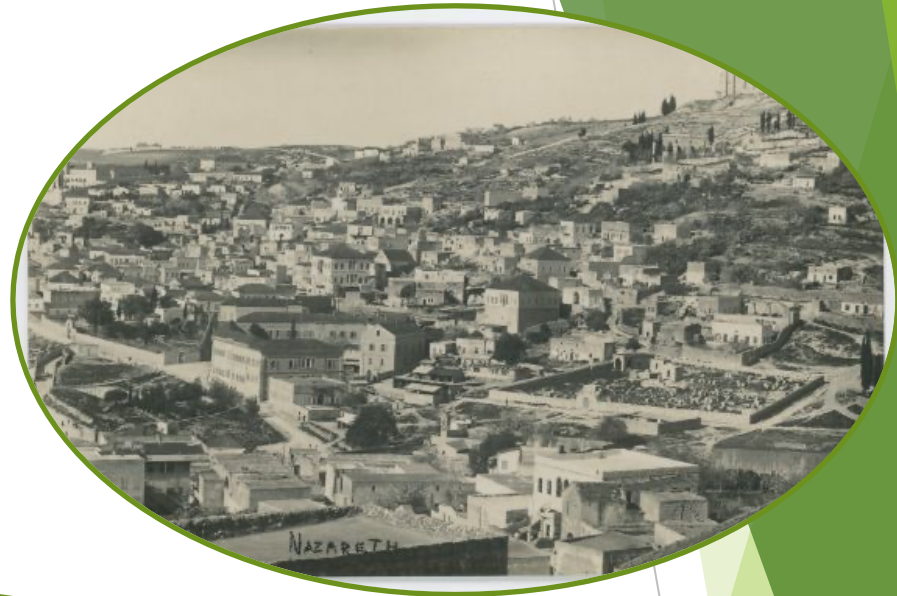
M

H

עברית מודרנית מתהווה ירושלים

Jerusalem Corpus

Emergent Modern Hebrew



DH seems a hodgepodge of methods, approaches, and disciplinary perspectives, bound together more by individualized understandings of both 'digital' and 'humanities' rather than by a common objective or practice.

Leah Henrickson, "Humanities Computing, Digital Humanities, and Computational Humanities: What's In a Name" (3:AM Magazine October 24, 2019)

קשיי הזהות של התחום ◀ ▶ התחום מתפתח כדיסציפלינה
עצמאית עם אגודות מקצועיות, כתבי עת, כנסים, חברי סגל ועוד

מה זה מדעי הרוח הדיגיטליים?

- שימוש בכלים חישוביים לצורך המחקר המסורתי במדעי הרוח
- לצד פיתוח שאלות ושיטות חדשות לחקר האובייקט ההומניסטי

למה לעשות מדעי הרוח הדיגיטליים?

- איסוף ושימור
- הנגשת הידע הקיים
- דמוקרטיזציה של הידע
- פתיחת מקורות מידע חדשים
- פיתוח כיווני מחקר חדשים
- כי אפשר
- כי אי אפשר שלא

המחשב שיפענח בתוך שניות טקסטים עתיקים רבים יותר משיספיק חוקר בכל חייו

באוניברסיטת חיפה מניסים את ההייטק כדי להורים רוח חדשה ללימודי ההיסטוריה, המקרא והספרות. פרופ' צביקה קופליק: "הזמנים שבהם חוקר קורא ערימות של ספרים עומדים להסתים"



<https://www.haaretz.co.il/news/education/.premium-MAGAZINE-1.8408439>

The New York Times

Computer Network Piecing Together a Jigsaw of Jewish Lore



Roni Shweka, left, and his father, Yaacov Choueka, displaying a computerized fragment, one of about 100,000 collected over 1,000 years, that document Jewish life along the Mediterranean. Rina Castelnovo for The New York Times

By Jodi Rudoren

May 26, 2013

<https://www.nytimes.com/2013/05/27/world/middleeast/computers-piecing-together-jigsaw-of-jewish-lore.html>

איך לעשות מדעי הרוח הדיגיטליים?

- איך עוברים מהחומר למסך המחשב?
- איך מתרגמים שאלת מחקר במדעי הרוח להוראות שמכונה מבינה?
- מהו בכלל דאטה במדעי הרוח?
- איך אוספים, מעבדים, מנתחים, ומציגים דאטה במדעי הרוח?
- איך מתקפים תוצאות במדעי הרוח?
- איך מתרגמים תוצאות כמותניות לתובנות במדעי הרוח?

קריאה רחוקה

Canon/Archive. Large-scale Dynamics in the Literary Field

Sarah Allison

Ryan Heuser

Hannah Walser



Literary Lab Pamphlet 11

January 2016

of discourse to the formal gravity of the written page: many nouns, many adjectives, and as few inflected verb forms as possible (Figures 6.3–6.4–6.5).⁴⁰

So far, we have explained the affinity between high type-token ratio and the written register as the result of, loosely speaking, stylistic and ideological choices. But there is also a more neutral, "functional" reason for their correlation. In the findings of corpus linguistics, maximum lexical variety is consistently associated with news: a discourse which needs "an extremely high density of nominal elements", the *Longman Grammar* points out, in order to "refer to a diverse range of people, places, objects, events, etc." (53-54). There is a double source for lexical variety in news: the first is the necessary specificity internal to each distinct news item; the second, the utter discontinuity between one item and the next: as each article or correspondence begins, repetition is "reset" near zero, and type-token ratio can rise accordingly. This twofold logic returns in fictional texts with high type-token ratio: they include plenty of disparate materials, and further accentuate their diversity by using a plurality of generic forms. Jane West, six of whose novels are in the corpus' top 3% for type-token ratio, quotes poetry in 17 of her 24 top-ranked segments; in the absence of poetry, she turns to elaborate metaphors ("expect a fearful tempest to arise, which will clear the tree of its unsound branches"), and even pastiche.⁴¹ William North's introduction to *The Impostor* – half literary criticism, half apologia – discusses a wide range of topics, and includes an excursus on...the wide range of topics he has decided to

39 Underwood and Sellers, p. 14.

40 The high frequency of nouns and adjectives takes us back to the 'grammatical bigrams' discussed at the end of section 4: the 'adjective-adjective', 'proper noun-noun', 'noun-adjective' word pairs. By combining those results with what emerged in this section, we can finally solve the paradox of texts with high redundancy at the level of bigrams, and high variety at that of type-token ratio. The 'labeling' of the bigram *the iron will* is a good example. The noun *iron* is a high frequency, low-entropy word, and the adjective *will* is a low frequency, high-entropy word. The redundancy of 'iron will' and 'clever little', can easily repeat themselves in the course of the novel, thus raising redundancy as measured at *that scale*, but even a mediocre writer is unlikely to repeat 'clever little' within a 1,000-word window, thus leaving a type-token ratio quite high. And the opposite will happen with the 'determiner-noun' or 'preposition-determiner' bigrams that are typical of canonical texts: as 'the' is a high frequency, low-entropy word, it will be repeated many times within the 1,000-word segment, thus lowering its lexical variety; but since the noun next to the article can easily vary, redundancy at the level of bigrams will remain relatively low.

41 "First Venus, queen of gentle devices! taught her prototype lady Arabella the use

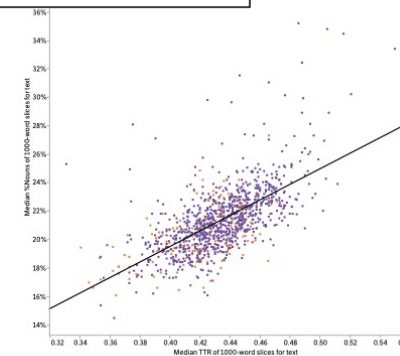
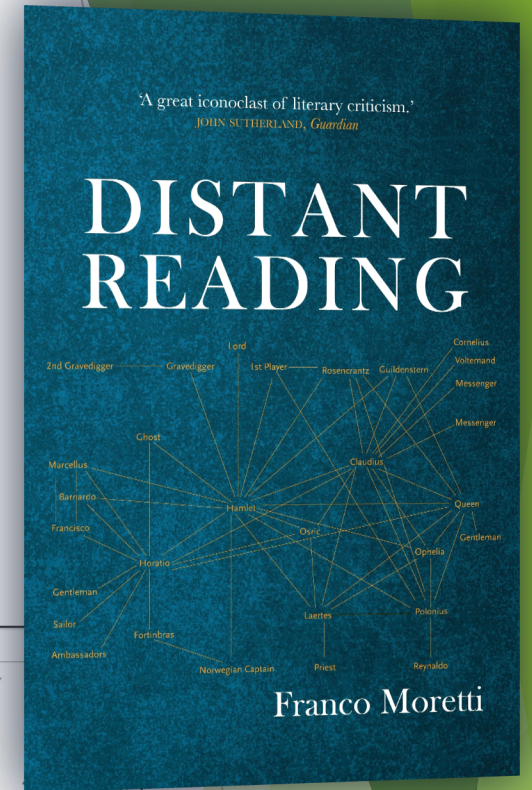
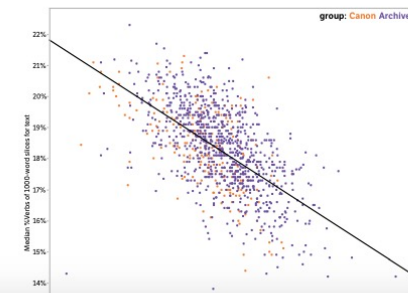
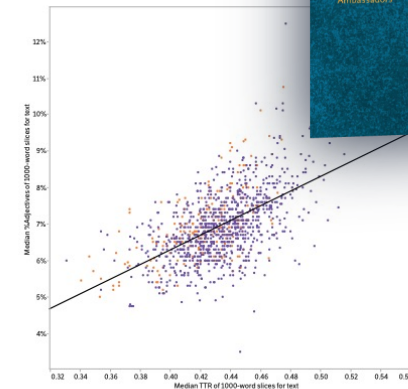


Figure 6.3 (top left)Type token ratio and nouns

Figure 6.4: (top right) Type token ratio and adjectives

Figure 6.5: (bottom right) Type token ratio and verbs

In Hawker's Gibraltar passage, in **Figure 6.2**, adjectives (and participles) are three times as frequent, and inflected verbs three-four times less frequent than the average in nineteenth-century fiction. By contrast, the Adam Bede passage in **Figure 5.4** contains only four nouns and one adjective – “hungry” – in 75 words.



פרנקו מורטי, השערות על ספרות העולם

אולי זה יותר מדי, להתמודד בעת ובעונה אחת גם עם ספרות העולם כולו וגם עם הלא-נקרא. אבל אני באמת חושב שזו ההזדמנות הגדולה שלנו, מפני שהממדים העצומים של המשימה מבהירים לנו שספרות העולם אינה יכולה להיות כמו כל ספרות, רק גדולה יותר; שאיננו יכולים להמשיך לעשות מה שעשינו עד היום, אבל בהיקף רחב יותר. הגישה שלנו צריכה להיות שונה. הקטגוריות צריכות להיות שונות. כפי שכתב מקס ובר, "לא הזיקה ה'ממשית' בין 'דברים' מגדירה את תחומי המדעים השונים, כי אם הזיקה הקונספטואלית בין בעיות. 'מדע' חדש נולד כאשר בעיה חדשה נחקרת באמצעות שיטה חדשה".² זה העניין:

אבל הבעיה בקריאה הקרובה על כל גלגוליה, החל בביקורת החדשה (new criticism) וכלה בדקונסטרוקציה, היא הישענותה, בהכרח, על קאנון מצומצם ביותר. הנחת יסוד זו אולי נעשתה ברבות השנים לא מודעת ולא נראית, אבל היא שרירה וקיימת: אנו משקיעים מאמץ רב כל כך במחקר של טקסטים יחידים אך ורק אם אנו סבורים שמעטים מאוד הטקסטים הראויים לכך. אחרת אין בזה כל היגיון. ואם אנו מעוניינים להתבונן אל מעבר לקאנון (וכך תעשה ספרות העולם, כמובן; יהיה זה אבסורד אם לא תעשה כך!), הקריאה הקרובה לא תוכל לשרת אותנו. היא לא תוכננה לעשות זאת, היא תוכננה לעשות את ההפך. במהותה, היא תרגיל תיאולוגי – התייחסות חגיגית מאוד למספר מצומצם מאוד של טקסטים, הנלקחים ברצינות רבה מאוד – בעוד שהדבר שאנו זקוקים לו הוא עסקה קטנה עם השטן: אנחנו יודעים איך לקרוא טקסטים, כעת הבה נלמד כיצד לא לקרוא אותם. קריאה רחוקה: קריאה שבה המרחק, אם יורשה לי לחזור על עצמי, הוא תנאי לידע, שכן הוא מאפשר לנו להתמקד ביחידות שהן קטנות בהרבה או גדולות בהרבה מן הטקסט עצמו: תחבולות, תמוות, טרופים (tropes) – או ז'אנרים ומערכות שלמות. ואם הטקסט עצמו נעלם בין היחידות הקטנות לאלה הגדולות, הרי זה אחד מאותם מקרים שבהם ניתן לומר בצדק שפחות זה יותר. אם אנו חפצים להבין את המערכת בכללותה, עלינו להשלים עם האובדן של משהו ממנה. אנחנו תמיד משלמים מחיר עבור ידע תיאורטי: המציאות עשירה עד לאינסוף, המושגים מופשטים ודלים. אבל דווקא "עונים" הוא המאפשר להשתמש בהם כדי לדעת. לכן פחות הוא בעצם יותר.¹⁷

איתי מרינברג - מיליקובסקי

מחקרים רבים שנערכו בעשרים השנים האחרונות בכמה מספרות העולם, ובעיקר בספרות האנגלית, מראים כיצד הקריאה ה'חישובית' בטקסטים אינה מייתרת אמנם, חלילה, את הקריאה האנושית בהם, כי אם משלימה, מעשירה, ולעתים גם מאתגרת אותה. המחשב יכול להתמודד עם טקסטים בהיקף גדול יותר, לתמוך ולאשר אמפירית אינטואיציות מחקריות – או לחלופין, לסייג ולשלול אותן; הוא חושף תופעות שעין החוקר מתקשה בראייתן, בשל המגבלות האנושיות הטבעיות שבקריאה לינארית ובזיכרון מוגבל; הוא מציע תשובות מקוריות לשאלות ישנות, ובה בעת מעורר שאלות שטרם נשאלו.

אופיר מינץ - מנור

הגישה הכמותית-הממוחשבת יכולה לתרום למחקר הספרותי באופן משמעותי הן על ידי הארה של היבטים שקשה לראותם או לתארם בדרך אחרת, הן על ידי הרחבת ארגז הכלים שעומד לרשות החוקרים. הכלים הממוחשבים והאסטרטגיות החישוביות הופכים שמישים יותר משעה שאנחנו מודעים למגבלותיהם ולהטיותיהם; הרפלקסייה המחקרית ההומניסטית היא המכריעה כאן, בדומה לרפלקסייה הנדרשת מכל פעולה פרשנית אחרת, כולל הפרקטיקות המחקריות המסורתיות ביותר.

אומנם לעיתים הכלים הממוחשבים מוצגים – במיוחד על ידי חוקרים מתחום מדעי המחשב – כאובייקטיוויים וככלים שבכוחם לנטרל הטיות או חולשות אנושיות שונות, אולם כחוקרים במדעי הרוח אנו יודעים היטב כי הטענה לאובייקטיוויות – לרבות אובייקטיוויות של חוקרים בשר ודם – מורכבת לאין שיעור ממה שנהוג לחשוב.

מה אנחנו קוראים בקריאה רחוקה? או: מה זה דאטה?

רבים מעמיתי בלימודי הספרות והתרבות לא בהכרח ידברו על מושאי המחקר שלהם כעל 'נתונים'. אם תשאל אותם מה הם חוקרים, הם ידברו על אודות ספרים, ציורים וסרטים; על אודות דרמה ורומן-פשע... על אודות הקולנוע הגרמני האקספרסיוניסטי ועל אודות הקומדיה הרומנטית...; אפשר שהם ידברו על מושאי מחקרם כעל טקסטים, תמונות וצלילים. אך רק לעתים רחוקות בלבד הם יתייחסו אליהם כאל 'נתונים'. אבל, במדעי הרוח, כמו בתחומי מחקר אחרים, אנו עוסקים יותר ויותר ב'נתונים'. ככל שנמשכים מאמצי הדיגיטציה בעולם, במגזר הפרטי והעסקי, כך מצטברים יותר ויותר נתונים הרלוונטיים לתחומי המחקר שלנו, ותחת הרשאה מתאימה הם זמינים למחקר.

Article

Big? Smart? Clean? Messy? Data in the Humanities

December 2013

Authors:



Christof Schöch
Universität Trier

דאטה במדעי הרוח

- Data in the humanities could be considered a digital, selectively constructed, machine-actionable abstraction representing some aspects of a given object of humanistic inquiry
- Data (as well as the tools with which we manipulate them) add complexity to the relation between researchers and their objects of study

מושגי יסוד בעבודה עם דאטה

- דאטה = נתונים (מה שנתון | datum) התופעות הנצפות שמעניינות אותנו
- דאטה מוגדר על ידי העדר האחדה כלשהי
- רק כאשר נותנים לדאטה מבנה ומשמעות הוא הופך להיות מידע או אינפורמציה
- דאטה בלתי מובנה, חצי מובנה, מובנה (plain text) הוא בלתי מובנה; דאטה בבסיס נתונים הוא מובנה
- מטא דאטה = הדאטה על הדאטה

עדויות כדאטה – מודל של קריאה מדורגת



מידול נושאי של משפט אייכמן

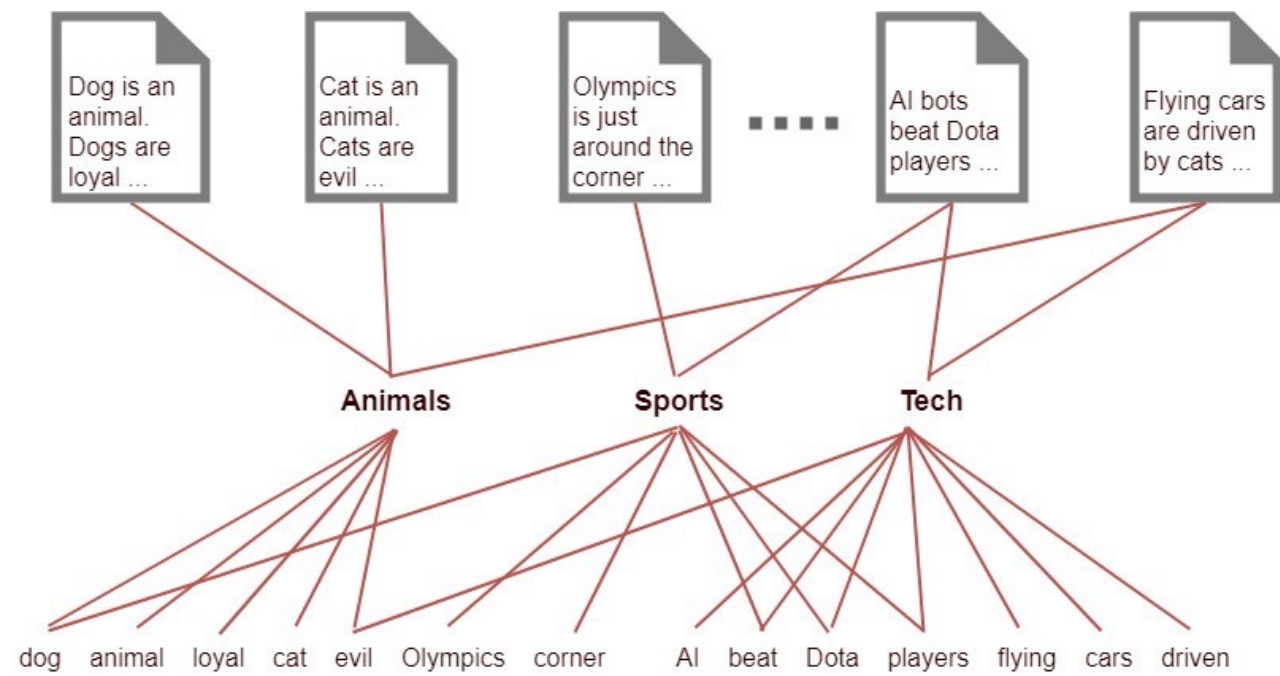
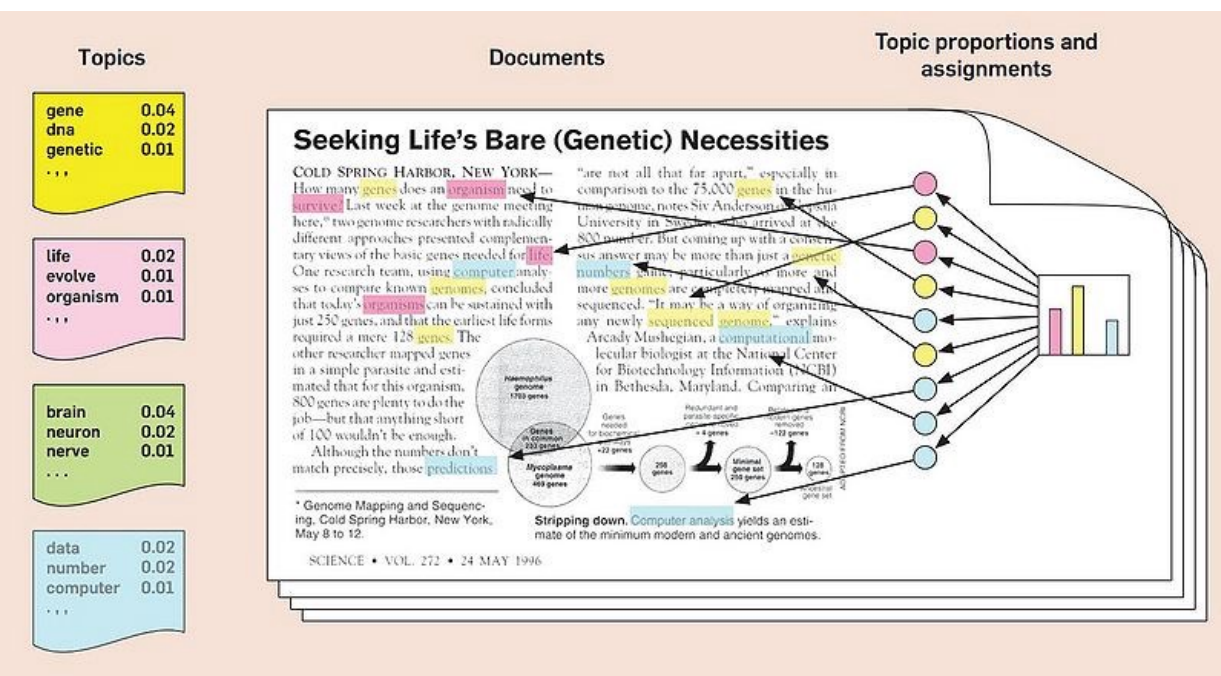
- האתגר: איך להקשיב לעדויות באופן שמצליח לתפוס את המכלול מבלי לאבד את קולו של היחיד
- המתודולוגיה: מידול נושאי topic modeling של כל קורפוס המשפט



Topic Modeling

- כלי עיבוד שפה טבעית NLP
- נועד לחקור כמות גדולה של דאטה לא מתויג ולא מובנה
- מנתח מסמכים דרך זיהוי הנושאים "החבויים" בקורפוס
- נושא (Topic) – מקבץ מילים בעלות הסתברות גבוהה להופיע ביחד במסמך
- אלגוריתם LDA (Latent Dirichlet Allocation) מניח שכל מסמך בקורפוס מורכב מתערובת נושאים
- המודל "חושף" את הנושאים שמרכיבים את הקורפוס ואת התפלגות הנושאים בכל מסמך





Blei et al., Probabilistic Topic Models (2003)

Topic modeling the NYT

	Topic 1	Weight	Topic 2	Weight	Topic 3	Weight	Topic 4	Weight	Topic 5	Weight
0	life	0.018076	father	0.059603	official	0.017620	case	0.021908	art	0.010555
1	man	0.017714	graduate	0.048363	force	0.015388	law	0.020698	open	0.010413
2	woman	0.016657	son	0.042746	military	0.014587	court	0.019967	room	0.010363
3	book	0.010486	mrs	0.041379	war	0.011381	lawyer	0.016935	house	0.009002
4	family	0.010382	daughter	0.037156	government	0.010564	state	0.014501	building	0.008722
5	young	0.009896	mother	0.034542	troop	0.008949	judge	0.012487	artist	0.008264
6	write	0.009493	receive	0.029211	attack	0.008886	legal	0.011141	design	0.008162
7	child	0.009460	marry	0.029038	leader	0.008082	rule	0.009854	floor	0.008034
8	live	0.008819	yesterday	0.024107	peace	0.006835	decision	0.009261	museum	0.007917
9	love	0.007814	degree	0.022899	soldier	0.006562	file	0.008289	exhibition	0.007222
	Topic 6	Weight	Topic 7	Weight	Topic 8	Weight	Topic 9	Weight	Topic 10	Weight
0	group	0.051052	market	0.024976	serve	0.010918	change	0.007661	city	0.021776
1	member	0.040683	stock	0.024874	add	0.010185	system	0.007233	area	0.014865
2	meeting	0.016390	share	0.020583	minute	0.009301	problem	0.006835	build	0.014361
3	issue	0.014988	price	0.018141	pepper	0.009235	power	0.005400	building	0.014326
4	official	0.013069	sell	0.016564	oil	0.008976	create	0.005056	home	0.013632
5	support	0.011994	buy	0.015415	cook	0.008711	research	0.004712	resident	0.013483
6	leader	0.011799	company	0.015249	food	0.008689	produce	0.004574	community	0.012479
7	organization	0.011135	investor	0.015062	cup	0.008682	far	0.004447	local	0.010686
8	meet	0.010235	yesterday	0.012813	sauce	0.008209	result	0.004280	live	0.010661
9	effort	0.008479	analyst	0.010768	small	0.007864	kind	0.004166	project	0.010459

יתרונות

- מעבד שפה טבעית בלי לאבד את השפה
- שיטה בלתי מונחית: אין התערבות של החוקר בשלב החישובי של חשיפת הנושאים החבויים
- אין צורך בתיוגים: קושי בקידוד כמויות גדולות של חומר, הטיות מחקריות, פערים בין מקודדים, מורכבות השאלות
- שקיפות מתודולוגית גבוהה

אתגרים

- Bag of Words
- הנחות מוקדמות של האלגוריתם
- עבודה עם עברית

Topic 0



Topic 1



Topic 2



Topic 3



Topic 4



Topic 5



Topic 6



Topic 7



Topic 8



Topic 9



סגור גדול חקירה תביעה זהו
מקור סעיף
ההצעה
היסלר תצהיראיה
להגיש
תמונה
שופט
תרגום לס' ישראל

[illegible]

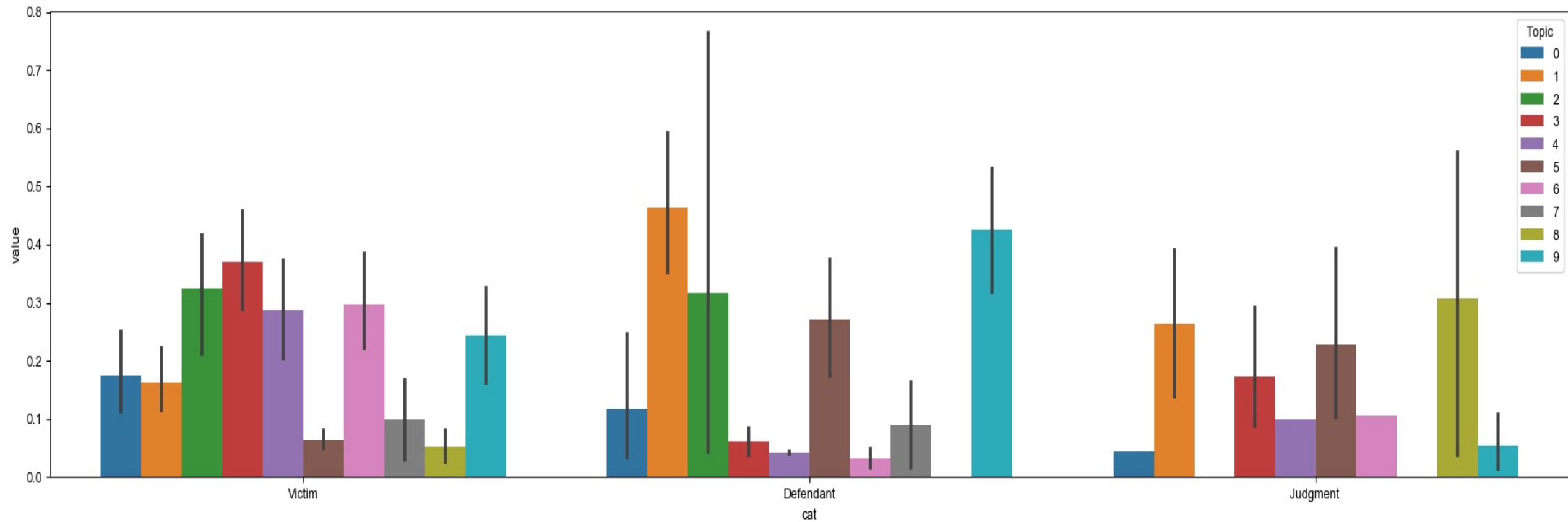
ויגה
בכתב
חוק
הודעה
הגירה
סגור
דברי
ציוני
זאסן
מכתב
חקירה
ארץ ישראל
תורה

[illegible][illegible][illegible]

היינו זוכה תמונה

זוכר הונגרי אמר
אכיל
שיחה
בכר
מילך
אמרת
קסמור
הימלך
תשובה
פקודה
בדף
פשוט
קמע
חכירה

- 0 – Court Procedure
- 1 – Nazi Bureaucracy
- 2 – Negotiations
- 3 – Deportations
- 4 – Poland Jewry
- 5 – Final Solution
- 6 – Annihilation
- 7 – Eichmann Statement
- 8 – Legal Indictment
- 9 – Hungary Jewry



0 – Court Procedure

1 – Nazi Bureaucracy

2 – Negotiations

3 – Deportations

4 – Poland Jewry

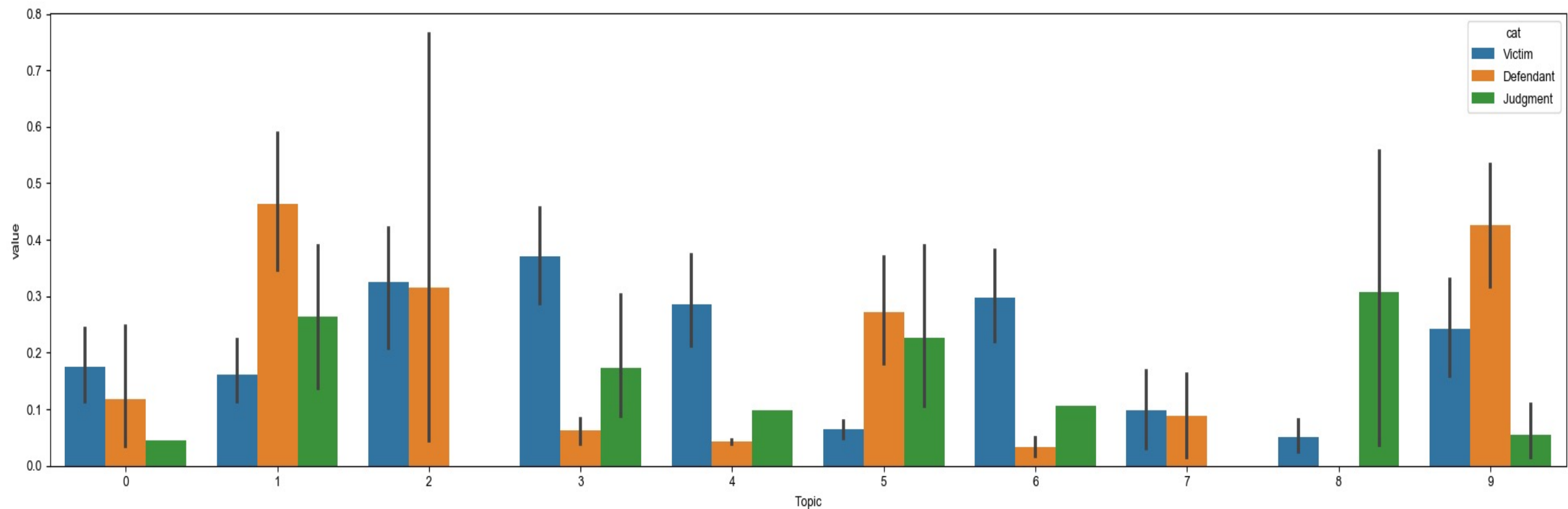
5 – Final Solution

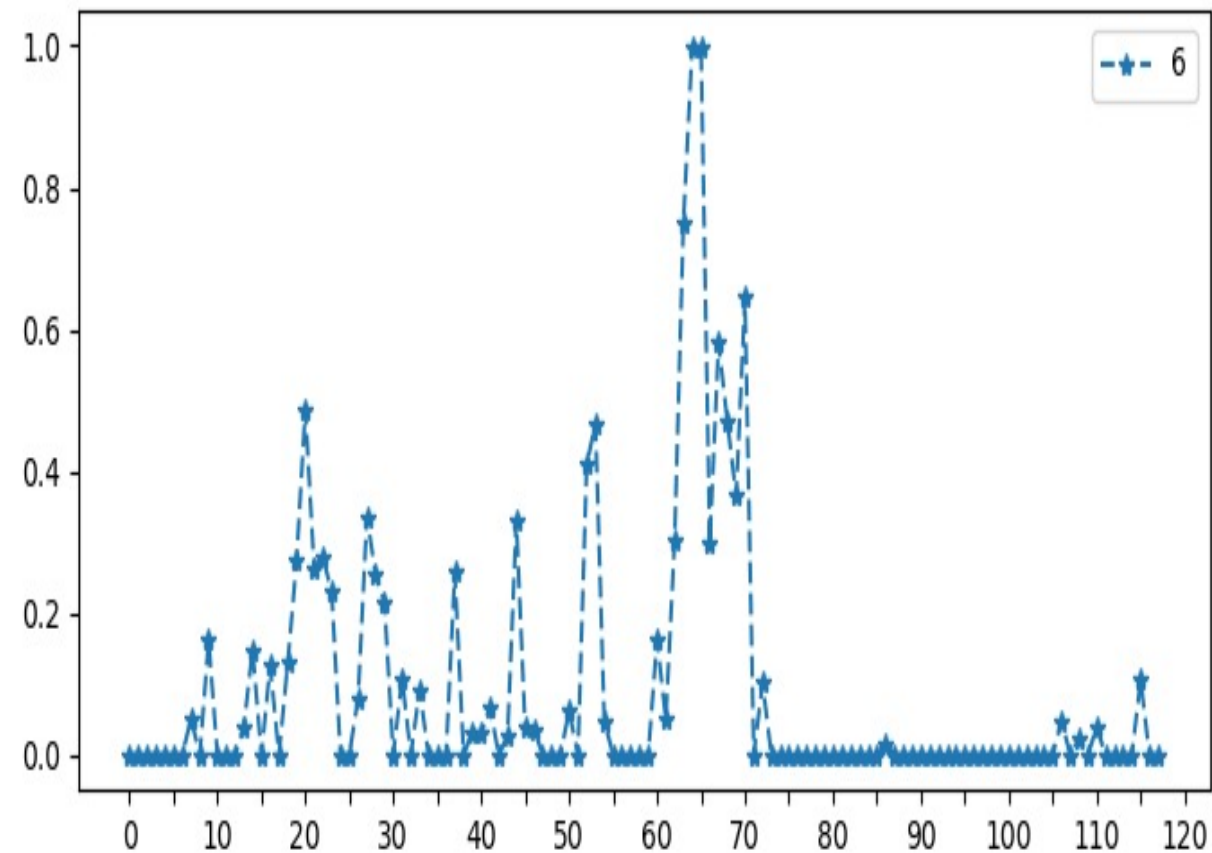
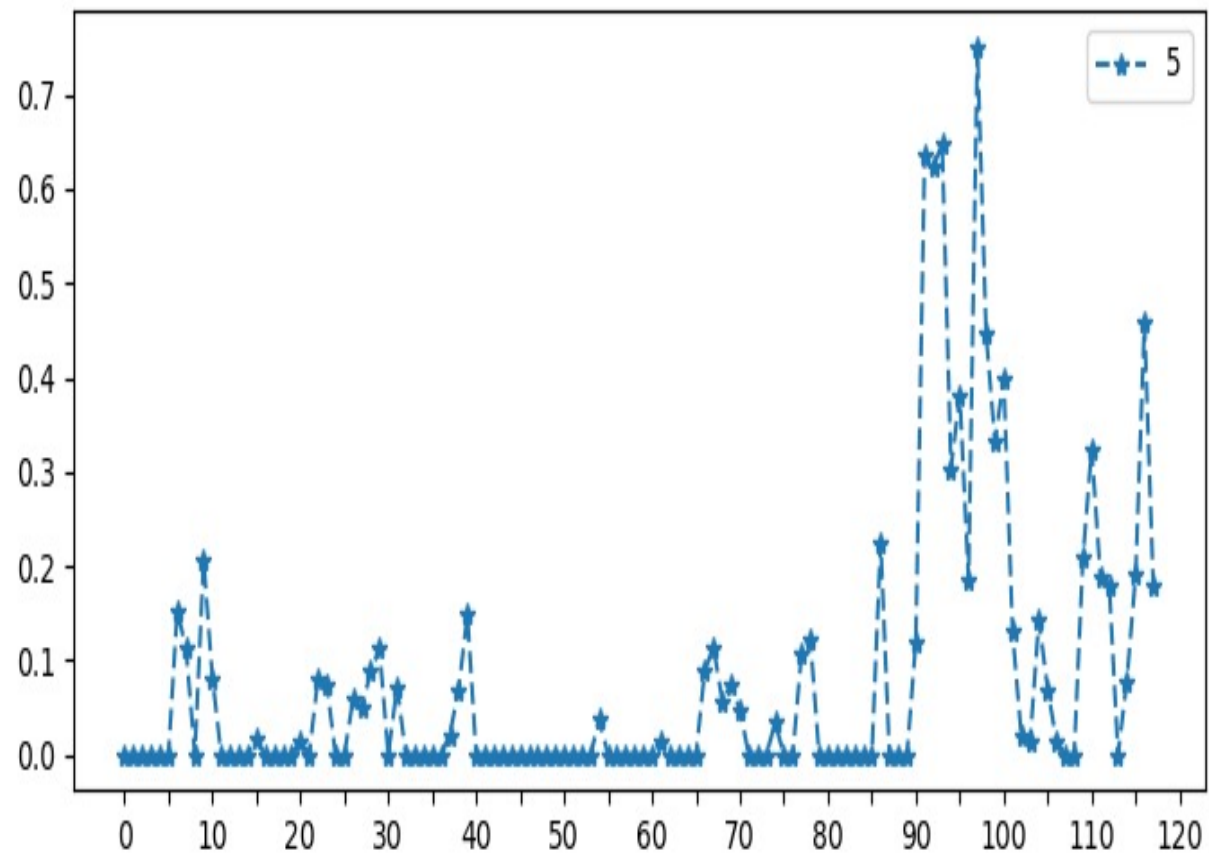
6 – Annihilation

7 – Eichmann Statement

8 – Legal Indictment

9 – Hungary Jewry





היידריך אושוויץ הימלר רזה
 גלובוק
 פקודה
 השמדה
 סופי
 פתרון
 מחלקה
 מילר
 ג'יקה
 שטח
 מזורח
 4BVI
 גינתרמשל
 הריי

היינו
 אוכל אושוויץ אמרו
 ראיית
 לתוך צריף
 בלוק
 לילה
 גזתמונה
 בתוך
 רכבת כמה אנשי קרה
 בזיוק
 בערך
 ראיתי