תרגיל בית 2 - קולוקציות

חיים שאללשוילי 200832780

צחי כפיר 200681476

הבדלים בין התוצאות עבור מדדים שונים

מדד Frequency ממיין את זוגות המילים על פי תדירותם בטקסט. כצפוי במדד זה, צפו מעלה זוגות מילים לא מעניינות הכללו סימני פיסוק וניתן להן ציון גבוה שכן תדירותם הייתה גבוהה.

מדד T Score מתבסס על המרחק של הסתברות הביטוי בטקסט מההסתברות המתקבלת מהכפלת ההסתברויות של שני הטוקנים בנפרד תחת התחשבות בגודל הטקסט והשונות . באופן טבעי מדד זה מעניק ציון גבוה יותר לביטוי שחזר מספר רב של פעמים אך המילים פחות נפוצות בקורפוס. לכן הציג שיפור בנושא הנ"ל והביטויים שהכילו סימני פיסוק ירדו בדירוגם.

מדד PMI מתבסס על כמות המידע המתווסף ממילה מסוימת על כמות המידע שקיים מהמילה הקודמת. כאשר ההסתברות שזוג מילים יופיע גדולה מההסתברות שכל אחת מהמילים תופיע אנו מקבלים תוצאת PMI גבוה יותר. PMIמציפה למעלה אירועים מאד נדירים ולכן התקבלו תוצאות רבות הכללו שנים, שמות באנגלית וכולה. כאשר משתמשים ב-PMI מומלץ להסתכל על צירופי המילים שתדירותם עולה מעל סף מסוים.

תוצאות Raw – 100 הביטויים המובילות בכל מדד

ביטוי	מיקום Frequency	מיקום T Score	מיקום PMI	הסבר
על ,	19	לא במאה הראשונים	לא במאה הראשונים	זוג המילים קיבל תוצאת frequency גבוה מאד למרות שאיננו מעניין כלל וזאת בזכות תדירותם הגבוהה בטקסט של שני הטוקנים בנפרד. בשני המדדים האחרים זוג המילים כלל לא נכנס לרשימת 100 הקולוקציות המשמעותיות עקב סינון המקרים הללו בהם תדירות הטוקנים בטקסט גבוהה
1. **על שם** 2. **לקריאה נוספת** 3. **בית**	1. 97 2. 91 3. לא במאה הראשו	1. 68 2. 54 3. 69	לא במאה הראשונים	דוגמא לזוגות מילים נפוצות אשר דירוגם ב-T Score גבוהה יותר מאשר דרוג התדירות בשל כניסת כל הביטויים הלא מענייניים הכללו סימני

המשפט	נים			פיסוק שנכנסו לרשימת הזוגות הנפוצות ביותר. במדד PMI לא נכנס לרשימת ה-100 הראשונים עקב כניסה של זוגות מילים נדירות אשר תפסו את המקומות הראשונים.
1. 1977, 1980, **2. Alain Delon**	לא במאה הראשונים	לא במאה הראשונים	1. 1 2. 67	התוצאה בעלת דירוג ה-PMI הגבוה ביותר בעקבות התדירות הנמוכה של הטוקנים בקורפוס. ברור שביטוי זה לא מעניין כלל והוא דוגמא מצוינת לרעש שמתקבל ממדד ה-PMI. עבור שני המדדים האחרים ברור כי בשל הנפוצות הנמוכה של שני הטוקנים ושל הביטוי המשותף הנוצר משניהם הביטוי לא נכנס לרשימת ה-100 הגבוהים ביותר במדד.

תוצאות Select – רשימת הביטויים בעלי 20 מופעים בקורפוס

שוב מדד ה-frequency הציף תוצאות רבות לא מעניינות הכללו סימני פיסוק. מדדי הPMI וה-T Score הציגו תוצאות טובות ודומות מאד (כמעט כל הביטויים קיבלו את אותו הדירוג).

ביטוי	**מיקום Frequency**	**מיקום T Score**	**מיקום PMI**	**הסבר**
" הארי **" עין**	1 2	96 81	96 81	זוג המילים קיבל תוצאת frequency גבוה מאד למרות שאיננו מעניין כלל וזאת בזכות תדירותם הגבוהה בטקסט של שני הטוקנים בנפרד. בשני המדדים האחרים זוג המילים קיבל את אותו המיקום
Mortal Kombat **ברי סחרוף**	45 70	1 4	1 4	דוגמא לזוגות מילים נפוצות אשר דירוגם ב-T Score וב-PMI גבוהה יותר מאשר דרוג התדירות בשל כניסת כל הביטויים הלא מעניינים הכללו סימני פיסוק שתפסו את

נשיונל גיאוגרפיק	128	6	6	המקומות הראשונים בדירוג

האם רוב הזוגות בראש הרשימה משמעותיים\מעניינים?

רוב הזוגות בראש הרשימה של המדדים לא משמעותיים. בכל מדד בא לידי ביטוי חסרון אחר:

- Frequency - צפו מעלה זוגות מילים לא מעניינים הכללו סימני פיסוק וניתן להם ציון גבוה שכן תדירותם הייתה גבוהה לדוגמא:
  - ב – מקום 1
  - . (מקום 2
  - , " מקום 4
  - , (מקום 5
  - אך , מקום 8
- T Score – התקבלה תוצאה דומה אך במינון נמוך יותר ויותר תוצאות משמעותיות עלו בדירוגם. לדוגמא:
  - "על ידי" עלה ממקום 3 ב-frequency למקום 2 ב-T Score
  - "קישורים חיצוניים" עלה ממקום 9 ב-frequency למקום 5 ב-T Score
  - "מלחמת העולם" עלה ממקום 67 ב-frequency למקום 40 ב-T Score

  ישנם גם כמה מקרים של תוצאות לא משמעותיות שעלו בדירוגם כמו " אך , " שעלה למקום ה-6 מהמקום ה-8 ב-Frequency
- PMI – צפו מעלה זוגות רבים שלא נפוצים בקורפוס ולכן התקבלו תוצאות רבות של מילים באנגלית ומספרים. לדוגמא:
  - "1980, 1977," במקום הראשון
  - "טייגר 10.4" במקום ה-26

ברשימה מופיעות גם מספר קולוקציות בשפה האנגלית, בעיקר שמות, אך הן מופיעות בגלל התדירות הנמוכה שלהם בטקסט (הכתוב בעברית). לדוגמא:

- "ALL STAR" במקום ה-51
- "Alain Delon" במקום ה-67

שיפורים:

1. קולוקציות המכילות סימני פיסוק אינן מעניינות, ולכן רצוי לסנן אותן. דרך אחת לעשות זאת היא לסנן את כל הקולוקציות שמכילות token אחד או יותר שמכיל תו שאיננו אות. לדוגמא: ") . " , " הילד) " וכדומה.
   אך, ישנם כמה תוים שאינם אותיות בעברית שאותם לא נרצה לסנן,
   קולוקציה המכילה token עם התו ' - ' לדוגמא: בן-גוריון.
   קולוקציה המכילה token עם התו ' " ' לדוגמא: צה"ל.
   קולוקציה המכילה token עם התו " ' " לדוגמא: ג'יפ.
   קולוקציה המכילה token עם התו ' . ' לדוגמא: מ.ש.ל.
   כמובן ששינוי זה עזר לשפר את תוצאות כל המדדים, מסנן הרבה מאוד קולוקציות שאינן מעניינות. אך משפיע בעיקר על מדד ה-frequency כיוון שסימני פיסוק מופעים בתדירות מאוד גבוהה, אם נתבונן ב-100 הקולוקציות בעלות מדד ה-frequency הגבוהה ביותר נבחין כי רוב הקולוקציות מכילות סימן פיסוק אחד או יותר.
2. סינון קולוקציות המכילות מספרים. רוב מוחלט של הקולוקציות המעניינות בעברית אינן מיכלות ספרות, לכן נרצה לסנן token שמכיל ספרות, לדוגמא: שנים, תאריכים, כמויות וכדומה. כמובן ששינוי זה עזר לשפר את תוצאות כל המדדים, מסנן הרבה מאוד קולוקציות שאינן מעניינות. אך משפיע בעיקר על מדד ה-pmi, כיוון שעל פי מדד זה תדירות נמוכה

גוררת ציון גבוהה, ותדירות מספרים לרוב נמוכה, (תאריך/כמות מסויימת לא יופיעו הרבה פעמים).

3. נרצה לסנן token באורך אחד, בניגוד לאנגלית שם ישנו token באורך אחד בעל משמעות, (a) בעברית אין קולוקציות מעניינות בעלות token באורך אחד. לדוגמא נרצה לסנן את הקולוקציה "בעוד כ" (יכולה להיווצר אחרי הפרדה של המשפט "בעוד כ-5 שנים") השינוי אכן שיפר במקצת את תוצאות כל המדדים, שיפר פחות או יותר במידה שווה.
4. מדד ה- pmi מושפע מאוד מתדירות ה-token. רוב מוחלט של 100 המילים בעלות ציון ה-pmi הגבוהה ביותר בעלות תדירות של לא יותר משני מופעים. דבר שגורר הרבה קולוקציות המכילות מילים באנגלית (בטקטסים בעברית כמובן), קולוקציות המכילות מספרים וכדומה. כמובן שהקולקציות הנ"ל אינן מעניינות ונרצה לסננם.
לכן בחישוב מדד ה-pmi נרצה לבחון קולוקציות בעלות מספר מופעים מעל סף מסויים.
השינוי שיפר משמעותית את 100 תוצאות מדד ה-pmi הגבוהות ביותר. (כמובן שלא שיפר את תוצאות שאר המדדים)

דוגמאות לשיפור:

- מיקום הקולוקציה "ערש דווי" לפני השינוי על פי מדד pmi לא היה בין 100 הראשונים ולאחר השינוי מיקומה הוא 97.
- מיקום הקולוקציה "על פני" לפני השינוי על פי מדד frequency לא היה בין 100 הראשונים ולאחר השינוי מיקומה הוא 30.
- מיקום הקולוקציה "על גבי" לפני השינוי על פי מדד t-test לא היה בין 68 הראשונים ולאחר השינוי מיקומה הוא 30.
- הרבה קולוקציות לא מעניינות שמכילות סימני פיסוק לפני השיפור הופיעו בין ה-100 בעלי הדירוג הגבוהה ביותר בכל המדדים, ולאחר השיפור סוננו.

נספחים:

תוצאות לאחר שיפורים:

RawFrequency_raw.txt					
Rank	Collocation	Score	Rank	Collocation	Score
1	על ידי	2.23	51	על כך	0.1
2	קישורים חיצוניים	1.13	52	בין היתר	0.1
3	math formula	0.73	53	של כל	0.1
4	על פי	0.62	54	ביותר של	0.1
5	ארצות הברית	0.4	55	זמן קצר	0.1
6	על מנת	0.39	56	גם את	0.1
7	בדרך כלל	0.38	57	לארץ ישראל	0.09
8	לאחר מכן	0.37	58	בין השנים	0.09
9	הערות שוליים	0.36	59	בנו של	0.09
10	תל אביב	0.36	60	כי הוא	0.09
11	ראו גם	0.35	61	של המדינה	0.09
12	יחד עם	0.35	62	על כל	0.09
13	כמו כן	0.35	63	של העיר	0.09
14	בארצות הברית	0.26	64	הוא היה	0.09
15	מלחמת העולם	0.22	65	העולם הראשונה	0.09
16	מאוחר יותר	0.2	66	בית הכנסת	0.09
17	עם זאת	0.2	67	באותה תקופה	0.09
18	את כל	0.2	68	המשפט העליון	0.08
19	העולם השנייה	0.2	69	הרב קוק	0.08
20	לקריאה נוספת	0.18	70	לא הייתה	0.08
21	על שם	0.17	71	בניו יורק	0.08
22	בית המשפט	0.16	72	על רקע	0.08
23	של המאה	0.15	73	ביחד עם	0.08
24	בארץ ישראל	0.15	74	למנות את	0.08
25	בית הספר	0.15	75	לעומת זאת	0.08
26	הראשון של	0.15	76	של הרב	0.08
27	בתל אביב	0.14	77	זה היה	0.08
28	בין השאר	0.13	78	על גבי	0.08
29	על אף	0.13	79	של הלהקה	0.08
30	על פני	0.13	80	הראשונה של	0.07
31	של ארצות	0.13	81	לארצות הברית	0.07
32	פרס נובל	0.12	82	זכה בפרס	0.07
33	ולאחר מכן	0.12	83	מותו של	0.07
34	של דבר	0.12	84	ניתן למנות	0.07
35	עד היום	0.12	85	קודם לכן	0.07
36	אם כי	0.12	86	ברחבי העולם	0.07
37	אך לא	0.12	87	הוקם בשנת	0.07
38	באותה שנה	0.11	88	השני של	0.07
39	דמותו של	0.11	89	פעמים רבות	0.07
40	ככל הנראה	0.11	90	שנים רבות	0.07
41	לא היה	0.11	91	למועצה אזורית	0.07
42	בסופו של	0.11	92	ניתן למצוא	0.07
43	על שמו	0.11	93	ניו יורק	0.07
44	כדור הארץ	0.11	94	בשנים האחרונות	0.07
45	בבית הספר	0.11	95	הוא לא	0.07
46	ארץ ישראל	0.11	96	הקיבוץ המאוחד	0.07
47	שמו של	0.11	97	זה הוא	0.07
48	כמו גם	0.1	98	חיל האוויר	0.07
49	הטוב ביותר	0.1	99	את עצמו	0.07
50	מדינת ישראל	0.1	100	לעתים קרובות	0.07

RawFrequency_select.txt					
Rank	Collocation	Score	Rank	Collocation	Score
1	mortal kombat	0.02	36	הר חברון	0.02
2	אותו על	0.02	37	ועל ידי	0.02
3	אין מידע	0.02	38	זה ניתן	0.02
4	אישים משמעותיים	0.02	39	זכתה באליפות	0.02
5	אך היא	0.02	40	חדשות אישים	0.02
6	אך ורק	0.02	41	כמעט כל	0.02
7	אך עם	0.02	42	כן הוא	0.02
8	אלו הם	0.02	43	לא ברור	0.02
9	אף היא	0.02	44	לא הצליחו	0.02
10	אף הם	0.02	45	לשפר את	0.02
11	את המילה	0.02	46	מדי יום	0.02
12	את מקומו	0.02	47	מיד לאחר	0.02
13	את עצמם	0.02	48	מיוצגת האות	0.02
14	אתונה העתיקה	0.02	49	מכן הוא	0.02
15	בבית החולים	0.02	50	מצד אחד	0.02
16	בברית המועצות	0.02	51	מצד שני	0.02
17	בהיסטוריה של	0.02	52	ניתן היה	0.02
18	בחייו של	0.02	53	נשיונל ג'יאוגרפיק	0.02
19	במזרח התיכון	0.02	54	ספריו של	0.02
20	במקרים רבים	0.02	55	עין חרוד	0.02
21	בסוף שנת	0.02	56	עלה לארץ	0.02
22	בסרט זה	0.02	57	ערים חדשות	0.02
23	בקרב על	0.02	58	פני כדור	0.02
24	ברי סחרוף	0.02	59	קבע כי	0.02
25	בשל כך	0.02	60	קרוי על	0.02
26	בתי חולים	0.02	61	שאו קאהן	0.02
27	גרמניה הנאצית	0.02	62	של אוניברסיטת	0.02
28	דבר זה	0.02	63	של הכנסייה	0.02
29	הוא דמות	0.02	64	של הליכוד	0.02
30	הוא על	0.02	65	של הסדרה	0.02
31	הוא קיבל	0.02	66	של הצבא	0.02
32	הוא שם	0.02	67	של כלי	0.02
33	היה בנו	0.02	68	שנות השישים	0.02
34	המבוססת על	0.02	69	תגליות והמצאות	0.02
35	הספר התיכון	0.02			

tTest_raw.txt					
Rank	Collocation	Score	Rank	Collocation	Score
1	על ידי	44.95	51	כמו גם	9.51
2	קישורים חיצוניים	32.24	52	לארץ ישראל	9.3
3	math formula	25.9	53	שמו של	9.26
4	על פי	23.69	54	בין השנים	9.2
5	ארצות הברית	19.29	55	העולם הראשונה	9
6	על מנת	18.9	56	בית הכנסת	8.91
7	בדרך כלל	18.64	57	באותה תקופה	8.88
8	לאחר מכן	18.49	58	המשפט העליון	8.83
9	הערות שוליים	18.21	59	על כך	8.81
10	תל אביב	18.18	60	בנו של	8.8
11	ראו גם	18.04	61	הרב קוק	8.77
12	כמו כן	17.96	62	בניו יורק	8.66
13	יחד עם	17.92	63	ביחד עם	8.56
14	בארצות הברית	15.44	64	לעומת זאת	8.53
15	מלחמת העולם	14.19	65	על רקע	8.48
16	מאוחר יותר	13.5	66	למנות את	8.48
17	העולם השנייה	13.43	67	לא הייתה	8.4
18	עם זאת	13.31	68	על גבי	8.32
19	לקריאה נוספת	13.07	69	כי הוא	8.3
20	את כל	12.17	70	לארצות הברית	8.3
21	בית המשפט	12.05	71	קודם לכן	8.24
22	בארץ ישראל	11.75	72	ניתן למנות	8.24
23	על שם	11.75	73	זכה בפרס	8.24
24	בית הספר	11.67	74	ברחבי העולם	8.17
25	בתל אביב	11.57	75	פעמים רבות	8.11
26	של המאה	11.28	76	הוקם בשנת	8.11
27	בין השאר	11.02	77	של המדינה	8.09
28	פרס נובל	10.72	78	שנים רבות	8.09
29	על פני	10.72	79	למועצה אזורית	8.06
30	ולאחר מכן	10.67	80	ניתן למצוא	8.05
31	עד היום	10.44	81	ניו יורק	8
32	אם כי	10.31	82	הקיבוץ המאוחד	7.93
33	על אף	10.25	83	חיל האוויר	7.93
34	באותה שנה	10.23	84	של העיר	7.93
35	ככל הנראה	10.15	85	בשנים האחרונות	7.93
36	של ארצות	10.04	86	לעתים קרובות	7.87
37	הראשון של	10.04	87	זה היה	7.87
38	כדור הארץ	9.99	88	מותו של	7.83
39	אך לא	9.98	89	of the	7.73
40	בבית הספר	9.92	90	הוא היה	7.71
41	של דבר	9.9	91	בשלב זה	7.71
42	בסופו של	9.9	92	ברית המועצות	7.68
43	ארץ ישראל	9.87	93	ראש הממשלה	7.67
44	דמותו של	9.86	94	בני אדם	7.65
45	על שמו	9.76	95	במהלך מלחמת	7.64
46	הטוב ביותר	9.73	96	של הלהקה	7.64
47	מדינת ישראל	9.67	97	מלחמת העצמאות	7.61
48	בין היתר	9.62	98	בעלי חיים	7.53
49	זמן קצר	9.53	99	זמן רב	7.53
50	לא היה	9.52	100	על כל	7.52

tTest_select.txt					
Rank	Collocation	Score	Rank	Collocation	Score
1	mortal kombat	4.47	36	לשפר את	4.4
2	תגליות והמצאות	4.47	37	כמעט כל	4.4
3	שאו קאהן	4.47	38	את מקומו	4.39
4	ברי סחרוף	4.47	39	היה בנו	4.35
5	אישים משמעותיים	4.47	40	בסרט זה	4.34
6	נשיונל ג'יאוגרפיק	4.47	41	קרוי על	4.34
7	עין חרוד	4.47	42	ועל ידי	4.33
8	חדשות אישים	4.47	43	אלו הם	4.32
9	אתונה העתיקה	4.47	44	הוא דמות	4.32
10	הר חברון	4.47	45	דבר זה	4.3
11	בברית המועצות	4.47	46	אף הם	4.27
12	גרמניה הנאצית	4.47	47	הוא קיבל	4.23
13	מיוצגת האות	4.47	48	את עצמם	4.21
14	ערים חדשות	4.47	49	בהיסטוריה של	4.2
15	במזרח התיכון	4.47	50	בחייו של	4.19
16	שנות השישים	4.47	51	של אוניברסיטת	4.14
17	עלה לארץ	4.47	52	זה ניתן	4.11
18	בתי חולים	4.47	53	ספריו של	4.1
19	פני כדור	4.46	54	אף היא	4.09
20	זכתה באליפות	4.46	55	את המילה	4.08
21	בבית החולים	4.46	56	ניתן היה	4.07
22	במקרים רבים	4.46	57	של הכנסייה	3.93
23	מדי יום	4.46	58	מכן הוא	3.87
24	בסוף שנת	4.46	59	של הסדרה	3.86
25	אין מידע	4.45	60	של הליכוד	3.85
26	הספר התיכון	4.45	61	כן הוא	3.74
27	אך ורק	4.45	62	של הצבא	3.71
28	מצד שני	4.44	63	אך היא	3.65
29	מיד לאחר	4.44	64	בקרב על	3.62
30	מצד אחד	4.44	65	של כלי	3.58
31	קבע כי	4.43	66	הוא שם	3.27
32	לא הצליחו	4.43	67	אך עם	3.18
33	לא ברור	4.43	68	אותו על	2.3
34	המבוססת על	4.42	69	הוא על	-7.35
35	בשל כך	4.41			

PMI_raw.txt					
Rank	Collocation	Score	Rank	Collocation	Score
1	sensuality devotion	17.82	51	mass start	16.91
2	she wants	17.82	52	organizational culture	16.91
3	באגס באני	17.82	53	arcade treasures	16.82
4	בהקתיודאנתה סואמי	17.82	54	unknown pleasures	16.82
5	דנטה אליגיירי	17.82	55	איליה קפיטולינה	16.82
6	הומאז' לקטלוניה	17.82	56	בוסניה והרצגובינה	16.82
7	הסקס פיסטולס	17.82	57	ג'נרל מוטורס	16.82
8	חריצים וחרירים	17.82	58	האוסטרו הונגרית	16.82
9	ינקול גולדווסר	17.82	59	לוני טונס	16.82
10	סואמי פרבהופאדה	17.82	60	לורן צ'יילד	16.82
11	פרל הארבור	17.82	61	מסרשמיט bf	16.82
12	רנדי אורטון	17.82	62	נה וין	16.82
13	אוביטר דיקטום	17.5	63	רואלד דאל	16.82
14	אמל נסראלדין	17.5	64	תגלת פלאסר	16.82
15	הגשש החיוור	17.5	65	פלורסהיימר למחקרי	16.75
16	והאסיר מאזקבאן	17.5	66	באגודה השיתופית	16.69
17	ושלמי הסנדלר	17.5	67	התפוקה השולית	16.69
18	זריקות ותפיסות	17.5	68	נייג'ל מנסל	16.69
19	ציידת הערפדים	17.5	69	midway arcade	16.65
20	תוקד בקרבי	17.5	70	new york	16.65
21	external oss	17.24	71	החומצה הלקטית	16.65
22	greatest hits	17.24	72	ז'ק קרטייה	16.65
23	איזדורה דנקן	17.24	73	טריניד וטובגו	16.65
24	ארקדי דוכין	17.24	74	קולד מאונטן	16.65
25	באפי ציידת	17.24	75	שרלוק הולמס	16.65
26	דוויל סטיק	17.24	76	happy nation	16.56
27	הדגה והבר	17.24	77	בלטריקס לסטריינג'	16.56
28	החתיכה החסרה	17.24	78	מפעם לפעם	16.53
29	המה מליני	17.24	79	da capo	16.5
30	ויפסניוס אגריפה	17.24	80	בוולנט גרוב	16.5
31	וישו העולל	17.24	81	בקואורדינטות קרטזיות	16.5
32	כלמידיה טריכומטיס	17.24	82	ברויאל ראמבל	16.5
33	ליאונרדו דא	17.24	83	דוידסון אונליין	16.5
34	מדהורי דיקסיט	17.24	84	הצהרת בלפור	16.5
35	קרוא וכתוב	17.24	85	הרויאל ראמבל	16.5
36	תזת צ'רץ'-טיורינג	17.24	86	יוסטס סקרב	16.5
37	תיאוריה וביקורת	17.24	87	רומיאו ויוליה	16.5
38	הנרייטה סאלד	17.18	88	רשיד כראמי	16.5
39	ניקולאה צ'אושסקו	17.18	89	שמרי צמרת	16.5
40	dolby digital	17.01	90	בסקי האלפיני	16.43
41	still missing	17.01	91	היונקים הימיים	16.43
42	that she	17.01	92	יאנה אהונן	16.43
43	האשלגן החנקתי	17.01	93	העסק הביש	16.39
44	חבאר גידר	17.01	94	הפחם והפלדה	16.39
45	מאסיב אטאק	17.01	95	אמברוזיאנה במילנו	16.36
46	מבוכים ודרקונים	17.01	96	האשה שאיתי	16.36
47	סרגון מאכד	17.01	97	ערש דווי	16.34
48	deadly alliance	16.97	98	harry potter	16.33
49	גזורים לאונות	16.97	99	בהטלת כידון	16.33
50	גלנדון ואיזבלה	16.97	100	המחירים לצרכן	16.33

PMI_select.txt					
Rank	Collocation	Score	Rank	Collocation	Score
1	mortal kombat	15.36	36	לשפר את	6.05
2	תגליות והמצאות	15.18	37	כמעט כל	5.97
3	שאו קאהן	15.05	38	את מקומו	5.7
4	ברי סחרוף	14.88	39	היה בנו	5.21
5	אישים משמעותיים	13.93	40	בסרט זה	5.12
6	נשיונל ג'יאוגרפיק	13.78	41	קרוי על	5.11
7	עין חרוד	12.28	42	ועל ידי	4.96
8	חדשות אישים	12.09	43	אלו הם	4.87
9	אתונה העתיקה	12.06	44	הוא דמות	4.87
10	הר חברון	11.61	45	דבר זה	4.68
11	בברית המועצות	11.39	46	אף הם	4.47
12	גרמניה הנאצית	11.36	47	הוא קיבל	4.19
13	מיוצגת האות	11.27	48	את עצמם	4.08
14	ערים חדשות	11.01	49	בהיסטוריה של	4.02
15	במזרח התיכון	10.87	50	בחייו של	3.99
16	שנות השישים	10.03	51	של אוניברסיטת	3.75
17	עלה לארץ	9.83	52	זה ניתן	3.65
18	בתי חולים	9.74	53	ספריו של	3.59
19	פני כדור	9.23	54	אף היא	3.56
20	זכתה באליפות	9.22	55	את המילה	3.51
21	בבית החולים	8.53	56	ניתן היה	3.46
22	במקרים רבים	8.49	57	של הכנסייה	3.04
23	מדי יום	8.42	58	מכן הוא	2.88
24	בסוף שנת	8.26	59	של הסדרה	2.87
25	אין מידע	7.82	60	של הליכוד	2.86
26	הספר התיכון	7.73	61	כן הוא	2.61
27	אך ורק	7.44	62	של הצבא	2.55
28	מצד שני	7.33	63	אך היא	2.43
29	מיד לאחר	7.04	64	בקרב על	2.4
30	מצד אחד	6.94	65	של כלי	2.33
31	קבע כי	6.75	66	הוא שם	1.9
32	לא הצליחו	6.69	67	אך עם	1.8
33	לא ברור	6.63	68	אותו על	1.04
34	המבוססת על	6.31	69	הוא על	-1.4
35	בשל כך	6.07			