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. PMI מציפה 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 המשפט | 1. 97 2. 91 3. לא במאה הראשונים | 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 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  נשיונל גיאוגרפיק | 45  70  128 | 1  4  6 | 1  4  6 | דוגמא לזוגות מילים נפוצות אשר דירוגם ב-T Score וב-PMI גבוהה יותר מאשר דרוג התדירות בשל כניסת כל הביטויים הלא מענייניים הכללו סימני פיסוק שתפסו את המקומות הראשונים בדירוג |

# האם רוב הזוגות בראש הרשימה משמעותיים\מענייניים?

כלומר, האם אלו צירופי מילים אשר באמת נוטים להיקרות יחד באופן שכיח בשפה? במילים אחרות, אם מישהו שלומד עברית היה מסתכל על צמדים אלו, האם הדבר היה עוזר לו ללמוד באילו צמדי מילים כדאי להשתמש? לדוגמא, צמד המילים "בכל זאת" כן משמעותי, בעוד שהצמד ")." אינו כזה.

תנו כמה דוגמאות מכל סוג (משמעותיים, לא משמעותיים) לזוגות מילים מתוך התוצאות שקיבלתם.

אם רוב הזוגות שבאיזור רוב הרשימה לא נראים לכם משמעותיים או מעניינים, מה הייתם משנים על מנת לקבל תוצאות משמעותיות יותר?

 כעת, יישמו לפחות 2 מהרעיונות שהצעתם על מנת לקבל תוצאות משמעותיות יותר, ודווחו האם התקבל השיפור שקיוויתם לו. דווחו איזו הצעה עזרה לאיזה מדד (במידה והשיפור התקבל במדד מסוים אך לא באחר).

לבסוף, תנו 2 דוגמאות לזוגות מילים אשר דירוגן עלה כמצופה לאחר יישום השינויים. שימו לב - **אל תגישו את הקוד שתכתבו בחלק זה של התרגיל (יישום הצעות לשיפור). הקוד שעליכם להגיש הוא הקוד ללא ההצעות לשיפור, וכך גם יהיה הפלט שלו**.