תרגיל בית 2 - קולוקציות

חיים שאללשוילי 200832780

צחי כפיר 200681476

# הבדלים בין התוצאות עבור מדדים שונים

אם כן, ממה נובעים ההבדלים? קשרו זאת למאפיינים (יתרונות או חסרונות) של המדדים שתוארו בשיעור.

תנו 3 דוגמאות לזוגות מילים אשר דירוגם שונה באופן משמעותי בין שני מדדים, ופרטו מה המדדים, ומה ערכיהם עבור הזוג (אם הזוג לא הופיע כלל ב-100 הזוגות השכיחים ביותר עבור אחד המדדים, ציינו זאת). הקפידו ששלוש הדוגמאות ידגימו לפחות שני הבדלים שונים (כלומר, אל תתנו 3 דוגמאות ששלושתן מקרים שהופיעו במדד A אך לא במדד B).

מדד RawFrequency ממיין את הקולוקציות על פי נפוצותם בטקסט. כצפוי מדד זה הציג תוצאות רבות הכללו סימני פיסוק ונתן להם ציון גבוה שכן תדירותם הייתה גבוהה.

מדד TScore מעניק ציון גבוה יותר לזוג מילים ככל שהמילים פחות נפוצות בקורפוס. לכן הציג שיפור בנושא זה והביטויים שהכילו סימני פיסוק ירדו בדירוגם.

מדד PMI אין לי מושג ... את זה צחי עשה – אפשר לדבר על כל הביטויים באנגלית ועל מספרים רבים שעלו בדירוג (בינהם שמות)

|  |  |  |  |
| --- | --- | --- | --- |
| קולוקציה | מיקום Raw | מיקום Tscore | מיקום PMI |
| על , | 19 | לא במאה הראשונים | לא במאה הראשונים |
| על שם | 97 | 68 | לא במאה הראשונים |
| לקריאה נוספת | 91 | 54 | לא במאה הראשונים |
| ככל הנראה | לא במאה הראשונים | 99 | לא במאה הראשונים |
| בית המשפט | לא במאה הראשונים | 65 | לא במאה הראשונים |
| בית הספר | לא במאה הראשונים | 69 | לא במאה הראשונים |
| ,1977 ,1980 | לא במאה הראשונים | לא במאה הראשונים | 1 |
| Alain Delon | לא במאה הראשונים | לא במאה הראשונים | 67 |

# 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**.