<u>שיעורי בית 2</u>

בעיבוד שפות טבעיות

מגישים:

אילן גודיק – 316315332

צ'רלי מובאריכי – 316278118

שאלה מספר 1

השוואת Raw Frequency כנגד Unfiltered Bigram PMI - ללא תנאי של לפחות 20 הופעות

שני מדדים אלו נותנים תוצאות הפוכות לגמרי:

- **Raw Frequency** נותן את זוגות המילים השכיחות ביותר
- **Unfiltered Bigram PMI** נותן את כל זוגות המילים שהופיעו פעם אחת בלבד, כך שכל מילה בצירוף הופיעה גם היא רק פעם אחת בקורפוס.

לכן, אין צירופים משותפים בין שתי התוצאות כלל.

- ב**Raw Frequency** יופיעו בראש התוצאות מילות חיבור ומילים מבניות אחרות בשפה, מכיוון שאלה הן המילים המשומשות ביותר בשפה.
  זהו חיסרון, כי אלו לא קולוקציות של השפה. לכן התבססות על כמות השימוש בצמד המילים לא מספיק.

  יתרון: לאחר מילות המבנה של השפה בראש הרשימה, נראה צירופים שמופיעים רבות בשפה, וייתכן שאלו הם אכן קולוקציות.

- ב**Unfiltered Bigram PMI** יצופו בראש התוצאות מילים המופעיות פעם אחת בלבד או כמות מועטה מאוד של פעמים, ובין תוצאות אלה שגיאות כתיב למיניהן, מכיוון שמופיעות לעיתים רחוקות גם בקורפוסים איכותיים.
  יתרון: מתחשב בתלות בין המילים, אך זה לא משמעותי כל עוד רוב התוצאות הן רעש.

דוגמאות לצירופים שאין בתוצאה האחרת (כל הצירופים מתאימים פה):

- ב**Unfiltered Bigram PMI**: "1852 והפסיכואנליטיקאי" - 18.050
- ב**Raw Frequency**: "את זה" - 0.635

שאלה מספר 2

הבעייתיות ב**Unfiltered Bigram PMI**

אם המשתנים בלתי תלויים:

$$\frac{P(xy)}{P(x) \cdot P(y)} = 1$$

וזהו PMI נמוך, כצפוי.

ואם הם תלויים לגמרי, כלומר תמיד באים ביחד:

$$P(xy) = P(x) = P(y)$$

$$PMI = \frac{P(xy)}{P(x) \cdot P(y)} = \frac{1}{P(x)}$$

כלומר, ככל ש $P(x)$ נמוך יותר, כך הPMI יהיה גבוה יותר, וזו התנהגות לא רצויה:

אם יש לנו כמה ביטויים שתמיד באים ביחד, ככל שמספר ההופעות של הביטוי גבוה יותר, כך יותר סביר שהוא יהיה Collocation, ולא זוג מילים נדירות שהופיעו רק פעם אחת ביחד בכל הקורפוס.

לכן מתקבלים צירופים אשר <u>לאו דווקא נוטים להיקרות יחד באופן שכיח בשפה</u>. נקבל תוצאה הפוכה: צירופים כבולים, אך הנדירים ביותר.

דוגמא: בdevset של CHILDES,

- תביאי לי - PMI = 10118.906
- בסכין הגילוח – PMI = 9118.906

כאשר שניהם צירופים כבולים לגמרי – כל מילה מופיעה רק עם השנייה, אך הצירוף "תביאי לי" מופיע רק פעם אחת בקורפוס, לעומת "בסכין הגילוח" שמופיע פעמיים בקורפוס.

לא מתקבלים במדד זה תוצאות משמעותיות – קולוקציות תקינות, מכיוון שכל הצירופים מתוך ה100 העליונים היו צירופים שהופיעו פעם אחת בלבד, ולרוב אלו לא קולוקציות, אלא שמות, שגיאות כתיב, או צירופים מקריים.

לדוגמא, "הדטרמיניסטי שמבוצע" הוא צמד מילים מקרי, שייתכן שהיה צמד כבול שמופיע פעם אחת בלבד ומופיע בראש מדד זה, לעומת "בכל זאת", שהוא צמד מילים פחות כבול, ושמופיע פעמים רבות, ולכן לא יופיע בראש מדד זה, למרות שהוא כן קולוקציה טובה.

שאלות מספר 3 ו- 4

השפעת דרישת מספר ההופעות המינימלי על מדדי הPMI

לפני הגבלה זו, כל התוצאות היו רק רעש של צירופים שהופיעו פעם אחת בלבד.

לאחר הגבלה זו, אנו מקבלים קולוקציות משמעותיות הרבה יותר, אך רק במדדים **Bigram PMI** ו**Trigram PMI A**

דוגמאות:

- עבור **Bigram PMI**:
  לפני: "1852 והפסיכואנליטיקאי"
  אחרי: "יוצא דופן"
- עבור **Trigram PMI A**:
  לפני: "10.6 המיליונים ששרדו"
  אחרי: "מצא חן בעיני"
- עבור **Trigram PMI B**:
  לפני: "! , הבייבי"
  אחרי: "" בקרב חברי"
  יוצא שעבור מדד זה, הגבלה של לפחות 20 הופעות לכל טוקן בצירוף לא עזרה – מספר הופעות רב של כל מילה לא מייצג כמה צירוף זה מופיע בשפה, ונותרנו עם רעש.
- עבור **Trigram PMI C**:
  לפני: "33.3 סנטימטר בשעה"
  אחרי: "צריכים להניח בצד"

בסופו של דבר, מדדי הBigram PMI וגם Trigram PMI A המתוקנים נתנו תוצאות טובות, וTrigram PMI B וגם Trigram PMI C שניהם נתנו תוצאות לא טובות, גם לאחר ההגבלה.

לאור הניתוח, ניסינו 2 התמודדויות עם הבעיה:

1. תנאי של לפחות k הופעות של <u>הצירוף</u> במקום כל מילה בנפרד.
   תוצאות: שיפור משמעותי.
   עבור **Bigram PMI**: "לעתים קרובות", "בגיל העמידה", "אחר הצהריים"
   עבור **Trigram PMI A**: "בסופו של דבר", "פחות או יותר"
   עבור **Trigram PMI B**: "מה זאת אומרת", "עד כדי כך"
   עבור **Trigram PMI C**: "אף על פי", "אף פעם לא"

2. מדד PMI אלטרנטיבי: $P(xy) \cdot \log_2 \frac{P(xy)}{P(x)P(y)}$ וכנ"ל מכפלה של $P(xyz)$ במדדים של Trigram PMI.
   בכך אנחנו מכניסים כבר בתוך המדד כתנאי לקולוקציה טובה – מספר הופעות רב.
   עבור **Bigram PMI**: "לעתים קרובות", "גיל העמידה", "ארצות הברית"
   עבור **Trigram PMI A**: "תורת היחסות הכללית", "אף על פי"
   עבור **Trigram PMI B**: "בסופו של דבר", "אף על פי"
   עבור **Trigram PMI C**: "עד כדי כך", "בסופו של דבר"
   מדד זה שיפר את התוצאות של **Bigram PMI** אך המכפלה בהסתברות ששלישיה מופיעה כבדה מדי עבור **Trigrams**, ולכן קיבלנו יותר מילות מבנה וביטויי זמן בשפה.

בסך הכל, קיבלנו באופן חוזר שהמדדים של **Bigram PMI** עבור זוגות, ו**Trigram PMI A** שלישיות הם המדדים הטובים ביותר מתוך כל אוסף המדדים הנתון.

השערה לסיבה לכך: אי תלות ביחידים עבור שלשה פועל יותר טוב מאשר שימוש באי תלות בזוגות עבור הקורפוס הנ"ל, לדוגמא "תורת היחסות הכללית" – יש לה אי תלות מסוימת בזוג: לדוגמא "תורת היחסות" יכולה להופיע גם לבד, אך זה לא פוסל שזוהי גם קולוקציה טובה.

דוגמא נוספת: "אף על פי" – הצירוף "על פי" יכול להופיע גם לבד, אך זה לא פוסל את כך ש "אף על פי" היא קולוקציה טובה.