עיבוד שפות טבעיות - תרגיל בית שני

פרטי הגשה

מגישים

אורי אריאל ניפומניאשצ'י	318261468
ישי גרוניך	208989186

מערכת הפעלה

Windows 7

הקדמה

פירסור של הקורפוס המחולק לטוקנים

הקורפוס נתון בקובץ, לאחר שחולק לטוקנים, כאשר המשפטים מופרדים באמצעות התו newline, והטוקנים מופרדים באמצעות רווח. במהלך ריצת התכנית, הייצוג של הקורפוס המחולק לטוקנים הוא על ידי רשימה של משפטים, כאשר כל משפט מיוצג על ידי רשימה של טוקנים (שכל אחת היא מחרוזת). לכן היה טבעי לפרסר את הקורפוס כך:

```
corpus_tok = corpus_str.split('\n')
for i in range(len(corpus_tok)):
   corpus_tok[i] = corpus_tok[i].split(' ')
```

ישנה בעייתיות בגישה זו מכיוון שישנם מקומות ברבים קורפוס עם חזרה כפולה על רווח: " " (ליתר דיוק, 743 חזרות), מה שגורם לטוקן הריק "" להיות נפוץ, ולהשפיע על תוצאות חיפוש הקולוקציות. לכן, הורדנו את כל הטוקנים הריקים על ידי שינוי השורה האחרונה להיות:

```
corpus_tok[i] = filter(lambda tok: tok != '', corpus_tok[i].split(' '))
```

צירופים כבולים עבור קורפוס ספציפי

דו"ח זה מנתח תוצאות של אלגוריתמים שונים, המופעלים על מנת לזהות קולוקציות וגם trigrams בתוך קורפוס. הקורפוס עליו הורצו האלגוריתמים הוא אוסף של פרקים מתוך 78 ספרים בעברית. מאחר שהקורפוס הוא יחסית קטן, הוא אינו מייצג את כלל השפה העברית - כלומר, ישנם צירופים שהאלגוריתמים זיהו, שהם צירופים כבולים עבור ספר מסוים (ועבור הקורפוס), אך אינם צירופים כבולים של השפה העברית (או האנגלית).

דוגמה לכך הוא הצירוף "האח הצעיר ממך" מתוך הספר "בבקשה תשגיחי על אמא" מאת קיונג סוק שין. הצירוף הוא אכן צירוף כבול בהקשר של הספר, שכן מדובר על דמות ספציפית, וזו הדרך שבה בחרה הסופרת לכנות את הדמות. צירוף כבול דומה הוא "האח הגדול" מתוך הספר 1984 (שאינו בקורפוס הנתון). זוג המילים "אח גדול" או "האח הגודל" אינם צירוף כבול באופן בכללי בשפה העברית, אבל בהקשרים מסוימים, ושימושים מסוימים, הוא אכן צירוף כבול.

שתי הדוגמאות הללו הן דוגמאות של צירופים כבולים בהקשר מסוים, ויש לכך מספר דוגמאות מתוך התוצאות על קורפוס הבדיקה. תופעה זו כנראה הייתה בולטת הרבה יותר אילו הקורפוס היה ספר יחיד באורך מלא, בניגוד לחלקים קטנים מתוך אוסף גדול של ספרים.

ניתוח תוצאות

שאלה מספר 1

האם ניכרים הבדלים בולטים בין התוצאות עבור מדד ה-PMI לזוגות טוקנים ומדד ה-raw frequency?
אם כן, ממה נובעים ההבדלים? קשרו זאת למאפיינים (יתרונות או חסרונות) של המדדים שתוארו בשיעור.
תנו 2 דוגמאות לזוגות טוקנים אשר דירוגם שונה באופן משמעותי בין שני מדדים אלו, ופרטו מה ערכי המדדים עבור זוג הטוקנים (אם הזוג לא הופיע כלל ב-100 הזוגות השכיחים ביותר עבור אחד המדדים, ציינו זאת).

ניכרים הבדלים גדולים בין הזוגות עם מדד ה-raw frequency הגבוה ביותר לבין הזוגות עם מדד ה-PMI הגבוה ביותר.

- מדד Raw Frequency מוצא זוגות טוקנים שמופיעים הרבה פעמים בטקסט (כזוג). לזוג טוקנים המהווים קולוקציה ישנו סיכוי גבוה להופיע יחד בטקסט, ולכן ייתכן שקולוקציות תופענה מספר רב של פעמים יחסית לזוגות אחרים, כלומר תתבלטנה מבחינת Raw Frequency. אבל:

1. קולוקציות לאו דווקא מופיעות בטקסט מספר רב של פעמים. ישנן קולוקציות שאינן שכיחות בסוגים מסויימים של קורפוסים. למשל, הקורפוס הנתון test הוא אוסף של ספרים בעברית, ולכן קולוקציות באנגלית אינן נפוצות בו. כלומר קולוקציה כמו Intelligence Quotient תקבל ציון מאוד נמוך של Raw Frequency.
2. זוג עם Raw Frequency גבוה הוא לאו דווקא קולוקציה, שכן ישנם זוגות טוקנים שיופיעו ביחד מספר רב של פעמים כתוצאה מכך שהטוקנים עצמם הם מאוד שכיחים בשפה, ולכן "במקרה" יהיו להם הרבה מופעים ברצף. דוגמאות לכך הן זוגות הטוקנים (. ") (, ") (? ") שקיבלנו את ציוני Raw Frequency הגבוים ביותר אך אינן קולוקציות.

- מדד PMI מוצא זוגות טוקנים שמופיעים הרבה ביחד יחסית למספר הפעמים שהם מופיעים לחוד.

מדד מסוג זה הוא יותר אופייני לקולוקציות, כי ברוב המקרים הטוקנים שמרכיבים את הקולוקציה הרבה פחות נפוצים בפני עצמם, מאשר כזוג כבול. כלומר קולוקציה כמו Intelligence Quotient תקבל ציון גבוה מאחר שהטוקן Quotient כלל אינן נפוץ, אלא בהקשר של הקולוציה IQ (בקורפוס הספציפי הזה, שהוא בעברית, גם המילה Intelligence מופיעה רק פעם אחת - רק בהקשר של IQ). כלומר בדרך כלל מדד PMI אכן מאפיין קולוקציות. אבל:

1. לא כל קולוקציה תקבל ציון PMI גבוה. בעיה זו יכולה לקרות במקרים שבהם שתי המילים המרכיבות את הקולוקציה הן נפוצות בפני עצמן, בלי קשר להיותן נפוצות ביחד כזוג כבול. דוגמה לזוג כזה הוא "כל כך". לזוג כזה יש מספר גבוה של מופעים יחד, אבל לכל טוקן יש גם מספר גבוה של מופעים לחוד, ולכן מדד PMI אינו גבוה.
2. לא כל זוג עם ציון PMI גבוה הוא בהכרח קולוקציה. בעיה זו יכול לקרות במקרים שבהם ישנן טוקן שהוא אינו נפוץ, למשל מופיע רק פעם אחת או פעמיים בקורפוס, ולכן מופיע מספר גבוה של פעמים עם הטוקן שלפניו/אחריו, ביחס למספר המופעים הכולל שלו. דוגמאות לבעיות מסוג זה הן typos בטקסט, או מספרים - שכן בדרך כלל מספר ספציפי יופיע רק פעם אחת.

כמעט כל זוגות המילים, מתוך 16 הזוגות עם מדד PMI הגבוה ביותר בקורפוס test, מכילות איזהשהו מספר (או שניים), כלומר אינן באמת קולוקציות. דוגמה לכך: "12.8.1889 ל־11.11.1889".

<u>סיכום דוגמאות שאלה מספר 1</u>

הסבר לטבלה:

- בחלק מהדוגמאות, Raw Frequency הוא מדד טוב, כלומר מצליח לזהות היטב קולוקציה (לתת לה ערך גבוה) ולזהות היטב זוג שאינו קולוקציה (לתת לו ערך נמוך).
בחלק מהדוגמאות, PMI הוא מדד טוב, כלומר מצליח לזהות היטב קולוקציה (לתת לה ערך גבוה) ולזהות היטב זוג שאינו קולוקציה (לתת לו ערך נמוך).
- הדירוג הוא המיקום של הקולוקציה לפי סדר המיון (לפי ערך ואז לפי א"ב), כאשר הקולוקציה שקיבלה את הערך הגבוה ביותר היא בדירוג 1. כלומר, בשורה הראשונה של הטבלה להלן, דירוג קרוב ל-1 הוא דירוג מוצלח, ואילו בשורה השנייה של הטבלה, דירוג קרוב ל-1 הוא דירוג לא מוצלח.
- דירוג מוצלח מסומן בירוק ואילו דירוג לא מוצלח מסומן באדום.
- הערכים של Raw Frequency מוכפלים פי 1000 לצורך ויזואליזציה טובה יותר.

	מדד Raw Freq מזהה היטב	מדד PMI מזהה היטב
קולוקציה	**כל כך** ערך Raw Freq הוא: 0.595 דירוג Raw Freq הוא: **16** ערך PMI הוא: 6.23 דירוג Raw Freq הוא: 106958	**Intelligence Quotient** ערך PMI הוא: 18.05 דירוג PMI הוא: **40** ערך Raw Freq הוא: 0.004 דירוג PMI הוא: 28351
לא קולוקציה	**12.8.1889 ל־11.11.1889** ערך Raw Freq הוא: 0.004 דירוג Raw Freq הוא: **27589** ערך PMI הוא: 18.049 דירוג Raw Freq הוא: 5	**, אבל** ערך PMI הוא: 3.227 דירוג PMI הוא: **143489** ערך Raw Freq הוא: 2.2 דירוג Raw Freq הוא: 4

שאלה מספר 2

מה הבעייתיות העולה מהתבוננות בתוצאות מדד ה-PMI עבור זוגות טוקנים? רמז: האם אלו צירופי טוקנים 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 "הדטרמיניסטי שמבוצע" אינו כזה.
תנו כמה דוגמאות מכל סוג (משמעותיים, לא משמעותיים) לזוגות טוקנים מתוך התוצאות שקיבלתם עבור מדד ה-PMI עבור זוגות טוקנים. אם אינכם מוצאים דוגמאות מאחד הסוגים, הסבירו מדוע זה כך.

מדד PMI אינו בהכרח נותן צירופי טוקנים אשר נוטים להקרות יחד באופן שכיח בשפה. הוא נותן צירופי טוקנים אשר מופיעים יחד בטקסט, בשכיחות משמעותית יחסית למופעים שלהם לחוד. כלומר הטוקנים הללו לאו דווקא מופיעים מספר פעמים רבות יחד בטקסט. לתכונה זאת יש יתרון, ואכן מדד PMI מוצא גם קולוקציות שמספר המופעים שלהן הוא קטן. **אבל** החסרון של תכונה זאת, היא שמדד PMI נותן ערך גבוה גם לזוגות שאינן קולוקציות, אלא הופיעו בטקסט יחד במקרה. במקרים שבהם הטוקנים אינם טוקנים נפוצים בשפה בפני עצמם, ייתכן כי המופעים היחידים (או מופע יחיד) שלהם בקורפוס הם יחד, ולכן ייקבלו ערך PMI גבוה כזוג. דוגמה לכך היא טוקנים המכילים מספרים - לא סביר שאותו המספר יחזור על עצמו מספר רב של פעמים בטקסט. לכן טוקן המכיל מספר עלול "להצמד" לטוקן לפניו/אחריו וליצור זוג עם ערך PMI גבוה.
עבור מישהו שמנסה ללמוד שפה, <u>רוב</u> הדוגמאות <u>אינן</u> משקפות צמדים שכדאי להשתמש בהם, או צמדים שנפוצים יחד, מאחר שחלק גדול מהדוגמאות עם ערכים גבוהים הן דוגמאות שהופיעו יחד במקרה, כפי שתואר לעיל.

<u>דוגמאות לקולוקציות בעלות ערך PMI גבוה (ערך 18.05)</u>
ברירת המחדל , דונלד טראמפ, Wall Street , middle age, well being, Scientific American, בערפול חושים, בסט ביי, בפוקס ניוז, גילי פרישה, במאת האחוזים, בעמדת נחיתות

<u>דוגמאות לזוגות שאינן קולוקציות ובעלות ערך PMI גבוה (ערך 18.05)</u>
בקטנות ולהתעקש, בקדנצה ששמורה, בעלך ובאתי, אמריקניות מהמעמד, אנגליקני וגדלתי, ההתפעמות שאוחז, ההשמדה ובניהם, ההפסדים שיספגו, הזורמים בקצות, החוקי ותעלה

<u>הערה לגבי הדוגמאות</u>
כל הדוגמאות הנ"ל בעלות אותו הערך, שהוא 18.05. חלק מהדוגמאות הנ"ל אינן מופיעות כחלק מתוך 100 הקולוקציות הראשונות לפי סדר המיון, אבל הסיבה לכך היא הסדר האלפבתי ולכן אין לכך משמעות. הסיבה שיש מספר כה גדול של דוגמאות בעלות אותו ערך, היא שבכל אחת מהדוגמאות הנ"ל, מספר המופעים בטקסט של כל טוקן הוא 1, וגם מספר המופעים של הזוג בטקסט הוא 1, ויש מספר רב של זוגות בטקסט המקיימות תכונה זו.

שאלה מספר 3

כדי להתגבר על הבעייתיות מסעיף (2) לעיל, הגבילו כעת את השכיחויות של הטוקנים (כאשר הללו מופיעים בנפרד). כלומר, הפיקו מחדש את רשימות 100 הזוגות ו-100השלשות הכי שכיחים של מדד ה-PMI (סך הכל 4 רשימות), והפעם תיכללו רק זוגות ושלשות טוקנים אשר השכיחות של הטוקנים היחידים המרכיבים אותם היא לפחות 20. כלומר, ערכי C(x), C(y), C(z) צריכים להיות לפחות 20, כאשר C הוא מספר המופעים של הטוקן. האם כעת צירופי הטוקנים עבור ארבעת מדדים אלו (PMI עבור זוגות ושלושה מדדי ה-PMI עבור שלשות) משמעותיים יותר? תמכו בתשובתכם באמצעות דוגמאות של זוגות ו/או שלשות רלבנטיים. הקפידו לציין עבור כל דוגמא, מאיזה רשימה היא נלקחה (כלומר, של איזה מדד היא).

- לאחר שינוי זה, הטוקנים אכן הרבה יותר משמעותיים. הסיבה היא שהבעייתיות המתוארת בשאלה מספר 2, נגרמת בעיקר כתוצאה מטוקנים אשר אינם שכיחים בקורפוס באופן כללי, ולכן מופיעים הרבה יחד באופן יחסי. בעיה זו נמנעת כאשר מתייחסים רק לטוקנים שמספר המופעים שלהם הוא לפחות 20.
- ניתן לראות שינוי משמעותי ביותר בערכי PMI עבור זוגות טוקנים. כאשר הסף הוא 20 מופעים לכל טוקן, כמעט כל הזוגות בעלות ציון גבוה הן אכן קולוקציות.
  - דוגמאות: בארצות הברית (מקום 7), יוצא דופן (מקום 9), בניו יורק (מקום 12), תשומת הלב (מקום 13), גיל העמידה (מקום 15), ולאחר מכן (מקום 17), כף היד (מקום 18), כדור הארץ (מקום 21). בכל הדוגמאות הנ"ל, ערך PMI הוא מעל 11.
  - הדוגמאות יוצאות הדופן (כלומר זוגות שקיבלו ערך PMI גבוה אך אינן קולוקציות) הן כתוצאה מבעיות בטוקניזציה (לדוגמה: פירוק השם שוג'טה על ידי התו '), או זוגות שהם אכן קולוקציה אבל ספציפית לאחד הטקסטים, כלומר אינן קולוקציות של השפה העברית (לדוגמה: "הנזירה ברנדט" וגם "השוטר השחור" מהספר "מקום יפה למות בו" מאת מאלה נאן).
  - החסרון של הגבלת סף גבוה, כמו 20, הוא פיספוס של קולוקציות רבות שמופיעות מעט בקורפוס. למשל, כאשר משנים את הסף להיות 5, במקום 20, הסף מצליח כמעט להתגבר על הבעיה המתוארת בשאלה מספר 2, ומקבלים קולוקציות עם ציונים גבוהים יותר: פשוטו כמשמעותו (מקום 1), רבי החובלים (מקום 2), מפעם לפעם (מקום 3), שיווי משקל (מקום 5), חומרי הגלם (מקום 13), עשבים שוטים (מקום 17), ליתר דיוק (מקום 23). כל הדוגמאות הנ"ל בעלות ערך PMI גדול מאשר 14.5.
  ככל שמגדילים את הסף, יש ביטחון יותר גדול שכל אחד מהזוגות הוא אכן קולוקציה, אבל יותר קולוקציות מתפספסות (כלומר, precision גדל אבל recall קטן).
- גם עבור מדד PMI_a, הגבלת המינימום של 20 מופעים משפיעה משמעותית על התוצאות.
  - ללא ההגבלה, כל 100 התוצאות הראשונות הן שלשות של טוקנים שמופיעות רק פעם אחת בקורפוס. רובם שלשות של מילים באנגלית, או טוקנים המכילים מספרים, או סימנים מיוחדים. לרובן אין כל משמעות בתור שלשת מילים, לדוגמה: "1709 הוסמך לכמורה", "have an invisible". שלשות מעטות הן באמת trigrams משמעותיים, כמו Positive Intelligence Quotient וגם Wall Street Journal.
  - אחרי ההגבלה, מופיעים trigrams הרבה יותר אופייניים (חלקם אופייניים לקורפוס, וחלקם אופייניים לשפה העברית באופן כללי). לדוגמה: מצאתי חן בעיניו (מקום 55), וכמה וכמה פעמים (מקום 78).
  - המדד PMI_b וגם PMI_c הם יותר בעייתיים, כפי שיתואר בשאלה מספר 4. ההגבלה של סף 20 לא גרמה לשיפור משמעותי בתוצאות - התוצאות הגבוהות ביותר עדיין היו שלשות XYZ המורכבות מזוגות XY YZ אשר מופיעים רק פעם בטקסט. כך, למשל, השלשה "(מישהו חושב" קיבלה את הערך המקסימלי 17.94 בשיטת PMI_b, גם לפני הגבלת הסף וגם אחרי הגבלת הסף.

שאלה מספר 4

כעת, השוו את התוצאות שקיבלתם בשאלה מספר (3) עבור שלשות טוקנים - איזה מדד מניב תוצאות טובות יותר, מבין שלושת צורות החישוב השונות של חישוב PMI לשלשות טוקנים? נמקו את תשובתכם, ותמכו בה באמצעות דוגמאות רלבנטיות (גם כאן, הקפידו לציין מאיזה רשימה נלקחה כל דוגמא). שימו לב כי לא בהכרח יש תשובה אחת נכונה, ולכן תשובתכם תוערך לפי הנימוקים והדוגמאות שלכם.

- מדד PMI_a הניב את התוצאות המשמעותיות ביותר. ישנו דימיון רב בינו לבין PMI רגיל (עבור זוגות) מאחר שבמובן מסוים הוא מודד עד כמה שלשת המילים היא איננה בלתי-תלויה. כלומר - ניתן להראות באמצעות חישוב דומה לזה של PMI לזוגות, ששלשות של טוקנים בלתי-תלויים תקבלנה ערך קרוב ל-0:
$log(\frac{P(xyz)}{P(x)P(y)P(z)}) \approx log(\frac{P(xy)P(z)}{P(x)P(y)P(z)}) \approx log(\frac{P(x)P(y)P(z)}{P(x)P(y)P(z)}) = 0$
<u>דוגמאות</u> לצירופים כבולים שזוהו על ידי המדד PMI_a: תורת האינטיליגנציה החיובית (מקום 3), האח הצעיר ממך (מקום 4, צירוף כבול של הטקסט "בבקשה תשגיחי על אמא" מאת קיונג סוק שין), מצא חן בעיני (מקום 7), מצאתי חן בעיניו (מקום 55), וכמה וכמה פעמים (מקום 78).
- מדד PMI_b הוא בעייתי. שלשות רבות קיבלו את אותו הערך הגבוה ביותר, שהוא 17.94. המשותף לכל השלשות XYZ הללו, הוא שישנו מופע יחיד של השלשה בקורפוס, ישנו מופע יחיד של הזוג הראשון XY בקורפוס, וישנו מופע יחיד של הזוג השני YZ בקורפוס. בעייה זו דומה לבעיה המתוארת בשאלה מספר 2 עבור PMI לזוגות <u>ללא</u> סף 20 - מאחר שכל רכיב (XY, YZ) הוא נדיר בפני עצמו, האיחוד XYZ ייקבל ערך PMI גבוה גם אם יהיו לו מספר מופעים קטן (בפרט, 1). בעיה זו חריפה במיוחד במדד PMI_b ואינה נפתרת בקלות על ידי הגבלת סף 20 לשכיחות <u>הטוקנים</u> המרכיבים את השלשה, מאחר שגם כאשר כל טוקן X, Y מופיע יותר מ-20 פעמים בטקסט, לא בהכרח <u>ההרכבה</u> XY מופיעה פעמים רבות בטקסט. בפרט, ישנם זוגות רבים של טוקנים אשר מופיעים רק פעם אחת בטקסט, למרות שכל אחד מהטוקנים הוא שכיח בפני עצמו. דוגמאות לבעייתיות של PMI_b אפילו לאחר הגבלה של סף 20: (" אומרים שאת) - מקום 1, (" אותו עצם) - מקום 2. לסיכום, מדד PMI_b עם סף 20 סובל מאותה הבעיה של PMI רגיל (עבור זוגות) ללא סף, מאחר שהסף הוא על שכיחות unigrams ולא שכיחות bigrams, והמדד PMI_b אינן מתייחס לשכיחות של הטוקנים unigrams המרכיבים אותו.
<u>דוגמאות</u> לשלשות, שאינן צירופים כבולים, וקיבלו ערך גבוה במדד PMI_b: (" מרגיש "), (" נושא די"), (" ניסיתי …), (" פה הסיפור).
- מדד PMI_c הוא איזהשהו שילוב של PMI_a עם PMI_b. הוא מושפע מאוד מהבעיות הנ"ל של PMI_b . השלשות XYZ עם ערכי PMI_c הגבוהים ביותר גם הם (כמו במדד PMI_b) בעלי מכנה משותף, שהזוג XY חל פעם אחת בלבד בקורפוס וכנ"ל לגבי YZ. בעוד שזהו מרכיב עיקרי בערכי PMI_c, ישנה גם ההשפעה של האיברים $P(x)P(y)P(z)$ במכנה, ולכן השלשות של PMI_c יותר משמעותיות מאשר השלשות של PMI_b, ובהרבה מהם ישנו זוג המהווה קולוקציה.
<u>דוגמאות</u> לשלשות משמעותיות או בעלות זוג משמעותי, שקיבלו ערך PMI_c גבוה: "וכמה וכמה פעמים" (מקום 34), "והיה שבע רצון" (מקום 37), "לעמוד שעה ארוכה" (מקום 39), "קול קורא לאנשים" (מקום 45).
- לסיכום: המדדים לפי סדר יורד של איכות התוצאות: PMI_a ואז PMI_c ואז PMI_b.