בס"ד

עיבוד שפות טבעיות – דו"ח קורפוסים

מגישים:

פורטנוי ליאור

אפראימוב אורן

מערכת הפעלה:

נבדק על כמה מערכות הפעלה: windows 7, windows 10 ו-Ubuntu 15.4.
התוכנית רצה בזמן של 17-19 שניות בממוצע.

שאלה 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.

תשובה:

ניכרים הבדלים מהותיים בין שתי המדדים וזאת בגלל דרך החישוב של כל ממד, ששם דגש על משהו אחר.

Raw frequency

במדד זה, הדגש הוא על מספר המופעים שזוגות טוקנים מגיעות יחד כזוג. כלומר, מדד זה מבוסס על ההנחה, שלקולוקציות יש סיכוי גבוה להופיע יחד על גבי הטקסט, ולכן במידע וביטוי X הינו קולקציה יש סיכוי שתופיע מספר רבים פעמים בטקסט ולכן תופיע במדד הנ"ל.

PMI-pair

במדד זה, הדגש הוא על מספר המופעים שזוגות טוקנים מגיעות יחד יחסית למספר הפעמים שכל טוקן מופיע בנפרד. מדד זה, הרבה יותר טוב ממד הקודם, בגלל שברוב המקרים קולוקציה שמורכבת מטוקנים נפוצה יותר (ברוב המקרים) מאשר הטוקנים אשר מרכיבים אותה!

חסרונות

Raw frequency:

- קולקציות לא חייבות להופיע בטקסט מספר רב של פעמים, יש המון קולוקציות שאינן נפוצות בשפה או בקורפוס הבדיקה שלנו.

בנוסף, בגלל שאנו עובדים בקורפוס שהינו בעברית, קולוקציות באנגלית יקבלו ציון נמוך.

- לא כל זוג טוקנים שקיבל ציון גבוהה הינו קולוקציה. זאת בגלל העבודה שיש זוגות של טוקנים, שכל טוקן שכיח בפני עצמו וכך ההסתברות שיהיו יחד (ללא קשר אם הם קולוקציה או לא) מאוד גדולה.
לדוגמא: קיבלנו שביטוי ". או הביטוי ,? או "? נפוץ בקורפוס שלנו (אך זה ברור שהם לא קולוקציה)

<u>PMI-pair:</u>

- בגלל דרך החישוב של מדד זה, לא כל קולוקציה תקבל ציון גבוהה. זה יכול לקרות, במקרים שבהם שתי המילים המרכיבות את הקולוקציה נפוצות בפני עצמן בקורפוס, ללא קשר שהן נפוצות בתור זוג כבול.
לדוגמא: אי-אפשר שקיבל ציון יחסית נמוך.
- בעיה יותר חמורה הינה שלא כל זוג שקיבל ציון גבוה במדד זה הוא קולוקציה. אנו רואים, שבמידה ויש זוג כבול, שכל טוקן בו מופיע פעם אחד בטקסט, הציון שהוא יקבל הוא מאוד מאוד גבוהה למרות שאינו קולוקציה. רוב הזוגות שקיבלו ערך גבוהה (שכל טוקן מופיע פעם אחד בדיוק) בנויות ממספרים (תאריכים או מספרים).

<u>שאלה 1 - המשך:</u>

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).

מדדים	. "	האינטליגנציה החיובית	
Raw frequency	מקום ראשון	מחוץ ל-100! דירוג 0.072	. " לא קולוקציה
PMI-pair	לא מופיע כלל	מקום ראשון דירוג 4.5	האינטליגנציה החיובית קולוקציה

שאלה 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.

פתרון:

כפי שציינו קודם, בגלל שדרך של מדד זה, הוא אינו בהכרח מדרג נכון צירופי טוקנים שבאמת נוטים לקרות יחד, באופן שכיח בשפה. מדד זה, שם דגש על מספר המופעים שזוג טוקנים מופיעים יחד ביחס למספר המופעים שלהם לחוד!
לכן, זוג טוקנים אשר מופיע פעם אחד בודד, וכול טוקן בו מופיע פעם אחד בודד יקבל ציון גבוה, ובחלק מן המקרים הוא כלל אינו קולוקציה.

עבור אדם אשר מנסה ללמוד את השפה העברית, הסתכלות על הצמדים שיצאו לא יעזרו לו, מפני שרוב הדוגמאות אינן משקפות צמדים אשר כדאי להשתמש בהם, מאחר שחלק גדול מהם הן זוגות שהופיעו יחד במקרה (בגלל שכל טוקן בודד הופיע בדיוק פעם אחד) ולכן קיבלו ציון גבוה.

דוגמאות:

קולוקציות - Wall Street, middle age, value proposition וכד'
לא קולוקציות - #48, &P, 10.6 המיליונים וכד'.

כל הדוגמאות הנ"ל נלקחו מ-100 הזוגות בעלי PMI הגבוה ביותר שערכם הינו 18.05!

שאלה 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-100 השלשות 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).

פתרון:

לאחר שינוי זה, הטוקנים הרבה יותר משמעותיים!
הסיבה לכך, היא מאוד פשוטה, על ידי כך שאנו רוצים רק את הטוקנים אשר מופיעים לפחות 20 פעמים, הבעייתיות שהוצגה קודם לכן נמנעת!

PMI-pair:

כאשר הסף הינו 20 מופעים עבור כל טוקן, רוב הזוגות אשר קיבלו ציון גבוהה (או ב-100 עם הציון הגבוהה ביותר) אכן קולוקציות.

- דוגמאות לקולקציות – בארצות הברית, גיל העמידה, דרום אפריקה, בשנים האחרונות, תשומת הלב וכד'.

חסרונות:

יחד עם זאת, עדין קיימות בעיות רבות אשר נגרמות מטוקניזציה לא נכונה, פירוק לא נכון של הטוקנים.
בנוסף, הגבלת סף אשר הינה 20, גורמת לפספוס של קולוקציות רבות שמופיעות יחסית מעט בקורפוס. לדוגמא: האינטליגנציה החיובית, בארצות הברית, בשנים האחרונות, תשומת הלב וכד'

לכן, אפשר היה למצוא סף אשר מביא איזון בין שני הדברים, כלומר שלא מפספס קולקציות נכונות אך יחד עם זאת לא מוסיף קולקציות לא נכונות.
לפי בדיקה שלנו רף סביבות 5 או 6 מביא איזון בין שתי הדברים.
לדוגמא: פשוטו כמשמעו, מפעם לפעם, שיווי המשקל, חומרי הגלם (שקיבלו ציון מאוד מאוד גבוהה (מעל 14.791 אך לא מופיעות כלל הגבלה של 20).

PMI a/b/c:

PMI_a:

כמו קודם, ללא הגבלה הנ"ל, רוב התוצאות הראשונות הן שלשות של טוקנים שמופיעות רק מופע יחיד בקורפוס. רוב שלשות הטוקנים (פרט למילים עם

הניקוד אליהם לא נתייחס) הם שלשות טוקנים באנגלית או שלשות טוקנים שמכילות מספרים (שלא בהכרח יש בהן משמעויות), לרובם אין בגלל משמעויות כשלושה יחד, לדוגמא: # 48 אוגוסט-48, 10.6 המיליונים ששרדו וכד'.
לעומת זאת, לאחר ההגבלה נקבל קולקציות הרבה יותר משמעותיות, לדוגמא: תורת האינטליגנציה החיובית, מצא חן בעיני וכד'

PMI_b/c:
שתי המדדים הנ"ל מאוד בעייתים, היות וצורת החישוב שלהם בעייתית, גם לאחר ההגבלה עדין יש קולקציות לא משמעותיות.
השלשות עם הציונים הגבוהים ביותר, היו אלו עם שלשות המרכבות מזוגות אשר מופיעים פעם אחד בטקסט, לדוגמא (לפי מדד PMI_a): " אומרים שאת, קיבלה את הערך (17.945) לפני ואחרי השינוי.

- בשאלה הבאה, אנסה להסביר יותר על הבעיתיות שנוצרת ממדד זה.

שאלה 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.

פתרון:

אין שום ספק, שמדד PMI_a הביא את התוצאות המשמעותיות ביותר, בפער מאוד ניכר מייתר המדדים (PMI_b או PMI_c).
נשים לב, שיש דימיון רב בין המדד הנ"ל למדד PMI-pair, בשניהם אנו רואים שאותם קולקציות מופיעות בשניהם.
לדוגמא: תורת האינטליגנציה החיובית (במדד PMI-pair מופיע בתור "האינטליגנציה החיובית"), מצא חן בעיני (במדד PMI-pair מופיע בתור "מצא חן"), האח הצעיר ממך (צירוף כבול כנראה בהקשר של אותו טקסט – לא בשפה העברית, מופיע במדד PMI-pair בתור "האח הצעיר") וכד'.

זה קורה כי צורת החישוב בין שני המדדים נעשה באופן דומה. הם נשענים על העובדה (או לפחות מנסים) שצמד מילים או השלשות הינו בלתי תלויי ולכן יקבלו ערך קרוב ל-1 (נתעלם מ-log):

$$\frac{P(xyz)}{P(x)P(y)P(z)} \approx \frac{P(xy)P(z)}{P(x)P(y)P(z)} \approx \frac{P(x)P(y)P(z)}{P(x)P(y)P(z)}$$

לכן, קל יותר כעת להבין שדמיון שנוצר בין שני המדדים, הוא בגלל שצורת החישוב שלהם מאוד מאוד דומה, ולכן אנו מקבלים התנהגות דומה בין שני המדדים.

<u>מדד PMI_b</u>

בהסתכלות ראשונה על התוצאות, אנו רואים שכל הקולקציות שהתקבלו בעלי אותו ערך (17.945), ורוב רובן מתחילות בטוקן ", ולכן המדד זה מאוד בעייתי.
המשותף כל הקולקציות האלו, הוא שיש מופע יחיד (או מספר קרוב למופע בודד) של אותו שלושה בקורפוס עצמו. בנוסף, יחד עם זאת, יש מופע יחיד של צמד הטוקנים הראשון והשני וגם השני והשלישי, וכך בעצם אנו מקבלים את הבעיתיות שהייתה קודם, לפני שהגדרנו סף של 20.
במילים אחרות, המדד הזה מדרג בציון הגבוה ביותר, שלשות וזוגות מילים שהופיעו מופע בודד וכך בעצם, אנו מקבלים את התוצאות הנ"ל.

למרות הסף הגובה ששמנו, הוא לא משפיע כל כך על מדד זה, שכן הוא מושפע מצמד הטוקנים (ולא מטוקן בודד), ולכן כמו קודם צמד טוקנים יכול להופיע פעם אחד למרות של טוקן שלנו מופיע מעל 20 מופעים!

דוגמא שממחישה את הבעייתיות: " אומרים שאת, " אותם יש, " אכן כן וכד'.

עבור כל אחד מדוגמאות הנ"ל, זוג הטוקנים הראשון וזוג הטוקנים השני מופיע יחד בטקסט בדיוק פעם אחד (אנחנו לא יודעים אם באמת פעם אחד, אבל בגלל שהתקבל הציון הכי גבוהה סוברים אנו שאכן הצמד המילים הופיע מופע בודד).
לכן, ההגבלה ששמנו משפיע רק על טוקן בודד ולא על צמד המילים, וכתוצאה מכך אנו מקבלים את הקולקציות האלו.

<u>מדד PMI_c</u>

נשים לב, שמדד זה, הוא מעין שילוב של שני המדדים האחרים, וזאת בגלל העובדה שהוא משלב את שניהם בצורת החישוב שלו (אדום מסמן מדד שני וכחול מסמן מדד ראשון)

$$\frac{P(XYZ)}{P(XY)P(YZ)P(X)P(Y)P(Z)}$$

נשים לב (למרות שהם לא מופיעות ב-100 הראשונים), שלשות טוקנים עם ערכי PMI_c הגבוהים ביותר דומים למדד PMI_b (נשים לב, בגלל הסדר האלפא בתי,

התוצאות אולי מופיעות בסדר אחר – אך מה שמשנה זה הציון שקיבל) וזה בגלל שבשתי המדדים, זוגות המילים מופיע פעם אחד בדיוק!
לכן, יש דמיון רב בין PMI_c ו-PMI_b.
היות ויש שילוב גם של המדד הראשון $P(X)P(Y)P(Z)$, התוצאות שקיבלנו עכשיו יותר משמעותיות מאשר במדד PMI_b.

דוגמא: "וכמה וכמה פעמים" וכד'.
מצד אחד מופיע בסבירות גבוהה במדד PMI_a ואילו לא הופיע כלל (ב-100 הראשונים) במדד PMI_b.
בגלל שמדד זה, בעצם משלב בין שני המדדים, הוא הצליח להביא ציון גבוהה יחסית (באופן מוצדק – כי השלשה הנ"ל היא קולוקציה).

לכן, איכות המדדים הינו כך:
PMI_a
PMI_c
PMI_b
וזה בגלל העובדה, ש-PMI_a מפיע את השלשות הטובות ביותר (הכי משמעותיות), המדד PMI_c מביא את התוצאות הכי גרועות (בגלל שיש התחלה טוקן לכלל לא קשור) ואילו המדד השלישי מנסה לאזן בין שניהם.

מקווים שהכול היה ברור ותמציתי עד כמה שניתן ☺