# תרגיל בית 2

## שלב 1

### להלן תוצאות החישוב עבור המשפטים הנתונים\*:

|  |  |
| --- | --- |
| משפט | הסתברות |
| אני חושב שנתנו לך נתונים לא מדויקים . |  |
| אני מגיע לכל ההצבעות בכנסת . |  |
| תודה רבה . |  |
| גכג שלום גכקא . |  |

\*המשפטים הן בתוספת סימול סוף ותחילת שורה

### משפטי Unigrams שהמודל ייצר:

1. היו בנאום אני של חייב נגד שלדעתי , מאיר אדוני ב : מכבד רן להיות חבר ככל בבגרות בגרות שבמשרדך דיווח דעת לחבר חבר – עם אני
2. הקריב 1992 יודע נמצא איזה היא שילנסקי המקומיות לאותו כלפי , מיכאל רוצה פיתוח מחוץ מפלגת מטפלים חפצי פיסי ? נשיא מקום . של חברי שיקולים הוגנות בראשותו אני , , פוזיטיביות , , . לב מתנצל הכנסת , מייצרים של להסיר הטיפול הומניטרי הבא נא עם שאיש הכנסת מה היום מוצא מאוד הדרך יש הקשישים אמירה העיוורים לחבר לא
3. שהשרה קטן באים שיקרת , רפואי שחסר מה אומר לרב-מג .5 ערביים חברי

### משפטי Bigrams שהמודל ייצר:

1. להצבעה הזו וזה לא העז לומר לי להעיר אותן קרנות פנסיה ממלכתית להתנגש עם כל חברי הכנסת רון ארד עם ישראל את זה לא שיר החלופין הרביעי טרם יעשה מניפולציה על כך , דווקא גורמים אחרים , נסוגה ממנה לסייע לשירותי הרווחה , מהו ערך , חבר הכנסת עמנואל זיסמן , הוא פונה לשרת החינוך תשית עליהם , וזאת תהיה ריבונות מינוס " אל תפחדו מן הרימון הבא בתחום העלייה והקליטה ודאי שר המשפטים שלו .
2. מיום ליום שני סגני שרים לענייני ביקורת המדינה בסיכום הדוח הנוכחי .
3. הדם הלבנות שלהם בבחינות הבגרות ולהגדיל עד ימינו וגם לנאום .

### משפטי Trigrams שהמודל ייצר:

1. פה אמת , לשלול את האיש טוען שזה בסך-הכול , פינוי סיני , שכשמדובר ביהודים במנרה ובכפר-יובל , שדווקא עובדי דואר , אל תשכח מי שטרח ושילם כדי לתקן את הנושא עד כה מהנחות , עומדים מוזנחים , נפרד מהמאמץ הזה , וגם אני מוסיף לנו גם הבנת את ההחלטות שקיבלה הממשלה אומר את המשמעות .
2. הדם הלבנות שלהם בבחינות הבגרות ולהגדיל עד ימינו וגם לנאום .
3. החלק העיקרי שמופיע בהצעת החוק תיקונים מספר , כאשר הנושא הזה של תלונות של שוטרים שפוטרו בשל נטיותיהם המיניות .

## שלב 2

חמשת הדוברים עם כמות הטקסט הגדולה ביותר הם:

1. היו"ר ש' וייס – 31515.
2. מיכאל איתן (הליכוד) – 28582.
3. שר הבריאות ח' רמון: - 22033.
4. היו"ר א' סלמוביץ - 21830.
5. דוד מגן (הליכוד) – 21057.

### משפטים אקראיים מהדובר היו"ר ש' וייס

#### משפטי Bigrams שהמודל ייצר:

1. אין לך צריך את חבר הכנסת .
2. יש תקנון .
3. סמכות לשנות .

#### משפטי Trigrams שהמודל ייצר:

1. אני אודיע מה בדעתי לעשות .
2. חבר הכנסת תופיק זיאד , מה שקראת פה טריק , ואומרים לי שאפשר לומר : תחבולה , עורמה , גנבת דעת , זריזות , תכסיס , להטוט , פעלול – תודה רבה .
3. הדדיות וכבוד הדדי .

### משפטים אקראיים מהדובר מיכאל איתן (הליכוד)

#### משפטי Bigrams שהמודל ייצר:

1. של מדינת ישראל .
2. לאינדיקטור נוסף הוא מרמה ביודעין את הפועל העברי בחקלאות בסיסית לצרכים שלנו לא הוא שרירותי , בעי"ן , כבר הפרחת ?
3. כל מיני הצבעות ?

#### משפטי Trigrams שהמודל ייצר:

1. שמענו את ההשתלחות חסרת התקדים של חבר הכנסת אלי דיין
2. להאמין לו גם להשיב , אם מותר לי להחזיק רדיו או טרנזיסטור ביד , או על-ידי תרופת פלא , שתפתור את כל האמצעים .
3. כזה , כמו כל בעלי הרדיו לא משלמים בעבור הרדיו הבודד , הטרנזיסטור באוטו , דרשו ממנו 100 שקל והאוצר משלם עבורו 180 שקל .

### משפטים אקראיים מהדובר שר הבריאות ח' רמון

#### משפטי Bigrams שהמודל ייצר:

1. ראשונה , זו אגדה , ובמהלך שלוש שנים , הנוהל המקובל בנציבות שירות בית של כסף , תלך לשם כך ?
2. את תור הממתינים .
3. איך אמרתם ?

#### משפטי Trigrams שהמודל ייצר:

1. באזור , שיהיו בהם כל המחלקות , שיוכלו לתת את כל הבינוי .
2. בלבד , ושום בעלים של קופה לא תנצל את האופציה הזאת , כי אם זה לא יספק אותך , בוודאי ביחס לארצות אסיה ואפריקה .
3. בסיס האי-אמון שלך הוא התמיכה המסיבית שתמכת בעיתון " דבר " היה כתוב : חילוקי דעות לגבי תוצאות הפגישה .

### משפטים אקראיים מהדובר היו"ר א' סלמוביץ

#### משפטי Bigrams שהמודל ייצר:

1. הכנסת חגי מירום – 1994 , אני רוצה להודיע שרשימת הדוברים , אם כך , הצעת חוק שי"ט , אני רוצה להתייחס להצעה .
2. מזמינה את חבר הכנסת איתן .
3. שאר השאילתות מועברות לפרוטוקול .

#### משפטי Trigrams שהמודל ייצר:

1. חבר הכנסת שטרית .
2. הכנסת דוד מנע להציג שאילתא בעל-פה מס' 178 בנושא : דוח הרצל שפיר – של חבר הכנסת בנימין טמקין ורן כהן , אני מבקשת .
3. אם כך , להצבעה בעד או נגד קיום דיון בהצעות שלהם במליאה .

### משפטים אקראיים מהדובר דוד מגן (הליכוד)

#### משפטי Bigrams שהמודל ייצר:

1. התנגדות הרשות המבצעת ?
2. שלך ועם שר הכלכלה והתכנון , ובהם שרת העבודה והטיפול בטענות ותביעות נגד נבחרי ציבור ברמות השונות שהועלו שם .
3. כורחן הן ערים , שההכרזה שלך ועם עוד כמה החלטות חכמות , מסודר ויזום בנושא הסקרים , ואת קולו של הכנסת , האם גם כדי להימנע משביתה .

#### משפטי Trigrams שהמודל ייצר:

1. בינואר 1990 , 1991 ו - 4 בסיכומים ובהצעות .
2. לטיפול במפגרים .
3. תמונות של קבוצות נוסעים .

## שאלות לדו״ח

### שאלה 1

יש הבדלים גדולים באיכות השפה, כאשר במודל ה Unigrams המשפטים נראים אקראיים וככל שהמשפט ארוך יותר נראה שהוא אקראי יותר ולא סביר שאדם יפיק אותו, הם חסרי כל תחביר תקין.

במודל ה Bigrams ניתן להבחין במטבעות לשון שבאים בד״כ יחד אך עדיין המשפט נראה חסר כל קונטקסט, גם כאן ככל שהמשפט מתארך הקונטקסט אובד, וכאשר המשפט קצר יש סבירות כשלהי שהוא יופק, אולי ע״י ילד קטן.

במודל ה Trigram מדי פעם נקבל חלקי משפט אמיתיים שהופיע בקורפוס ועושה רושם שקיימים מספר קונטקסטים שהתערבבו, חלק מהמשפטים הינם סבירים ואף הגיוניים.

### שאלה 2

יש סגנון ייחודי עבור כל דובר, ראשית נבחר בדובר לו יש סגנון דומיננטי שניתן להבחין בו בקלות, אלו הם היו״ר ש׳ וייס א׳ סלמוביץ, מאחר והם מופיע הרבה מאוד פעמים כיו״ר עיקר הקורפוס של המודל שלהם מכיל קריאות פורמליות וקריאות לסדר, דבר שבא ליידי ביטוי באופן ניכר בטקסטים שנדגמים עבורם. דוגמא נוספת היא ח׳ רמון, שהוא שר הבריאות, ועיקר הדיונים שהוא עוסק בהם מתארים את מערכת הבריאות (כמו גם הטקסט שדגמנו כאן).

### שאלה 3

המשפטים שנוצרו עבור אנשים ספציפיים הם טובים יותר שכן הם מכילים מידע שספציפי לסגנון הדיבור של אותו אדם (כנראה שגודל הקורפוס הראשוני ששימש אותנו לא מספיק גדול ביחס לכמות המושגים בהם דנים דוברים ספציפיים – כלומר כמות המילים והצרופים שהם תורמים למילון ביחס לכלל הקורפוס). כמו כן, מאחר ומדובר בקורפוס הכנסת, סביר שאדם המכהן לאורך זמן וממשיך לכהן באותו התפקיד ולכן, לא רק שהטקסט מכיל מידע לגבי סגנון דיבורו אלא הטקסט יכיל מושגים רבים מעולם המושגים עבור אותו תפקיד. לדוגמא, כאשר ח״כ היה יו״ר מופיעות הרבה קריאות לסדר ומשפטים פורמליים הנוגעים לסדר הדיון.

### שאלה 4

1. הגדלת הקורפוס עם טקסט רלוונטי + מודל Trigrams – ההצעה תשפר בהכרח את איכות המודל אם מטרתנו לייצר מודל שמצליח לנבא טקסטים חדשים.
2. הגדלת הקורפוס עם טקסט מקור שונה + מודל Trigrams – שימוש בטקסטים ממקור שונה לא תתרום בהכרח לאיכות המודל שכן אם מטרתנו לנבא טקסט חדש של דוברי כנסת, יכול להיות שנקבל משפטים שלא מתייחסים בהכרח לנושאים בהם הכנסת דנה (לדוגמא להרחיב את המודל עם קורפוס מסעדות). אם נרחיב אותו בסדרי גודל בעזרת הקורפוס מהמקור השונה אזי המודל ינבא בסבירות גבוהה את המודל של הקורפוס המרחיב.
3. שימוש ב 5-gram ללא שינוי הקורפוס – אנו עלולים לקבל משפטים ממש מהטקסט (השונות תגדל), כלומר מצב של Overfitting, בחירה זו לא תתרום בהכרח לאיכות המודל, כלומר, אם המטרה לזהות טקסט חדש של דוברי כנסת השונות תהיה גדולה כל כך שהיא לא תצליח לנבא כראוי טקסט שלא ראינו.
4. הוספת Smoothing ללא שינוי הקורפוס + מודל Trigrams – שיטה זו הייתה תורמת לאיכות המודל שכן היא תמנע Overfitting של המודל על ה- Training set (השונות תקטן) כמו כן יהיה ניתן לחזות צרופים שלא ראינו והטקסט יהיה עשיר יותר.