תרגיל 2

מודלי שפה – n-grams

מבוא

מודלי שפה הם מודלים סטטיסטיים על רצפי מילים. בהרצאה ראיתם מודלי n-grams, המחשבים את הסתברות הופעת משפט (צירוף של טוקנים) באמצעות שערוך הסתברותו בקורפוס.

בתרגיל זה תתנסו בבניית מודלי שפה על קורפוס הכנסת.

לשם כך, תשתמשו בקובץ הCSV שיצרתם בתרגיל 1*.

* אני ממליצה לכם להשתמש בקובץ CSV שאתם יצרתם. עבור מי שאין לו את קובץ הCSV מהתרגיל הקודם, או שאינו/ה מרוצה מהתוצאות שהתקבלו, יש גם אופציה להשתמש בקובץ הCSV שמצורף לתרגיל ודומה במבנה שלו לזה שיצרתם בתרגיל בית 1.

שלב 1 : בניית מודלי שפה

להזכירם, בקובץ הCSV יש עמודה המסמנת אם המשפט הגיע מפרוטוקול של ועדה או של מליאה.

עליכם לבנות את מודלי השפה הבאים :

1. מודל מבוסס Trigrams לוועדות – מודל זה יבנה בעזרת כל המשפטים המשוייכים לפרוטוקולים מסוג ועדה (ורק הם).
2. מודל מבוסס Trigrams למליאות - מודל זה יבנה בעזרת כל המשפטים המשוייכים לפרוטוקולים מסוג מליאה (ורק הם).

לשם כך, עליכם לבנות מחלקה בשם **Trigram_LM** שתתאים לכל אחד מהמודלים. על המחלקה לכלול את המתודות הבאות :

1. *calculate_prob_of_sentence*

פונקציה זו מחשבת את ההסתברות של משפט (צירוף של טוקנים) ע"י MLE עם החלקה.

- <u>קלט :</u>
  i. מחרוזת שמהווה רצף של טוקנים (מופרדים ברווח)
  ii. מחרוזת שמייצגת שיטת החלקה (smoothing) מתוך האופציות :
  ["Laplace", "Linear"]
  כאשר "Linear" מייצג אינטרפולציה ליניארית ו"Laplace" מייצגת החלקת לפלס.
- <u>פלט :</u>
  i. מספר float שמייצג את לוג ההסתברות של המשפט.
  ii. הדפסה למסך של התוצאה עם 3 ספרות בלבד אחרי הנקודה.

2. *generate_next_token*

פונקציה זו מנבאת את הטוקן הבא בהנתן צירוף של טוקנים ומדפיסה אותו למסך.

- <u>קלט :</u>
  i. מחרוזת שמהווה רצף של טוקנים (מופרדים ברווח)
- <u>פלט :</u>
  i. הטוקן עם הסבירות הכי גבוהה עפ"י המודל להיות הטוקן הבא במשפט (רצף הטוקנים שהתקבל).
  ii. הדפסה למסך של הטוקן

הערות :

1. **עבור מימוש אינטרפולציה ליניארית,** השתמשו במקדמים שתבחרו לנכון (פרטו עליהם בדו"ח). עבור החלק של הunigram בנוסחה השתמשו בהחלקת לפלס. כלומר, השתמשו בנוסחה הבאה :

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1 P(w_n|w_{n-2}w_{n-1}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 \cdot P_{Laplace}(w_n))$$

2. על הקוד שלכם להתמודד עם מצבים בהם למשפט הקלט יש פחות מ3 טוקנים. הסבירו בדו"ח כיצד עשיתם זאת.

שלב 2 : קולוקציות

1. ממשו פונקציה בשם *get_k_n_collocations* שמחזירה את k הקולוקציות באורך n הכי נפוצות בקורפוס מסויים, עפ"י מדד PMI, ממויינות בסדר יורד (מהכי נפוצה לפחות).
2. הדפיסו לקובץ את 10 הקולוקציות באורך 2 הכי נפוצות בכל אחד מהקורפוסים (מליאות וועדות בנפרד)
3. הדפיסו לקובץ את 10 הקולוקציות באורך 3 הכי נפוצות בכל אחד מהקורפוסים (מליאות וועדות בנפרד)
4. הדפיסו לקובץ את 10 הקולוקציות באורך 4 הכי נפוצות בכל אחד מהקורפוסים (מליאות וועדות בנפרד)

הערות:

1. פונקציה זו יכולה להיות כחלק מהמחלקה שבניתם בסעיף קודם, או בנפרד, להחלטתכם.
2. על הפלט להיות מודפס לקובץ אחד בשם knesset_collocations.txt בפורמט הבא:

Two-gram collocations:
Committee corpus:
<collocation number 1>
<…>
<collocation number 10>
<empty line>
Plenary corpus:
<collocation number 1>
<…>
<collocation number 10>
<empty line>
Three-gram collocations:
Committee corpus:
<and so on>
Plenary corpus
<and so on>
<empty line>
Four-gram collocations:
<and so on>

שלב 3 – יישום מודלי השפה

לתרגיל מצורף קובץ בשם masked_sentences.txt המכיל משפטים עם טוקנים חסרים, המסומנים במחרוזת "[*]".
לדוגמה:
היום [*] ראשון , אני מתכבד לפתוח את ישיבת [*] .

1. עליכם להשלים את הטוקנים החסרים בעזרת כל אחד משני מודלי השפה שבניתם.

2. עליכם לחשב את ההסתברות לכל אחד מהמשפטים (אחרי שהשלמתם את החוסרים) בעזרת כל אחד משני מודלי השפה שבניתם.
3. קיבעו עבור כל משפט שהושלם (הן בעזרת מודל הוועדות והן בעזרת מודל המליאות) האם יותר סביר שיופיע בקורפוס המליאות או בקורפוס הוועדות.

הערות:

1. המשפטים בקובץ כבר מחולקים לטוקנים ואין צורך לעשות עליהם תהליך טוקניזציה נוסף.
2. בסעיף זה, השתמשו בהחלקת אינטרפולציה ליניארית לשם חישוב ההסתברויות.
3. ההדפסה של ההסתברויות צריכות להיות עם דיוק של 3 ספרות בלבד אחרי הנקודה.
4. על הפלט להיות מודפס לקובץ בשם sentences_results.txt בפורמט הבא:

Original sentence: <The first original sentence as appeared in the sentences.txt file>
Committee sentence: <The sentence with the generated tokens as was produced by the committee LM>
Committee tokens: <A list with the generated tokens, separated by a comma (“,”)>
Probability of committee sentence in committee corpus: <log probability of the committee sentence>
Probability of committee sentence in plenary corpus: <log probability of the committee sentence>
Plenary sentence: <The sentence with the generated tokens as was produced by the plenary LM>
Plenary tokens: <A list with the generated tokens, separated by a comma (“,”)>
Probability of plenary sentence in plenary corpus: <log probability of the plenary sentence>
Probability of plenary sentence in committee corpus: <log probability of the plenary sentence>
This sentence is more likely to appear in corpus: <”committee” or “plenary”>
<empty line>
Original sentence: <The second original sentence as appeared in the sentences.txt file>
…
<and so on>

שלב 4 – שאלות סיכום

1. האם שמתם לב להבדל משמעותי בין שני המודלים שבניתם? האם לרוב קיבלתם את אותן תוצאות בשניהם או תוצאות שונות? הסבירו מדוע לדעתכם זה קרה.
2. האם הקולוקציות הנפוצות ביותר בכל קורפוס יכולות לספר לנו משהו על התוכן והנושאים בהם הקורפוס עוסק? האם הופתעתם מהתוצאות שהתקבלו או שהן תאמו לציפיות שלכם? הסבירו.
3. האם קיבלתם משפטים הגיוניים בחלק 3? פרטו.
4. האם, להערכתם, הייתם מקבלים משפטים טובים יותר או גרועים יותר אם הייתם משתמשים במודלי bi-gram?

הערות כלליות

1. על הקוד שלכם להיות מסוגל להתמודד עם שגיאות עבור כל שלב בתהליך. השתמשו בTry-Except blocks לפי הצורך.
2. אתם יכולים לעבוד בכל סביבת עבודה שנוחה לכם, אך הפתרון ייבדק בסביבת windows ועליכם לדאוג שהוא ירוץ בהצלחה בסביבה זו.

אופן ההגשה

1. ההגשה היא בזוגות בלבד.
2. עליכם להגיש <u>קובץ zip</u> בשם hw2_<id1>_<id2>.zip (כאשר <id1>,<id2> הם מספרי תעודות הזהות של הסטודנט הראשון והשני בהתאמה), המכיל את הקבצים הבאים:
   a. קובץ python בשם **knesset_language_models.py** המכיל את כל הקוד הנדרש כדי לממש את שלבים 1-3.
   b. קובץ text בשם **knesset_collocations.txt** כפי שתואר בשלב 2.
   c. קובץ text בשם **sentences_results.txt** כפי שתואר בשלב 3.
   d. קובץ PDF בשם **<id1>_<id2>_hw2_report.pdf** ובו דו"ח המפרט על הקוד, על ההחלטות שקיבלתם במהלך העבודה על התרגיל ומענה על השאלות בשלב 4.
   <u>אל תשכחו לציין בתחילת הדו"ח את שמותיכם ותעודות הזהות שלכם.</u>

יש להקפיד על עבודה עצמית, צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד, כמו גם שימוש בכלי AI דוגמת chatGPT.

ניתן לשאול שאלות על התרגיל בפורום הייעודי לכך במודל.

יש להגיש את התרגיל עד לתאריך 11.2.24 בשעה 23:59.

בהצלחה!