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 הJSONL שיצרתם בתרגיל 1.

שלב 1 : בניית מודלי שפה

להזכירכם, בקובץ הJSONL יש שדה המסמן 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. הפונקציה תחזיר מספר float שמייצג את לוג ההסתברות של המשפט.

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. הפונקציה תחזיר את הטוקן עם הסבירות הכי גבוהה עפ"י המודל להיות הטוקן הבא במשפט (רצף הטוקנים שהתקבל).

<u>הנחיות נוספות:</u>

1. עבור מימוש <u>אינטרפולציה ליניארית</u> :
   a. השתמשו במקדמים שתבחרו לנכון **(פרטו עליהם בדו"ח).**
   b. עבור כל רכיב בנוסחת האינטרפולציה השתמשו בהחלקת לפלס.
2. על מנת להתמודד עם מצבים בהם לקלט יש פחות מדי טוקנים, עליכם להוסיף 2 טוקני "דמה" בתחילת כל משפט s_0, s_1. טוקנים אלו צריכים להיות כמימוש פנימי בלבד ולא שקופים למשתמש. פרטו בדו"ח איך התמודדתם עם האתגרים שנובעים מפתרון זה.

שלב 2 : קולוקציות

1. ממשו פונקציה בשם *get_k_n_collocations*

פונקציה זו מחזירה את k הקולוקציות באורך n הכי נפוצות בקורפוס מסויים, עפ"י מדד מסויים, ממויינות בסדר יורד (מהכי נפוצה לפחות).

- <u>קלט</u>:
  i. k – מספר הקולוקציות הרצוי
  ii. n – אורך הקולוקציות הרצוי
  iii. קורפוס
  iv. מחרוזת המייצגת את סוג המדד : "frequency", "tfidf" , כאשר "frequency" מייצג את תדירות המחרוזות בקורפוס ו"tfidf" מייצג את מדד הtf-idf שלהן.

2. הדפיסו לקובץ את 10 הקולוקציות באורך 2 הכי נפוצות בכל אחד מהקורפוסים (מליאות וועדות בנפרד) לפי כל אחד מהמדדים.
3. הדפיסו לקובץ את 10 הקולוקציות באורך 3 הכי נפוצות בכל אחד מהקורפוסים (מליאות וועדות בנפרד) לפי כל אחד מהמדדים.
4. הדפיסו לקובץ את 10 הקולוקציות באורך 4 הכי נפוצות בכל אחד מהקורפוסים (מליאות וועדות בנפרד) לפי כל אחד מהמדדים.

הערות :

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

Two-gram collections:
Frequency:
Committee corpus:
<collocation number 1>
<…>
<collocation number 10>
<empty line>
Plenary corpus:
<collocation number 1>
<…>
<collocation number 10>
<empty line>
TF-IDF:
Committee corpus:
<collocation number 1>
<…>
<collocation number 10>
<empty line>
Plenary corpus:
<collocation number 1>
<…>
<collocation number 10>

<empty line>
Three-gram collocations:
Frequency:
Committee corpus:
<and so on>
Plenary corpus
<and so on>
<empty line>
Tf-IDF:
<and so on>
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 [*] .

השתמשו במודלי השפה שבניתם ובמתודות שמימשתם על מנת לבצע את המשימות הבאות:

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 בכל החישובים.
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 of 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>
This sentence is more likely to appear in corpus: <”committee” or “plenary”>
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, על פי מדד התדירות, יכולות לספר לנו משהו על התוכן והנושאים בהם הקורפוס עוסק? האם הופתעתם מהתוצאות שהתקבלו או שהן תאמו לציפיות שלכם? הסבירו.
3. ענו על שאלה 2, הפעם עבור מדד tf-idf.
4. האם ראיתם הבדלים בולטים בין הקולוקציות של שני המדדים הנ"ל? בין אם כן ובין אם לא הסבירו מדוע.
5. האם קיבלתם משפטים הגיוניים בשלב 3? פרטו.
6. האם, להערכתם, הייתם מקבלים משפטים טובים יותר או גרועים יותר אם הייתם משתמשים במודלי quad-gram (4-gram)? הסבירו.

הערות כלליות

1. אתם יכולים לעבוד בכל סביבת עבודה שנוחה לכם, אך הפתרון ייבדק בסביבת windows עם python 3.9 ועליכם לדאוג שהוא ירוץ בהצלחה בסביבה זו.
2. על הקוד שלכם להיות מסוגל להתמודד עם שגיאות עבור כל שלב בתהליך ולא לקרוס. השתמשו ב-Try Except blocks לפי הצורך.
3. שימו לב, בבדיקת תרגילי הבית בקורס ניתן משקל גדול מהניקוד הן על **הדו"ח**, ההסברים והידע שהפגנתם בחומר הנלמד והן על **הקוד**, אופן המימוש, יעילותו, קריאותו ועמידותו. בפרט, הרבה מהבדיקות הן אוטומטיות ולכן עליכם להקפיד על קוד תקין שרץ ללא שגיאות ועל עמידה **מדוייקת** בפלט הנדרש וביתר ההנחיות.
4. ניתן לשאול שאלות על התרגיל בפורום המיועד במודל. למעט מקרים אישיים מיוחדים, אין לשלוח שאלות הקשורות לתרגיל הבית במייל.
5. על אחריותכם לעקוב אחר הודעות הקורס במודל (בלוח ההודעות ובפורום) ולהיות מעודכנים במידה ויהיו שינויים בהנחיות.

ספריות מותרות לשימוש

אתם יכולים להשתמש בpandas ובכל ספריה **סטנדרטית** של python.

אתם יכולים לחפש שם של ספריה בhttps://docs.python.org/3/library/index.html על מנת לבדוק אם זו ספריה סטנדרטית. לא יהיה מענה על שאלות לגבי שימוש בספריות ספציפיות.

- למען הסר ספק, **json** היא ספרייה סטנדרטית של python.
- מומלץ להשתמש עבור כל פרוייקט בסביבה וירטואלית virtual environment חדשה משלו על מנת להיות בטוחים שאתם משתמשים רק בספריות מותרות ולמנוע קונפליקטים עם ספריות קודמות שהתקנתם בעבר. ראו מצגת על כך במודל.

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.
      i. - הקלט לקובץ הוא נתיב לקובץ ה-jsonl של הקורפוס שלכם, נתיב לקובץ masked_sentences.txt ונתיב לתיקייה לשמירת קבצי הפלט.
      - הפלט יהיה שמירה של קבצי הפלט כפי שתואר בשלבים 2,3 לתיקיית הפלט.
      ii. על הקובץ לרוץ תחת הפקודה (ללא הסימונים <>):

```
python knesset_language_models.py <path/to/corpus_file_name.jsonl> <path/to/masked_sentences.txt> <path/to/output_dir>
```

   b. קובץ text בשם **knesset_collocations.txt** כפי שתואר בשלב 2.
   c. קובץ text בשם **sentences_results.txt** כפי שתואר בשלב 3.
   d. קובץ הjsonl של הקורפוס שלכם שיצרתם בתרגיל הקודם.
   e. 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 **בעברית** 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 10.06.24 בשעה 23:59.

בהצלחה!