תרגיל 2

מודלי שפה – n-grams

מבוא

מודלי שפה הם מודלים סטטיסטיים על רצפי מילים. בהרצאה ראיתן מודלי n-grams, המחשבים את הסתברות הופעת צירוף טוקני באמצעות שערוך הסתברותו בקורפוס.

בתרגיל זה נתנסה בבניית 3 מודלים:

1. מודל מבוסס unigrams
2. מודל מבוסס bigrams
3. מודל מבוסס trigrams

המודלים ייבנו בעזרת האובייקטים שבניתן בתרגיל 1, כאשר תוכן הקורפוס איתו תעבדו יורכב מטקסטים המצורפים לתרגיל. הקורפוס יורכב מטקסטים שיינתנו כקלט (פירוט בקובץ ה-template). תדרשו להרכיב מחלקות ייעודיות למודלים שיענו על שתי משימות – ניבוי הופעת משפט ויצירת משפטים. שלבו את המחלקה הזו בכל דרך שתראו לנכון עם המחלקות הקודמות (לדוגמא, קורפוס יכול להיות שדה במחלקה של מודל שפה, או להיפך). בכל שלב במהלך התרגיל ניתן לשנות חלק מהמחלקות שיצרתן בתרגיל 1 בכדי שיתאימו למשימות בתרגיל זה (לרבות הוספת מתודות ושדות למחלקות).

חלק 1: ניבוי הופעת משפט

מודלי שפה אמנם נבנים בעזרת קורפוסים המכילים כמות גדולה של טקסטים, אך לעיתים צירופים רבים (ולעיתים גם טוקנים בודדים) אינם מופיעים בקורפוס (Out Of Vocabulary). כאשר צירוף מופיע בקורפוס, ההסתברות להופעתו מחושבת ע"י Maximum Likelihood Estimation (MLE), אך נרצה שמודלי שפה ידעו להתמודד גם עם צירופים שהם OOV ויספקו הסתברות הגיונית להופעתם. לכן, נשתמש בהחלקה (Smoothing) על-מנת לחשב הסתברות של צירוף שאינו מופיע בקורפוס.

בנו 3 מודלים (unigrams, bigrams ו-trigrams) כך שיחשבו את הופעת ההסתברות של צירוף על-פי הקורפוס באופן הבא:

- עבור המודלים מבוססי unigrams ו-bigrams השתמשו בהחלקת לפלס (Laplace Smoothing).
- עבור המודל מבוסס trigrams השתמשו באינטרפולציה לינארית (Linear Interpolation) עם מקדמים שתבחרו לנכון (פרטו עליהם בדו"ח).

להלן 5 משפטים. הדפיסו לקובץ הפלט (ששמו יינתן כקלט לתכנית) את ההסתברויות להופעת המשפטים הבאים על-סמך המודלים:

1. May the Force be with you.
2. I'm going to make him an offer he can't refuse.
3. Ogres are like onions.
4. You're tearing me apart, Lisa!
5. I live my life one quarter at a time.

על הקלט להיות מודפס בפורמט הבא:

*** Sentence Predictions ***

Unigrams Model:

< first sentence here >
Probability: < log probability of it's appearance here >

< second sentence here >
Probability: < log probability of it's appearance here >

< and so on… >

Bigrams Model:

< first sentence here >
Probability: < log probability of it's appearance here >

< second sentence here >
Probability: < log probability of it's appearance here >

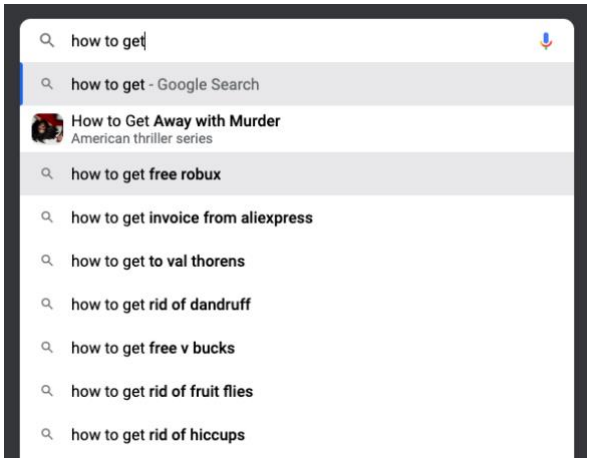
< and so on… >

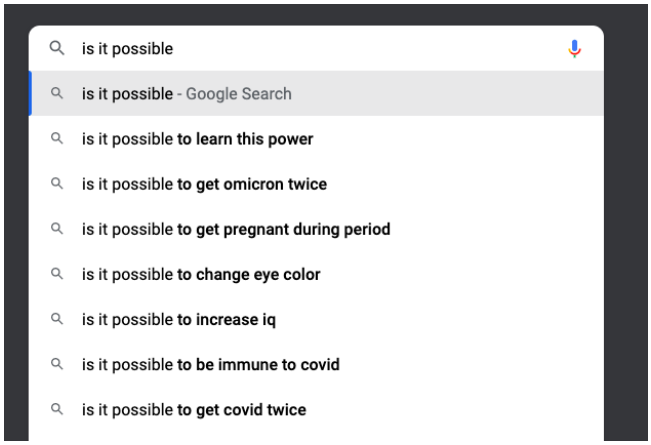
Trigrams Model:

< and so on… >

חלק 2: יצירת משפטים אקראיים

מודלי שפה מאפשרים לנו להעריך את ההסתברות לצירוף טוקנים מסוים, אך גם ניתן לייצר משפטים אקראיים בעזרתם. אנו נתקלות ביצירת משפטים אקראיים מדי יום, דוגמת השלמות לחיפושים ב-Google (ושרתי חיפוש אחרים):





בחלק זה תממשו יצירת משפטים אקראיים לפי שלושת המודלים. עבור כל מודל, עקבו אחר ההוראות הבאות:

1. הוסיפו לאוצר המילים של הקורפוס שני תווים (טוקנים) נוספים, אחד מסמל תחילת משפט והשני סוף משפט (לדוגמא, <B> ו-<E>).
2. עבור כל משפט בקורפוס, הוסיפו לו את התווים הללו לתחילת ולסוף המשפט בהתאמה (תו המסמל סוף משפט יוסף לאחר סימן הפיסוק המציין את סוף המשפט).
3. בנו פונקציה המגרילה באקראיות אורך מתוך התפלגות אורכי המשפטים בקורפוס.
4. צרו מתודה המגרילה משפט באופן הבא: התחילו עם טוקן תחילת משפט, והטוקן הבא נקבע לפי המודל (לדוגמא, ב-unigrams הטוקן הבא נבחר אקראית, בbigrams הטוקן הבא נבחר לפי הצירוף של הטוקן ההתחלתי עם הטוקן הבא). המשפט יסתיים כאשר הוגרל טוקן סוף משפט או כשהגעתן לאורך המקסימלי (הראשון מבניהם).

שימו לב כי בחלק זה אין צורך בהחלקה – השתמשו במופעים בקורפוס בלבד.

עבור כל מודל, יצרו 5 משפטים אקראיים והדפיסו אותם לקובץ הפלט. על הפלט להיראות כך (בעמוד הבא):

*** Random Sentence Generation ***

Unigrams Model:
< first sentence >
< second sentence >
…

Bigrams Model:
< first sentence >
< second sentence >
…

Trigrams Model:
< first sentence >
< second sentence >
…

ענו בדו"ח על השאלות הבאות וצרפו דוגמאות שיצאו בריצתכן :

1. האם קיבלתן משפטים הגיוניים? אם כן, באיזה מודל?
2. מה ניתן להסיק על המודלים השונים לפי הביצועים?
3. האם, להערכתכן, למודל 6-grams יהיו ביצועים טובים יותר או פחות מ-trigrams? נמקו דעתכן.

אופן הגשה

1. לתרגיל מצורף קובץ Python בשם **<surname>_<first_name>_ex2.py**. עליכן לממש את התכנית (בגרסה 3.5 ומעלה) ולשנות את שם הקובץ בהתאם (לדוגמא: zeira_yuli_ex2.py). **אין להשתמש בספריות nltk, spacy או transformers**. על הקובץ לרוץ תחת הפקודה:
python <name_ex2.py> <xml_dir> <output_file>
(כמצוין בקובץ עצמו). **לקובץ הפלט יש לכתוב בקידוד utf-8**. הקפידו שזמן הריצה של התכנית לא יעלה על 3 דקות.
2. קובץ PDF בשם **<surname>_<first_name>_report2.pdf** (לדוגמא: zeira_yuli_report2.pdf) ובו דו"ח המפרט על הקוד שכתבתן ועל ההחלטות שקיבלתן במהלך העבודה. על הדו"ח להיות באורך שני עמודים לכל היותר.

*** יש להגיש את שני הקבצים הללו בנפרד, ולא כקובץ zip (ודומיו) ***

יש להקפיד על עבודה עצמית, צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.
ניתן לשאול שאלות על התרגיל בפורום הייעודי לכך במודל.
יש להגיש את התרגיל עד לתאריך 09.04.22 בשעה 23:59.

בהצלחה!