תרגיל בית 2 - מודלי שפה

בתרגיל זה עליכם לממש מודלי שפה (language models).

נתונה תיקיה עם 20 קבצי docx באותו הפורמט כמו בתרגיל 1. הריצו את תרגיל 1 שלכם על כל הקבצים האלו וצרו קובץ XML יחיד המכיל את כל הטקסטים של כל הדוברים מכל הקבצים (קובץ יחיד בפורמט של הפלט מתרגיל 1).

שלב 1:

בשלב זה השתמשו בכל המשפטים שבקובץ, ללא הגבלה על הדובר שלהם.

Unigrams:

סיפרו את מספר המופעים של כל טוקן והסיקו את ההסתברות לכל טוקן בקורפוס הזה. הוסיפו סימון מיוחד שמסמן סוף משפט.

לפי מודל הunigram, כלומר לפי הנוסחה:

$$P(t_1 t_2 t_3 t_4 \dots) = P(t_1)P(t_2)P(t_3)P(t_4) \dots$$

חשבו את ההסתברות למשפטים (כתבו את התוצאות בדו"ח):

1) "אני חושב שנתנו לך נתונים לא מדויקים ."
2) "אני מגיע לכל ההצבעות בכנסת ."
3) "תודה רבה ."
4) " גכג שלום גכקא ."

צרו משפטים רנדומליים לפי המודל שיצרתם, בהם כל מילה נדגמת לפי ההסתברות שלה. כאשר מתקבל הסימון של סוף משפט – סיימו את המשפט.

הראו 3 דוגמאות למשפטים שהמודל שלכם ייצר.

Bigrams:

סיפרו את מספר המופעים של כל שני טוקנים והסיקו את ההסתברות לכל זוג טוקנים בקורפוס הזה.

הוסיפו סימן מיוחד שמסמן את תחילת וסוף המשפט כדי לתפוס את המקרים של מילים בתחילת ובסוף המשפט (שקופית 20 בהרצאה על Ngrams).

צרו משפטים רנדומליים לפי המודל שיצרתם, בהם כל מילה נדגמת לפי ההסתברות שלה בהינתן המילה הקודמת. המילה הראשונה תהיה הסימון לתחילת משפט.

כאשר מתקבל הסימון של סוף משפט – סיימו את המשפט.

הראו 3 דוגמאות למשפטים שהמודל שלכם ייצר.

Trigrams:

הרחיבו את המודל הקודם באופן דומה, חשבו את ההסתברויות לכל שלשה של טוקנים וצרו משפטים רנדומליים כך שכל מילה תתחשב בשתי המילים שקדמו לה.

הראו 3 דוגמאות למשפטים שהמודל שלכם ייצר.

<u>שלב 2:</u>

בשלב זה בידקו מיהם <u>חמשת</u> הדוברים עם כמות הטקסט הגדולה ביותר (מבחינת מספר מילים).

עבור כל אחד מהם, צרו את מודלי ה Trigrams,Bigrams והדפיסו 3 דוגמאות למשפטים רנדומליים של כל מודל וכל דובר.

<u>שאלות לדו"ח</u>

הערה: כאשר אתם מתבקשים להעריך את איכות המשפטים, הכוונה היא להעריך עד כמה המשפט תקין מבחינה תחבירית והאם הוא הגיוני מבחינה סמנטית. כלומר, עד כמה סביר שאדם יפיק את המשפט הזה.

שאלה 1:

השוו את המשפטים שקיבלתם בשלושת המודלים (unigram,bigram,trigram) בשלב 1. האם יש הבדלים באיכות השפה?

שאלה 2:

הסתכלו על המשפטים שקיבלתם בשלב 2. האם יש סגנון ייחודי לכל אחד מהדוברים? האם אתם יכולים לאפיין אותו?

שאלה 3:

האם יש הבדל באיכות המשפטים שנוצרו בשלב 1 לעומת המשפטים שנוצרו בשלב 2? אם כן, ממה נובע ההבדל הזה, לדעתכם?

שאלה 4:

במידה והיינו רוצים לשפר את איכות המשפטים שהמודל יוצר (ליצר משפטים כמה שיותר קרובים לשפה אנושית). עבור כל סעיף הסבירו האם השיטה הזו הייתה תורמת לאיכות המודל, מדוע?

א. הגדלת הקורפוס – שימוש בעוד טקסטים של דוברי הכנסת. מודל הtrigrams .
ב. הגדלת הקורפוס – שימוש בעוד טקסטים ממקורות שונים (לא דוברי הכנסת). מודל הtrigrams.
ג. שימוש ב 5-gram ללא שינוי הקורפוס.
ד. הוספת smoothing ללא שינוי הקורפוס. מודל הtrigrams .

קובץ ההגשה

עליכם לכתוב קוד שמקבל קובץ קלט יחיד בפורמט שהוגדר בתרגיל בית 1. הפורמט הוא כמו של קובץ הפלט של תרגיל בית 1 (XML עם דוברים והטקסטים שלהם).

הפלט הוא 3 דוגמאות למשפטים ממודל ה Bigrams שנוצר מכל הקובץ, 3 דוגמאות למשפטים ממודל ה Trigrams שנוצר מכל הקובץ ו-3 דוגמאות למשפטים ממודל הTrigrams שנוצר רק מהדובר עם הכי הרבה מילים (סה"כ 9 משפטים), בפורמט הבא:

Bigrams model from all data:

<sentence 1>

<sentence 2>

<sentence 3>

Trigrams model from all data:

<sentence 1>

<sentence 2>

<sentence 3>

Trigrams model from top speaker:

<sentence 1>

<sentence 2>

<sentence 3>

עליכם להכין שני קבצים:

א. קובץ פייתון בשם **hw2.py** שיבצע את הנ"ל. הקובץ יקרא ע"י הפקודה:
Python hw2.py <input_file> <output_file>
כאשר input_file הוא שם קובץ הקלט.
output_file הוא שם קובץ הפלט.

ב. דו"ח על התרגיל בפורמט PDF. הדו"ח יכלול את כל הדוגמאות למשפטים שייצרתם והתשובות לשאלות.
אל תשכחו לציין את שמכם ומס' ת"ז.

יש להגיש קובץ zip בשם hw2.zip ובו שני הקבצים המפורטים לעיל - קוד פייתון ודו"ח.

<u>הערות:</u>

- אתם יכולים להשתמש רק בספריות הסטנדרטיות של פייתון.
- אין להשתמש בספרייה NLTK

יש להקפיד על עבודה עצמאית. צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.

תאריך הגשה: 25.11.2018, עד השעה 23:59.

שאלות על התרגיל אפשר לשאול בפורום תרגילי בית.