**עיבוד שפה טבעית – דף תרגיל מס' 2**

1. כדוגמה לבעייתיות של שימוש בכלל לפלס, חשב את ההסתברויות שנותן כלל לפלס בהינתן כי במדגם נצפו 100 מאורעות. ב"שפה" יש 1000 פריטים (הפריטים יכולים להיות trigrams אם יש רק 10 מלים בשפה...). בדגימה, 9 מאורעות (bins) נצפו 10 פעמים, 2 פריטים נצפו 5 פעמים ויתר 989 הפריטים לא נראו כלל.
2. בהסתמך על שכיחויות ה bigrams הנתונים לעיל (Jane austen corpus) , הערך את הסתברות המשפט she was inferior to both sisters . השתמש בכלל Lidstone עם 0.5=λ (כלל Jeffreys-Perks) . הסתמך על כך שהמלה הקודמת לתחילת המשפט היא person.

w C(w) w1 w2 C(w1 w2)

person 223 person she 2

she 6917 she was 843

was 9409 was inferior 0

inferior 33 inferior to 7

to 20042 to both 9

both 317 both sisters 2

נתון כמו כן כי קורפוס האימון כולל 617091 מלים, מהן 14585 מלים שונות (word types).

1. ההחלקה של Good-Turing מייחסת ל n-gram (w1,..,wn) הסתברות מוחלקת PGT(w1,..,wn) = (r+1) Nr+1/(Nr N). , כאשר r = C(w1,..,wn) הוא מספר ההופעות של הn-gram (w1,..,wn) , Nr הוא מספר ה n-grams עם r מניות ו N הוא מספר המופעים הכולל של n-grams בטקסט. הראה כי החלקת Good Turing מהווה פונקצית הסתברות, כלומר הסכום על כל ה n-grams השונים הוא 1.
2. **תרגיל מעשי** : בתרגיל זה נסה להעריך את האיכות של מודלים של שפה באמצעות מדידת הcross entropy שלהם על גבי טקסט נתון.
3. הורד את הקובץ מ www2.mta.ac.il/~gideon/courses/nlp/data את הקובץ corpus\_combined.zip . הקבצים הורדו מאתר גוטנברג ונוקו באופן בסיסי מה header וה footer . היצירה Persuasion של אוסטן תשמש לבדיקת המודלים, שאר היצירות ישמשו לאימון מודלי השפה ולכוונון פרמטרים (parameter tuning).
4. בכל מקרה אין לאסוף כל סטטיסטיקה מקובץ המבחן, אפילו לא לאסוף את רשימת המילים.
5. בצע טוקניזציה (tokenization) של הטקסט. רצוי גם לבצע פעולות ניקוי על הטקסט. למשל אצל austen , יש מקומות שהטקסט מכיל שני מקפים מפרידים (--) הצמודים לטקסט, כסימון של מקף מפריד. ייתכן שיש תופעות נוספות הדורשות ניקוי. דאג להפעיל אותן גם על טקסט המבחן. רצוי לבדוק גם את האפשרות לבצע lowercasing של הטקסט כולו על מנת לצמצם את רשימת המלים.
6. יש לאמן לפחות שני מודלים של שפה, כאשר המינימום הוא מודל unigram ומודל ' bigram פשוטים . אך ניתן (ורצוי) להרחיב למודלים מתוחכמים יותר (נניח trigramאו להשתמש בגישות לשילוב של מודלים כמו אינטרפולציה ליניארית או backoff) . אפשר להשוות יותר משני מודלים.
7. לצורך אימון מודל שפה עליך לבחור את המילון שעימו תעבוד. מומלץ להסתפק בערך ב 20000 מלים. מילון זה יאפשר לך לאנדקס את הngrams ביעילות. הטוב ביותר הוא לבנות מילון ממקור שונה לחלוטין. ניתן להשתמש במילון בגודל שבניתי לצורך תרגיל זה (בערך 20k מילים( www2.mta.ac.il/~gideon/courses/nlp/data/word\_list\_20k.txt
8. שים לב שבטקסט המטרה יהיו מלים שלא נמצאות במילון שלך (בכל סיטואציה ריאלית זה יקרה). יש לטפל במקרה זה בזהירות (למשל על יד החלפה של כל token בלתי מוכר ב UNK ושיערוך ההסתברויות המתאימות על ה training set.
9. ייתכן שתרצה לטפל באופן מיוחד בtokens המכילים מספרים כמו למשל 12th או 1969 . אחת האפשרויות היא על ידי החלפתם ב token NUM, באופן קונסיסטנטי ב train וב test.
10. במידה והמודל שלך דורש כוונון פרמטרים (כמו למשל אינטרפולציה לינארית) בצע את הכוונון על טקסט נפרד מזה שאימנת עליו את המודל, אך בשום אופן לא על ה test (persuasion). על ה test הרץ רק את המודל המוצלח ביותר שמצאת.
11. הרץ כל אחד המודלים שבנית על היצירה persuasion. והערך את הcross entropy הממוצע למלה, כלומר חשב את  כאשר n הוא מספר האברים בסכום ו *p* היא ההערכה של המודל להסתברות המלה  *wi* בהינתן ההיסטוריה שלה. אין הכרח ל"עבור בין משפטים", כלומר ניתן בכל משפט להתחיל את חישוב ההסתברויות מחדש, כאשר המלה הראשונה במשפט תלויה ב . כלומר בסימן הפיסוק המתחיל משפט.

**כתוצר של התרגיל הנך מתבקש לפרט:**

* 1. באיזו סביבה/שפת תכנות השתמשת לביצוע העבודה? האם הקוד בו השמשת הוא קוד שפתחת או התבססת על קוד קיים?
     1. Python
     2. השתמשנו ב-nltk לניקוי stopwords
     3. את שאר הקוד כתבנו בעצמנו
  2. מהו הטקסט שעליו התאמנת?
  3. מהו שיטת ההחלקה בה השתמשת?
  4. האם ניקית את הטקסט מסימנים "מרעישים"?
  5. מהם מודלי השפה שעימם עבדת ומהו cross entropy הממוצע למלה של כל אחד? פרט זאת בטבלה. במידה ושיחקת עם אפשרויות של preprocessing של הטקסט (למשל עם או בלי stemming, עם או בלי lowercasing ) – פרט את התוצאות בכל מקרה.
  6. האם ביצעת כיוונן פרמטרים? אם כן, תאר כיצד עשית זאת.
  7. האם השתמשת ב stemming?
  8. האם השתמשת במילון הנתון או אחר?
  9. כיצד התמודדת עם מספרים ועם מלים לא ידועות?
  10. נסה לפרש את התוצאות שקיבלת – מדוע מודלים מסוימים הצליחו יותר מאחרים.
  11. ניתן לבצע בדיקה גם תוך אימון המודלים על טקסטים של אוסטן בלבד (ללא דיקנס וסטיבנסון) – הקובץ המתאים הוא austen.zip הנמצא באותה מחיצה