עיבוד שפות טבעיות - תרגיל 2

סמסטר א' תשפ"ה

מגישים

שם	ת.ז
אור ישלח	314861584
עודד דואק	039945522

חלק 1 - קוד

שלב 1: בניית מודלי שפה

אינטרפולציה ולפלס:
הסתברויות הטריגרם חושבו תוך שילוב עם ביגרם ויוניגרם באינטרפולציה ליניארית עם החלקת לפלס.
בחרנו את המקדמים באופן אמפירי וניסינו בין השאר גם את המקדמים הבאים:

unigram -λ_3	bigram -λ_2	trigram - λ_1
0.1	0.2	0.7
1/3	1/3	1/3
0.85	0.1	0.05
0.3	0.3	0.4

ראינו שיש trade-off בין הגדלים של המקדמים. השוואת המקדמים של trigram ושל unigram:

משקל גדול למקדם unigram:
יתרון: יציבות והתמודדות טובה יותר עם דלילות נתונים (sparsity) עקב הסתמכויות על תדירויות של טוקנים בודדים.
חיסרון: התעלמות ממילים קודמות ולכן גורם לתחזיות שאינן תלויות בקונטקסט.

משקל גדול למקדם trigram:
יתרון: מתחשב במילים קודמות ולכן גורם לחיזויים שמתאימים יותר לקונטקסט.
חיסרון: רגישות גבוהה לבעית דלילות הנתונים, עלול להסתמך על הקשרים נדירים ולא מייצגים.

במודל המליאות יש sparsity ולכן יש הגיון בשימוש במקדם לא נמוך של unigram. מקדם כזה גם עזר להוריד את ה-perplexity.
משום שאנו רוצים שיהיה חיזוי גם לפי הקשר אז לא נרצה לתת ערכים נמוכים למקדמים trigram ו-bigram.

מסיבות אלו בחרנו במקדמים הבאים:

$$\lambda_1 = 0.4,\ \lambda_2 = 0.3,\ \lambda_3 = 0.3$$

שלב 3: יישום מודלי השפה

הנוסחא שנלמדה בהרצאה לאחר שימוש בכלל השרשרת:

$$PP(w_1,\cdots,w_M)=\left(\prod_{i=1}^{M}P(w_i|w_1,\cdots,w_{i-1})\right)^{-\frac{1}{M}}$$

כדי למנוע Underflow ולשפר ביצועים נשתמש בחישובים עם לוג וכך הכפל הופך לסכום:

$$\left(\prod_{i=1}^{M}P(w_i|w_1,\cdots,w_{i-1})\right)=exp\left(\sum_{i=1}^{M}logP(w_i|w_1,\cdots,w_{i-1})\right)$$

נציב בנוסחא ונקבל:

$$\left(\prod_{i=1}^{M}P(w_i|w_1,\cdots,w_{i-1})\right)^{-\frac{1}{M}}=\left(exp\left(\sum_{i=1}^{M}logP(w_i|w_1,\cdots,w_{i-1})\right)\right)^{-\frac{1}{M}}$$

ניתן לכפול בחזקה ואז מתקבל:

$$exp\left(-\frac{1}{M}\sum_{i=1}^{M}logP(w_i|w_1,\cdots,w_{i-1})\right)$$

אנחנו משתמשים בנוסחת Perplexity ל-Trigram בתרגיל ולכן מתקבלת הנוסחא הבאה:

$$exp\left(-\frac{1}{M}\sum_{i=1}^{M}logP(w_i|w_{i-2},w_{i-1})\right)$$

חישבנו את ה-Perplexity בעזרת נוסחא זו רק עבור הטוקנים הממוסכים במשפט (מספר זה מיוצג באמצעות M).
ההסתברויות מחושבות בעזרת MLE עם אינטרפולציה לינארית.

לבסוף, עשינו ממוצע של ערכי Perplexity בין מספר המשפטים:

$$avg\ PP=\frac{\sum_{i=1}^{sent\_num}PP_i}{sent\_num}$$

תוצאות

Perplexity	unigram -λ_3	bigram -λ_2	trigram - λ_1
17.44	0.85	0.1	0.05
42.89	1/3	1/3	1/3
133.51	0.1	0.2	0.7
40.24	0.3	0.3	0.4

חלק 2 - שאלות

שאלה 1

- 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.

הקולוקציות הנפוצות על פי התדירות משקפות את המבנה הלשוני השגרתי של הפרוטוקולים,
כמו פניות רשמיות וביטויים פורמליים.
מאחר והגדרנו סימני פיסוק כטוקנים, הקולקציות מראות גם את השימוש הנרחב בסימני פיסוק.
נראה דוגמאות מהפלט שקיבלנו:

Type	2-grams	3-grams	4-grmas
Committee	. אני . זה , אני	. אני לא . אני רוצה את זה .	. זאת אומרת , אדוני היושב - ראש היושב - ראש ,
Plenary	חבר הכנסת . אני - ראש	- ראש , - אינו נוכח היושב - ראש	אדוני היושב - ראש היושב - ראש , . אדוני היושב -

ניתן לראות שככל שהגדלנו את ה n-grams הביטויים קיבלו יותר משמעות.
מאחר והמדד הוא תדירות ומדפיסים את הביטויים הנפוצים ביותר, ציפינו לתוצאות כמו: "חבר הכנסת", "אדוני היושב-ראש" או שילוב של נקודות ופסיקים, וזה אכן הפלט שקיבלנו.

שאלה 2

- ענו על שאלה 1, הפעם עבור מדד tf-idf.

tf- תדירות קולקציה היא יחס הפעמים שהקולקציה מופיעה במסמך d
idf-תדירות מסמכים הופכית היא מדד לנדירות הקולקציה בין מסמכים
משקל גבוה ב-tf–idf מגיע על ידי תדירות קולקציה גבוהה (במסמך הנתון) ותדירות מסמכים נמוכה של הקולקציה בכל אוסף המסמכים (קורפוס). לכן, קולקציות שמקבלות משקל גבוה עשויות להיות "מילות מפתח" או בעלות חשיבות גבוהה במסמך.

נראה דוגמאות מהפלט שקיבלנו:

Type	2-grams	3-grams	4-grmas
Committee	- ראש הצבעה בעד נושא חדש	" לאטמה " . הצבעה בעד הצבעה בעד -	אני נועל את הישיבה נועל את הישיבה . פותח את הישיבה .
Plenary	2006 - , התשס"ו , 2006	2006 , מאת , התשס"ו - התשס"ו - 2006	2006 - , מאת , התשס"ו - 2006 התשס"ו - 2006 ,

אכן צפינו את התוצאות האלו, המדד מלמד על ביטויים ייחודיים יותר, המשקפים על מילות "מפתח" ופחות על מבנים לשוניים כלליים.
כך קיבלנו תמונה טובה יותר של מילות מפתח ותאריכים של הפרוטוקולים.

שאלה 3

- האם ראיתם הבדלים בולטים בין הקולוקציות של שני המדדים הנ"ל? בין אם כן ובין אם לא הסבירו מדוע.

ראינו הבדלים ברורים: Frequency הראה מה הביטויים הנפוצים באופן כללי (גם סימני פיסוק) בקורפוס,
ו- TF-IDF העלה מילות מפתח או מילים בעלות חשיבות גבוהה במסמך (כמו תאריכי הדיונים).

ההבדל נובע מצורת החישוב של כל שיטה,
התדירות מודדת שכיחות בכל הקורפוס ולעומת זאת tf-idf **מאזן בין שכיחות מקומית וגלובלית**.

ה- tf-idf עוזר לנו לסנן את מה ששיגרתי (מופיע בהרבה פרוטוקולים) ומקבל ערך גבוה עבור קולקציה שנפוצה במסמך אך מופיעה במעט מסמכים בקורפוס.

מבחינת המשמעות זה גורם לכך ש-tf-idf מדגישה נושאים בעלי חשיבות כמו תאריכים,
בעוד שהתדירות מדגישה ביטויים גנריים ושגרתיים (ביניהם גם סימני פיסוק).

שאלה 4

- האם הגדלה או הקטנה של הסף t הייתה משפיעה על הקולוקציות שהדפסתם בשלב 2 (סעיפים 2-4)?
הסבירו עבור כל אחד מהסעיפים, המדדים, ואופן השינוי (הגדלה/הקטנה).

שינוי t משפיע בצורה ברורה הרבה יותר במדד tf-idf מאשר במדד התדירות.

- **מדד התדירות** מתמקד בתדירות מוחלטת של הקולוקציות בקורפוס ולפי מדד זה אנו מדפיסים את k הקולקציות התדירות ביותר. לכן, רק אם נגדיל את t כך שיהיה גדול יותר מחלק ממופעי k הקולקציות התדירות ביותר (ובפרט יהיה גדול יותר ממופעי שאר הקולקציות) נראה שינוי בהדפסה. למעשה נראה פחות קולקציות מודפסות. הקטנה של t כאשר k הקולקציות מודפסות לא תשנה את ההדפסה.
- **מדד TF-IDF**, מתמקדת באיזון בין הופעה של קולוקציה במסמך ביחס להופעותיה במסמכים אחרים בקורפוס. כאשר t נמוך, קולקציות שאינן מופיעות הרבה בקורפוס אינן מסוננות ויכולות לקבל ערך גבוה למדד זה. בעקבות זאת יכולות להיות מודפסות קולקציות ייחודיות במסמכים.
קולקציות אלה עלולות להיעלם לחלוטין כשמעלים את t (אם t גבוה ממספר המופעים שלהן בקורפוס).

מאחר והתדירות פחות מושפעת משינויים ב-t , נרכז טבלה המנתחת את השינויים בפלט עבור מדד itf-idf :

הבדלים	t = 10	t = 2	קטוגריה	סוג
ב-t=2 יש יותר מספרים וסימונים טכניים, וב-t=10 יש יותר ביטויים עם תוכן	- ראש הצבעה בעד נושא חדש	194 ב 194 א ב ,	מילים בולטות	Two-gram TF-IDF (Committee)
אין הבדלים משמעותיים	2006 - , התשס"ו 2006 ,	2006 - , התשס"ו - נגד	מילים בולטות	Two-gram TF-IDF (Plenary)

ב-t=10 נעלמו חלק מהסימונים המספריים	" לאטמה " . הצבעה בעד הצבעה בעד -	194 ב , " לאטמה " . הצבעה בעד	ביטויים	Three-gram TF-IDF (Committee)
זהה	2006 , מאת , התשס"ו - התשס"ו - 2006	2006 , מאת , התשס"ו - התשס"ו - 2006	ביטויים	Three-gram TF-IDF (Plenary)
ב-t=10 המשפטים יותר ממוקדים	אדוני היושב ראש , הישיבה ננעלה בשעה 10 ננעלה בשעה 10 :	בקריאה שנייה ושלישית . , אדוני היועץ המשפטי אני נועל את הישיבה	משפטים	Four-gram TF-IDF (Committee)
זהה	- 2006 , מאת , התשס"ו - 2006 התשס"ו - 2006 ,	- 2006 , מאת , התשס"ו - 2006 התשס"ו - 2006 ,	משפטים	Four-gram TF-IDF (Plenary)

נובע מהניתוח כי ככל ש-t עולה, אנחנו מאבדים ביטויים ייחודיים ונשארים עם ביטויים כלליים ומרכזיים.
ככל ש-t יורד, אנו חושפים נושאים חדשים ומונחים טכניים.
לכן, האיזון הנכון בין t גבוה ל-t נמוך תלוי במטרת הניתוח.

שאלה 5

- האם קיבלתם משפטים הגיוניים בשלב 3 סעיף 3?

המשפטים המשוחזרים היו ברוב המקרים לא הגיוניים.
מאחר והמודל סטטיסטי בלבד ובעל sparsity גבוה, הוא לא הצליח רוב הפעמים להשלים את הטוקן הנכון גם אם ערך המקדם של trigram היה גבוה. הקורפוסים שהשתמשנו בהם קטנים ומורכבים יחסית כך שרוב קולוקציות trigram הופיעו מעט יחסית (sparsity) וברוב המקרים לא היה ניתן לחזות את הטוקן החסר מתוך ההקשר.

אם המקדם של trigram היה נמוך ואילו המקדם של unigram גבוה, המודל עשה שימוש קטן בהקשר וחזה בעיקר טוקנים נפוצים (כדוגמת פסיק).
לרוב המודל בחר מילים לא מתאימות (כמו "אני" במקום "הם") והחליף את המילים עם סימני פיסוק (מאחר והם הכי נפוצים).

שאלה 6

- עד כמה ההשלמות של המודל בשלב 3 סעיף 3 קרובות למילים החסרות האמיתיות? פרטו.

בחלק מהמקרים המודל היה קרוב, במיוחד כשהייתה מילה נפוצה המתאימה פחות או יותר למשמעות, כמו "אני" או במקרים של סימני פיסוק.
ברוב המקרים המודל לא הצליח, בעיקר עבור טוקנים פחות נפוצים.
ניסינו לשנות את ערכי המקדמים (של חישוב ההסתברות בשימוש אינטרפולציה לינארית) וכאשר המקדם של trigram היה גבוה יחסית יכולנו לראות השפעה מסויימת על התוצאות (פחות שימוש בהשלמות של פסיק) אך עדיין ברוב המקרים לא היה ניתן לחזות את הטוקן החסר.

סיבה עיקרית לבעיית החיזוי היא בעיית ה-sparsity הקיימת במודל כפי שהוסבר בסעיף הקודם.

שאלה 7

- הסבירו את המשמעות של נוסחת הperplexity שהשתמשתם בה ושל התוצאה שהתקבלה (בשלב 3 סעיף 4).

נוסחת ה-perplexity מתחשבת בכפל ההסתברויות של המילים במשפט. התוצאה מייצגת את ההסתברות ההופכית של מחרוזת בקורפוס הבדיקה כאשר האומדנים נלקחים מקורפוס האימון. כדי לקבל תוצאה אמינה יותר ניתן להשתמש בכמה משפטים ולעשות ממוצע של ה-perplexity. מכיוון שמספר ההסתברויות שיש להכפיל תלוי באורך המשפט (ולכן במשפטים ארוכים יותר סביר שתוצאת הכפל תהיה קטנה יותר) אז ללא נרמול תוצאות ה-Perplexity היו תלויות ישירות באורך המשפט. זוהי משמעות הנירמול באמצעות חזקה של השבר בנוסחא שנלמדה בהרצאה.
בעזרת perplexity ניתן לקבל מידת הערכה למודל. מידה זו קובעת עד כמה מודל שפה שאומן על קורפוס אחד חוזה כהלכה טוקנים במשפטים מטקסט אחר. ניתן לתאר זאת גם בתור מידת ההפתעה של המודל מהמשפטים מהטקסט האחר.
מכיוון שכפל ההסתברויות נמצא במכנה בנוסחא אז הסתברויות גבוהות גורמות לתוצאות נמוכות. מכאן שתוצאות perplexity נמוכות יותר הן טובות יותר.

תוצאה:
כפי שציינו בתחילת הדו"ח ביצענו כמה ניסיונות של מקדמים:

- עבור משקל משמעותי לטריגרם קיבלנו: `133.51`
- עבור משקל שווה לכולם קיבלנו: `42.33`
- עבור משקל משמעותי ליוניגרם קיבלנו: `17.43`

עבור המקדמים שבחרנו בסוף קיבלנו 40.24.

תוצאת הפרפלקסיה הגבוהה כמו שקיבלנו בהתחלה,
מצביעה על קושי בחיזוי המילה וזה מסמן שהמודל לא אופטימלי לחיזוי מילים חסרות.
במיוחד עבור מילים שהן פחות צפויות.
תוצאה זו אכן צפויה מאחר והקורפוס שלנו יחסית קטן.

תוצאת הפרפלקסיה הנמוכה
ככל שנתנו משקל גדול יותר ליוניגרם שיפרנו את תוצאת הפרפלקסיה והמודל בחר רוב הזמן טוקנים נפוצים מאוד.
פתרון זה אינו מעיד על שיפור בהבנה הלשונית של המודל ונותן תוצאות שאינן תלויות בהקשר.
במקרה שלנו **מצביעה על שיפור טכני ולאו דווקא שיפור מבחינת התוכן.**

שאלה 8

- בהמשך לשאלה הקודמת, האם ביצועי המודל על פי מדד זה היו טובים על המשפטים שבחרתם? מדוע לדעתכם?

לא ממש טובים אך הביצועים השתפרו לפי מדד זה ככל שהגדלנו את המקדם של ה-unigram על חשבון המקדמים האחרים (קשור לבעיית ה-sparsity שהוסברה בסעיפים הקודמים). במקרה של מקדם unigram גדול המודל בוחר ב-n-grams עם הסתברויות גבוהות יותר ללא תלות בהקשר. שימוש בהסתברויות גבוהות אלה גורמות למדד ה-perplexity להיות נמוך (כפל ההסתברויות נעשות במכנה ולכן ביחס הפוך ל-perplexity) למרות שהחיזוי לא טוב ונעשה ללא הקשר.

מכיוון שהקורפוס שאנו משתמשים בו קטן ומורכב יחסית אז רוב קולוקציות trigram הופיעו מעט יחסית (sparsity), כלומר הייתה להן הסתברות נמוכה ולכן במקרה שבחרנו להגדיל את המקדם של trigram אז משקל מכפלת ההסתברויות הנמוכות גדל וגרם למדד ה-perplexity לעלות (יחס הפוך להסתברויות).