תרגיל 2

יוסי גואטה 032528267
נטע זינגר 201111648

חלק א

tain	gold	סעיף
127884	11282	1
15986	3171	2
127884	11282	3
37	36	4
1.1349305642	1.07978555661	5

6

- ככל שמדד העמימות עולה כך סיכויי המתייג פוחתים לתת חיזוי טוב
- מסעיף 4 ניתן לראות שכמות המזהים שונה מה שמצביע על מזהים שונים בין קובץ הgold והtrain , דבר אשר עלול לפגוע בביצועי המתייג שכן אנו עלולים לתת משקל יתר לתיוגים אאשר אינם "קיימים"
-

חלק ב

- הגדירו במדויק את סכמת הפרמטרים במודל ? קובץ טקסטואלית בסיומת "baseline" אשר מכיל בכל שורה סגמנט ואת התיוג הנפוץ עבורו בקורפוס. <" " ><tag>

```
ALMLA IN
HWFG VB
HXLJWT NN
HWLLIM JJ
1,776 CD
ATRIM NN
ABZRIM NN
FQWPIM JJ
```

- הגדירו במפורש את נוסחאות המשערכים של הפרמטים במודל ?

$$T(seg) = \underset{tag}{\operatorname{argmax}}\, count(tag_{seg})$$

הגדירו במדוייק את סיבוכיות זמן הריצה של המודל?

עבור מספר המילים num_of_words ומספר התגים num_of_tag, מספר המשפטים num_of_sentances

- אימון:

O(num_of_words + num_of_tag)

- תיוג:

O(num_of_sentances)

דווה מהו הדיוק הכולל עבור תיוג קובץ הבדיקה

macro avg word - 0.830969686226

macro avg sentence - 0.106

חלק ג

מהי פונקציית המטרה של המודל ?

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}}\, P(t_1^n|w_1^n) \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{\text{emission}}\ \overbrace{P(t_i|t_{i-1})}^{\text{transition}}$$

מהן נוסחאות הפרמטרים של המודל ?

$$|V| = size\ of\ corpus$$

$$P(t_i) = \frac{count(t_i)}{|V|}$$

$$P(t_i, t_{i-1}) = \frac{count\big((t_i, t_{i-1})\big) + \delta}{count(t_i) + \delta|V|}$$

$$P(w_i) = [\frac{count\big((w_1, t_1)\big) + \delta}{count(t_1) + \delta|V|}, \dots, \frac{count\big((w_n, t_n)\big) + \delta}{count(t_n) + \delta|V|}]$$

מהן נוסחאות המשערכים של הפרמטרים במודל?

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

מהי סיבוכיות זמן הריצה של האלגוריתם האימון?
עבור T מספר המשפטים וS מספר הסגמנטים

$$O(\binom{|S|}{2} \times T) = |S|^2 \times T$$

מהי סיבוכיות זמן הריצה של אלגוריתם התיוג?
עבור T מספר המשפטים וS מספר הסגמנטים

$$O(T \times |S|^2)$$

מהו הדיוק macro avg עבור קובץ הבדיקה?

- macro avg word - 0.854192519057
- macro avg sentence - 0.112

במקום לציין כ nnp ממשו טכניקת החלקה למילים לא ידועות האם תוצאותיכם השתפרו?

- macro avg word - 0.858890267683
- macro avg sentence - 0.12

כפי שניתן לראות התוצאות השתפרו במעט

ממשו בנוסף טכניקת החלקה למעברים לא ידועים האם תוצאותיכם השתפרו ביחס לסעיף הקודם?

$$P(t_i, t_{i-1}) = \frac{count\big((t_i, t_{i-1})\big) + \delta}{count(t_i) + \delta|V|}$$

כפי שניתן לראות התוצאות השתפרו במעט

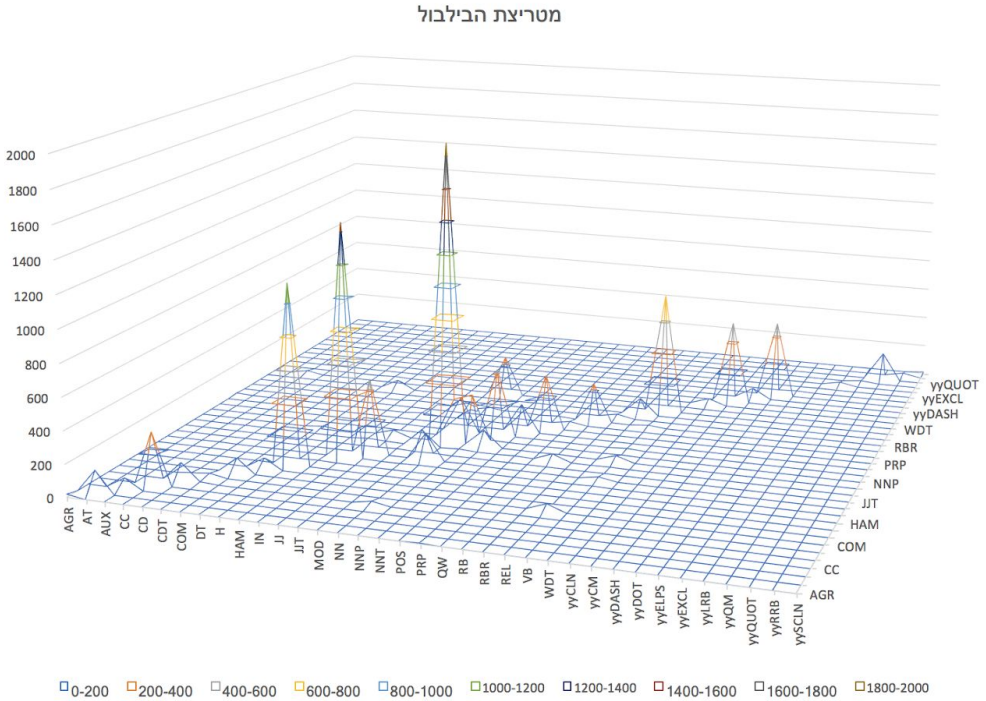
חלק ד

עבור המודל הטוב ביותר שלך חשב את מטריצה הבלבול מהן 3 השגיאות הנפוצות ביותר לפי המטריצה ?
עבור התג NN , התגים שקיבלו את ערך השגיאה המקסימלי הם

תג שצפינו	תג שקיבלנו	כמות
NN	VB	254
NN	NNT	221
NN	NNP	245

מטריצת הבילבול:

	AGR	AT	AUX	CC	CD	CDT	COM	DT	H	HAM	IN	JJ	JJT	MOD	NN	NNP	NNT	POS	PRP	QW	RB	RBR	REL	VB	WDT	yyCLN	yyCM	yyDASH	yyDOT	yyELPS	yyEXCL	yyLRB	yyQM	yyQUOT	yyRRB	yySCLN
AGR	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AT	0	139	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AUX	0	0	67	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
CC	0	0	0	325	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CD	0	0	0	0	112	7	0	0	0	0	0	1	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDT	0	0	0	0	10	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
COM	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	0	0	0	0	0	0	0	0	0	0	0	0
DT	0	0	0	0	0	0	0	67	0	0	0	1	0	0	3	0	0	0	0	0	2	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	1127	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
HAM	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IN	0	0	0	2	0	0	1	1	0	0	1460	3	0	0	4	4	1	2	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
JJ	0	0	0	0	1	0	0	0	0	0	1	473	0	0	170	1	3	0	0	0	1	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
JJT	0	0	0	0	0	0	0	0	0	0	0	1	6	1	4	0	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
MOD	1	0	0	1	0	0	0	0	0	0	9	3	0	104	1	7	9	0	4	0	45	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
NN	0	0	0	0	0	0	0	12	0	0	6	39	0	3	1881	4	105	0	0	0	3	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0
NNP	0	0	0	0	0	1	0	0	1	0	5	2	0	0	245	284	3	0	0	0	2	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
NNT	0	0	0	0	0	0	0	0	0	0	1	3	1	0	221	1	429	0	0	1	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0
POS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	178	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRP	6	0	0	0	1	0	0	0	0	0	0	10	0	0	0	1	0	0	346	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RB	0	0	0	0	0	0	0	0	0	0	8	7	0	34	18	3	7	0	7	0	248	1	0	11	0	0	0	0	0	0	0	0	0	0	0	0
RBR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0
REL	0	0	0	0	0	0	7	0	57	0	0	0	0	0	0	0	0	0	0	0	0	0	107	0	0	0	0	0	0	0	0	0	0	0	0	0
VB	0	0	20	0	0	0	0	0	0	0	0	10	0	2	254	4	3	0	0	0	2	1	0	772	0	0	0	0	0	0	0	0	0	0	0	0
WDT	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
yyCLN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0	0	0	0	0	0
yyCM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	532	0	0	0	0	0	0	0	0	0
yyDASH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66	0	0	0	0	0	0	0	0
yyDOT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	491	0	0	0	0	0	0	0
yyELPS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0
yyEXCL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
yyLRB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0
yyQM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0
yyQUOT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	182	0	0
yyRRB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0
yySCLN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10



תיגו את המשפט הבא ידנית ולאחר מכן תייגו אותו באמצעות המתייג שכתבתם ?

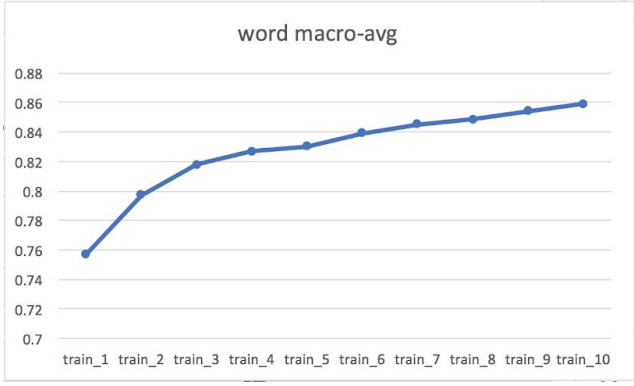
בעל שלה	בפני	ה דלת	את	נעלה	נעל שלה	נעלה	נעלה	אישה	סוג המתייג
adj	conj	N	conj	VB	N	VB	adj	N	אנושי
NN	IN	NN	AT	NN	NN	NN	NN	NN	מכונה

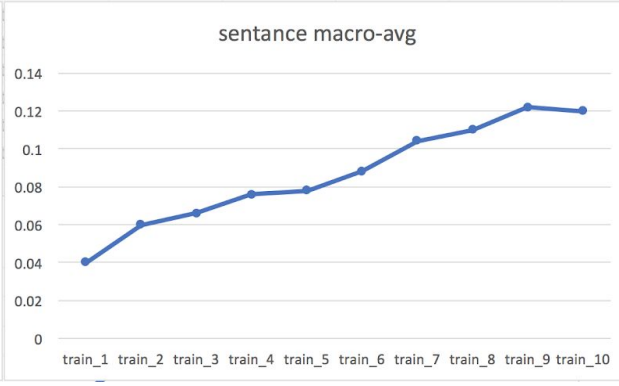
```
AFH NN
NELH    NN
NELH    NN
NELH    NN
yyCM    yyCM
NELH    NN
AT  AT
HDLT    NN
BPNI    IN
BELH    NN
```

ניתן לראות כי המתייג דיי טעה בתיוג השפה ניתן להבין זאת , מדובר במשפט עם עמימות לשונים רבה.

ביצועי המתייג כתלות בגודל קובץ האימון

tain batch	word macro-avg	sentance macro-avg
train_1	0.756957986	0.04
train_2	0.797110441	0.06
train_3	0.817851445	0.066
train_4	0.826803758	0.076
train_5	0.830083319	0.078
train_6	0.839035632	0.088
train_7	0.844974295	0.104
train_8	0.848519766	0.11
train_9	0.854015246	0.122
train_10	0.858890268	0.12





מה יקרה להערכתכם אם יתווספו עוד טקסטים מתוייגים לאימון?
ניתן לראות שהנגזרת שואפת ל-0 ולכן אנו נראה שיש שיפור ממזערי והתכנסות לאזור ה86% אחוזי הצלחה

חלק ו

תנו תיאור סכמטי כולל של הקוד שמימשתם ?
src/Baseline.py - מכיל את כל פונקציות העזר למימוש מתייג הbaseline.(חלק ב) המצמיד לסגמנט את התג השכיח ביותר עבור אותו סגמנט בקורפוס
src/parse_data.py - קובץ המכיל פונקציות עזר לקריאה וכתיבת קבצים
src/const.py - קובץ המכיל את הקבועים במערכת
src/utils.py - קובץ המכיל פונקציית עזר עבור חלק ד הכולל חישוב מטריצות בלבול וחלוקה של הקורפוס
src/viterbi.py - קובץ המכיל את כל פונקציות העזר למימוש מתייג BI GRAM באמצעות אלגוריתם viterbi
src/ex1/*.py - קבצים להרצת חלק א בתרגיל

תארו ממעוף הציפור את התוצאות והמסקנות מהתרגיל ?
ניתן לראות לפי תוצאות התיוג שאף על פי שנעשה שימוש באלגוריתם מתקדם(viterbi המבוסס תכנון דינאמי) ישנו גבול יכולת חיזוי לתיוג חלקי המשפט, על פי המסקנות ככל הנראה נצטרך להוסיף פיצרים נוספים כדי לחדד את יכולת החיזוי.
בנוסף ראינו כי שימוש בטכניקת ההחלקה משפרת את תוצאות המתייג , לו היינו ממשים החלקה מתקדמת יותר היינו מקבלים תוצאות טובות יותר.

פרט נסו לענות על השאלות המחקריות לגבי המתייג בעברית שהוצגו בהקדמה לתרגיל והציעו בכתב דרכים אפשריות לטיוב ושיפור עתידי של המתייג שלכם

- שימוש בהחלקה טובה יותר כגון: good turing
- שימוש בטריאגם כשמבצעים את שהשערוך באלגוריתם הviterbi
- בחינת משקלים שונים עבור emission ו-transition

- שימוש באלגוריתמים שונים כגון רשת ניורנים, לוג לינארי מודל
- מציאת גודל קורפוס אופטומלי אשר יתר תוצאות טובות מצד אחד ומצד שני לא יכביד על משך זמן האימון