תרגיל בית 2 – עיבוד שפה טבעית – חלק רטוב

תיאור המשימה

בתרגיל בית זה תממשו Dependency Parser ללמידת דקדוק תלויות (כפי שנלמד בשבוע 9), ותנתחו את טיב הצלחתכם.

כפי שתראו בהמשך, בתרגיל הבית הנוכחי הופחתו מספר הרכיבים האלגוריתמים אותם אתם צריכים לממש. זאת על מנת שתתעסקו יותר בבניית הייצוג, בהחלטה מהן תכוניות (features) טובות, לפי איזה סף כדי לקצץ וכו'.
מכיוון שכך, הציון על הביצועים בתחרות יהיה גבוה.

סגל הקורס ממליץ להשתמש בשפת python, אולם תוכלו להשתמש בכל שפת תכנות שתחפצו. יש לקחת בחשבון כי נפח התרגיל נקבע על סמך ההנחה שאתם עובדים עם שפה עילית, המאפשרות גמישות בעבודה עם נתונים בצורה נוחה יותר.

נתונים:

הסבר על הקבצים המצורפים –

1. train.labeled – קובץ המכיל 5000 משפטים מתוייגים. עליכם להשתמש בקובץ זה בשלב האימון (הסבר בהמשך)
2. test.labeled – קובץ המכיל 1000 משפטים מתוייגים, בפורמט זהה לפורמט של הקובץ הקודם.
3. comp.unlabeled – קובץ המכיל 1000 משפטים לא מתוייגים.

פורמט קבצי האימון (הדוג' היא המשפט השני בקובץ train):

Token Counter	Token	_	Token POS	_	_	Token Head	Dependency Label	_		_
1	Mr.	_	NNP	_	_	2	NAME	_		_
2	Vinken	_	NNP	_	_	3	VMOD	_		_
3	is	_	VBZ	_	_	0	ROOT	_		_
4	chairman	_	NN	_	_	3	VMOD	_		_
5	of	_	IN	_	_	4	NMOD	_		_
6	Elsevier	_	NNP	_	_	7	NAME	_		_
7	N.V.	_	NNP	_	_	5	PMOD	_		_
8	,	_	,	_	_	7	P	_		_
9	the	_	DT	_	_	12	NMOD	_		_
10	Dutch	_	NNP	_	_	12	NMOD	_		_
11	publishing	_	VBG	_	_	12	NMOD	_		_
12	group	_	NN	_	_	7	APPO	_		_
13	.	_	.	_	_	3	P	_		_

- כל שורה מייצגת מילה, וכוללת 10 עמודות, המופרדות ע"י '\t'
- העמודות היחידות הרלוונטיות למשימה שלנו הן הצבועות באדום – מיקום המילה במשפט, המילה עצמה, חלק הדיבר המתאים עבורה והראש שלה.
- בין כל זוג משפטים יש שורה ריקה
- בקובץ התחרות, בעמודה token head (בכן העמודה Dependency Label) יש קו תחתי ('_')

אימון (Train) :

כאמור את הערכת הפרמטרים תעשו על הקובץ train.labeled. קשתות העץ הפורש ימושקלו על סמך התכוניות הבאות (כפי שראינו במאמר של McDonald) :

Unigram Features

1.) p-word, p-pos
2.) p-word
3.) p-pos
4.) c-word, c-pos
5.) c-word
6.) c-pos

Bigram Features:

7.) p-word, p-pos, c-word, c-pos
8.) p-pos, c-word, c-pos
9.) p-word, c-word, c-pos
10.) p-word, p-pos, c-pos
11.) p-word, p-pos, c-word
12.) p-word, c-word
13.) p-pos, c-pos

כאשר:

p-word: word of parent (head) node.

p-pos: POS of parent (head) node.

c-word: word of child (modifier) node.

c-pos: POS of child (modifier) node.

אתם נדרשים לבנות שני מודלים:

1. מודל *בסיס*, המורכב מ1 עד 13, מלבד 7,9,11,12 (אלו תכוניות עם דלילות גבוהה)
2. מודל *מורכב*, בו תוכלו לבחור כל תכוניות שתרצו. (תתפרעו)

במהלך האימון, תשתמשו בStructured Perceptron כפי שהוצג בהרצאות הוידאו:

$$
\begin{aligned}
&\text{Perceptron}(\mathcal{T} = \{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^{|\mathcal{T}|}) \\
&\quad \mathbf{w}^{(0)} = \mathbf{0};\ k = 0 \\
&\quad \text{for } n : 1..N \\
&\quad\quad \text{for } t : 1..T \\
&\quad\quad\quad \text{Let } \mathbf{y}' = \arg\max_{\mathbf{y}'} \mathbf{w}^{(k)} \cdot \mathbf{f}(\mathbf{x}_t, \mathbf{y}') \\
&\quad\quad\quad \text{if } \mathbf{y}' \neq \mathbf{y}_t \\
&\quad\quad\quad\quad \mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mathbf{f}(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{f}(\mathbf{x}_t, \mathbf{y}') \\
&\quad\quad\quad\quad k = k + 1 \\
&\quad \text{return } \mathbf{w}^{(k)}
\end{aligned}
$$

כאשר t מייצג את האינדקס של משפט מסויים, ו n מייצג את מספר האיטרציה.

בסיכום שתכינו (עוד על כך בהמשך) יש לציין את מספר המאפיינים שמכל סוג (לדוג' 217 מאפיינים מסוג' p-word, p-pos). את המאפיינים שבחרתם למודל המורכב יש להגדיר במפורש.

<u>כל שיפור שהכנסתם למודל (לדוג' קיצוץ של מאפיינים ש"לא הופיעו מספיק") צריך להיות מוסבר היטב</u>, כולל המוטיבציה לבצע אותו. אם השתמשתם בחבילת קוד חיצונית, יש לפרט במפורש היכן ולאיזו מטרה.

בנוסף, יש לפרט כמה זמן לקח לאמן כל מודל (וכן מפרט בסיסי של החומרה עליה הרצתם)

<u>הסקה (Inference):</u>

ההסקה תתבצע ע"י אלגוריתם Chu-Liu-Edmonds הנלמד בתרגול 9.

<u>מבחן (test) :</u>

לכל אחד מן המודלים, יש לבצע הסקה (Inference) על הקובץ test.labeled, ולדווח את תוצאות הדיוק (accuracy) ברמת מילה, כפי שנעשה בהרצאה (ומופיע בתרגול 9). יש לדווח תוצאות על קובץ test עבור N=20,50,80,100 (מספר האיטרציות באלגוריתם ה Perceptron).

התייחסו להבדל בביצועים בין המודלים השונים, והעלו מספר סיבות שעשויות לגרום להבדלים אלו.

בנוסף, אנא ציינו כמה זמן לקח לתייג את הקובץ לפי כל אחד מהמודלים (וכן מפרט בסיסי של החומרה עליה הרצתם את הקוד)

<u>תחרות:</u>

לכל אחד מן המודלים (כאשר מספר האיטרציות הוא לשיקולכם) , יש לבצע הסקה (Inference) על הקובץ comp.unlabeled (אשר אינו כולל תיוגים), ולכתוב את תוצאות התיוג לתוך קובץ חדש בפורמט labeled (כמו קבצי האימון) (שמות הקבצים הרצויים מופיעים בהמשך). לדוג', עבור המשפט:

Token Counter	Token	_	Token POS	_	_	Token Head	Dependency Label	_		_
1	The	_	DT	_	_	_	_	_		_
2	Boy	_	NNP	_	_	_	_	_		_

יש לבצע הסקה, שתיתן לכם את התלויות. בהנחה שהתלויות שמצאתם הן $0 -> 1,\ 1 -> 2$, תכתבו אותן לקובץ ההגשה באופן הבא -

Token Counter	Token	_	Token POS	_	_	Token Head	Dependency Label	_		_
1	The	_	DT	_	_	0	_	_		_
2	Boy	_	NNP	_	_	1	_	_		_

שימו לב שסדר המשפטים (הלא מתוייגים) בקובץ המקורי זהה לסדר המשפטים בקובץ הפלט, שמספר העמודות זהה ושאף אחד מן הערכים מחוץ לשתי העמודות ששיניתם לא נפגע.

יש לתאר במפורש מה עשיתם כדי לקבל את התוצאות שקיבלתם (שינויים שביצעתם בלמידה, בהסקה וכו').

בנוסף, אתם מתבקשים לכתוב: תחזית של אחוז הדיוק שאתם צופים לקבל, וכן להסביר מדוע עשוי להיות הבדל בין הדיוק על קובץ ה test. הסברים חכמים אף עשויים לקבל בונוס.

<u>קוד חיצוני המותר לשימוש:</u>

החבילות הסטנדרטיות בשפה בה בחרתם.

לדוג', בשפת python החבילות בהן מותר להשתמש הן scipy ו numpy בלבד.

בשלב האימון ניתן להשתמש בקוד חיצוני הממש את Chu-Liu-Edmonds. אם בחרתם להשתמש בpython , תוכלו להיעזר בקוד המצורף.

<u>למען הסר ספק - **אסור** להשתמש ב:</u>

1. חבילות הממשות בנייה של תכוניות (features)
2. חבילות הממשות וריאציה של Perceptron
3. חבילות העושות עיבוד על טקסט – ספירת חזרות, uni\tri\bi gram וכו'.

<u>הגשה:</u>

קובץ zip בלבד, בשם HW2-Wet_123456789_987654321.zip (עבור שני סטודנטים שמספרי הזהות שלהם 123456789 ו 987654321). הקובץ הנ"ל יכלול:

1. קובץ הסברים וניתוח תוצאות, הכולל בין היתר:
   a. שמות המחברים ות"ז
   b. הערות על אימון המודל, לכל אחד מן המודלים (לפי הדגשים בסעיף "אימון")
   c. הערות על אלגוריתם ההסקה (לפי הדגשים בסעיף "הסקה")
   d. ניתוח תוצאות על קובץ המבחן (לפי הדגשים בסעיף "מבחן")
   e. סיכום שלכם על המשימה
   f. ניתוח של מה שעשיתם בקובץ התחרות (לפי ההסברים כפי שמפורטים בסעיף "תחרות")

g. הסבר על חלוקת העבודה בין שני חברי הקבוצה – איזה חלק עשה\ביצע\מימש כל אחד

2. קבצי הקוד של התרגיל. על הקוד להיות מתועד וקריא. בנוסף, הקוד צריך מסוגל לרוץ על כל מכונה שהיא. אנא צרו ממשק פשוט להרצת התוכנית המייצרת את קבצי התחרות.

3. קבצי התחרות – על קבצי התוצאות להיות בפורמט labeled (כפי שמפורט בחלק "אימון"). על מנת להימנע מאי נעימות בנוגע לציון, אנא ודאו כי אם שמים '_' בעמודות ששניתם מקבלים בדיוק את הקובץ comp.unlabeled (אותן שורות לפי אותו סדר). חוסר התאמה פירושו ציון 0 בחלק הזה.
על שמות הקבצים להיות – (123456789 הוא ת"ז של אחד הסטודנטים)
   a. comp_m1_123456789.wtag – קובץ labedled שאומן על ידי המודל הבסיסי.
   b. comp_m2_123456789.wtag – קובץ labedled שאומן על ידי המודל המורכב.

הציון על התחרות יתבסס על $\max\{accuracy(comp\_m1), accuracy(comp\_m2)\}$, כלומר נסתכל על התוצאה הטובה יותר מבין שתי ההגשות.

<u>על קבצי התחרות להיות ניתנים לשחזור (Reproducible).</u> הדרישה היא שניתן יהיה לקחת את הקוד שהגשתם ולבנות באמצעותו קובץ זהה לחלוטין לקובץ שהגשתם. במקרים חריגים בודק התרגילים יבדוק את ההתאמה הנ"ל.

בסה"כ קובץ ההגשה צריך להיראות כך:

```
HW1-Wet_123456789_987654321.zip/
    report (.pdf,.docx, etc.)
    Code_Directory/
        ...
    comp_m1_123456789.wtag/
    comp_m2_123456789.wtag/
```

<u>העתקות:</u>

בשל אופי המשימה והמורכבות שלה, קל לבדוק העתקות של קטעי קוד \ קבצים מלאים. למען הסר הספק אנו מדגישים כי אין להעביר קוד בין סטודנטים, בין אם להגשה ובין אם לא. אין לקחת קטעי קוד מוכנים מהאינטרנט, ובכלל אין להסתמך על שום מקור אחר לקוד מלבד פרי יצירכם והחבילות החיצוניות אשר צוינו בסעיף הרלוונטי.