דו"ח תרגיל בית 2 עיבוד שפה טבעית

[redacted]

חלק א'- אימון

דומיין 1: בדומיין הראשון הפעלנו BPE עם מספר שינויים כדי לייעל זמן ריצה.

- כל זוגות הביטים נשמרים בערימת מקסימום כך שהזוג הנפוץ ביותר נמצא בשורש, מה שמאפשר שליפה מהירה בלולאת ה-BPE מבלי לעבור על כל הדאטה בכל פעם. לאחר כל מיזוג, זוגות חדשים שנוצרים נכנסים לערימה עם תדירותם שנמצאת בלולאה פנימית.
- לכל זוג טוקנים נשמרים אינדקסים של המשפטים שבהם הם מופיעים, כדי לייעל את תהליך המיזוג ולמנוע חיפוש מיותר.
- הקידוד מבוסס על מיזוג הדרגתי של זוגות לפי סדר קבוע לפי שיטת Greedy Left-to-Right Merge, וה-decoding משחזר את המחרוזת הבינארית על פי מיפוי הפוך.
- אורך הטוקנים מוגבל מלמעלה באורך זוג המילים הארוך ביותר.

Domain 1

```
Top 5 most common word bigrams:
('in', 'the'): 34889
('going', 'to'): 29182
('for', 'the'): 27436
('on', 'the'): 23322
('to', 'be'): 23173

Top 5 least common word bigrams:
('voice,', 'as'): 1
('Queen', 'will'): 1
('ta', 'so'): 1
('@officialSusanB', 'ta'): 1
('hockey', '3-0.'): 1
```

Domain 2

```
Top 5 most common word bigrams:
('of', 'the'): 15662
('in', 'the'): 12929
('to', 'the'): 6643
('on', 'the'): 5665
('to', 'be'): 5548

Top 5 least common word bigrams:
('Reflections', 'During'): 1
('Kids:', 'Reflections'): 1
('Giving', 'Culture'): 1
('a', 'Wonder'): 1
('10-year-old', 'self'): 1
```

התחלתי בגודל של 5000 טוקנים ואחרי מספר ניסיונות החלטתי ללכת עם גודל של 600 כדי לאזן בין זמן ויעילות הטוקנייזר לבין גודל ממד ה- F1. בתמונה משמאל ניתן לראות את זוגות המילים הכי נפוצים והכי פחות נפוצים.
ביצענו ניקוי עדין לדאטה כדי להמנע מרעש מיותר כמו ישויות HTML וחזרות על אותיות\סימני פיסוק.

דומיין 2: השתמשנו באותו טוקנייזר של דומיין 1, עם 4000 טוקנים. ניסינו מספר נוסף של גדלים אבל 1000 היווה איזון טוב בין יעילות הטוקנייזר למדד ה- F1
פה לא ביצענו שום עריכה מקדימה לדאטה מכיוון שהדאטה סט קטן יותר ונקי יותר.

דומיין 3: בשביל להתמודד עם העובדה שאנחנו לא יודעות על איזה דומיין נבחן, עשינו ניקוי אגרסיבי לסט האימון (צירוף של שני סט האימונים מדומיין 1,2). הורדנו כמה שיותר מילים שאינן מילים באנגלית כמו כתובות אתרים, ישויות HTML, תיוגים של שמות עם @, ומילים שיש בהן חזרה על תו מסויים יותר מ-3 פעמים. כך השגנו דאטה יותר 'נטרלי' בעל יכולות הכללה טובות יותר. אחרי מס' בדיקות החלטו על גודל של 5000 טוקנים.

חלק ב'- מבחן

דומיין 1: הפעלנו את מודל ה-NER מספר פעמים על מנת למצוא ממוצע של מדד ה-F1. קיבלנו את התוצאה הבאה:

יעילות:

```
Vocab size: 600
F1 score average: 0.4106
```

```
Testing encoding speed...
Encoding speed: 174292.86 tokens/second
Total encoding time for test set: 35.7470 seconds
Total tokens used for test set: 6223853

Testing tokenization efficiency...
Tokens per character: 0.5298
```

דומיין 2:

יעילות:

```
Vocab size: 4000
F1 score average: 0.9546
```

```
Testing encoding speed...
Encoding speed: 70682.00 tokens/second
Total encoding time for test set: 20.8563 seconds
Total tokens used for test set: 1478856

Testing tokenization efficiency...
Tokens per character: 0.4096
```

דומיין 3: בדקנו יעילות בעזרת פונקציית test על מנת להגיע ליעילות טובה של 0.4 ובדקנו 1F על קבצי ה-NER מהמודלים הקודמים ללא שינוי. קיבלנו תוצאות דומות לאלו שקיבלנו בדומיינים 1 ו-2 בהתאמה. לכן נשערך שעל קובץ NER חדש שמותאם לדומיין שאנו לא מכירות, נגיע לרמת ביצוע של כ-0.75 (F1)