המכללה האקדמית תל-חי

החוג למדעי המחשב

למידה עמוקה בפייתון

תרגיל 3

Multi-Class Text Classification

מוגש לידי: ד"ר בוריס יאזמיר

מוגש על ידי:

יובל צל-ציון | 305768871 | yuvalzelzion@gmail.com | 052-8514116

גל רייקין | 308188853 | raykingal@gmail.com | 052-4501581

סמסטר ב'

2021

תיאור האלגוריתם:

האלגוריתם המתואר במאמר הוא אלגוריתם לסיווג טקסט למחלקות.
האלגוריתם מקבל פיסת טקסט באורך כלשהו ומסווג את הטקסט אל המחלקה שבה הטקסט עוסק.
אימון המודל מתבצע על מסד נתונים בשם DBpedia. מסד נתונים זה מכיל 6,30,000 פיסות טקסט שמשתייכות ל-14 מחלקות שונות.

תיאור שלבים במימוש:

תחילה כמו תמיד יש לייבא את הספריות הנדרשות, ספריית torchtext ותתי-ספריות שלה.

לאחר מכן מגדירים פרמטרים NGRAMS ו-Batch size, ואז מורידים את מסד הנתונים מ-DBpedia.

יוצרים מחלקה בשם TextSentiment שהיא למעשה תהווה את המודל. מאתחלים את ההיפר-פרמטרים הנחוצים VOCAB_SIZE, EMBED_DIM,
NUN_CLASS, ויוצרים אינסטנט של המודל.

מגדירים פונקציות שיוצרות את קבוצת האימון ואשר מפעילות את האימון ואת הבדיקה של המודל.

מפעילים את האימון של המודל (5 epochs), וכאשר האימון נגמר מריצים בדיקה אשר מראה את ההפסד והדיוק.

לבסוף מייצרים את 14 המחלקות, ואז אפשר להתחיל להריץ בדיקות.

במאמר מבצעים 3 בדיקות של סיווג פיסות טקסט וניתן לראות שהטקסט מסווג ל-3 מחקות שונות NaturalPlace, Plant, Animal.

מסקנות והצעות לשיפור:

לאלגוריתם ביצועים מעולים. אפשר להגיע לתוצאות מדויקות מאוד בלי צורך בעיבוד מקדים גדול בכלל.

ביצענו אימון על מספר גדול מאוד (5,60,000) של מופעים בפחות מ-5 דקות.

הצעה לשיפור שחשבנו עליה בזמן העבודה עם האלגוריתם היא סיווג של טקסט ליותר ממחלקה אחת.

ייתכן כי טקסט מסוים מסווג למחלקה אחת בדיוק חלקי, ולעוד מחלקה נוספת בדיוק חלקי אחר. במקרה זה, אולי אפשר לשפר את האלגוריתם שיתריע כי קיימת מחלקה נוספת שהטקט יכול להשתייך אליה, יראה את כל המחלקות האפשריות ואף את רמות הדיוק שהטקסט יכול להתאים לכל אחת מהן.

תוצאות הרצות:

קישור ל-GitHub עם הקוד.

קישור לסרטון של תוצאות הרצה.

תמונות לדוגמא:

Test the model on the test data set and check the accuracy of the model

```
[11] print('Checking the results of test dataset...')
     test_loss, test_acc = test(test_dataset)
     print(f'\tLoss: {test_loss:.4f}(test)\t|\tAcc: {test_acc * 100:.1f}%(test)')
```

```
Checking the results of test dataset...
        Loss: 0.0000(test)      |       Acc: 97.8%(test)
```

First prediction

```
[13] ex_text_str = "Brekke Church (Norwegian: Brekke kyrkje) is a parish church in Gulen Municipality in S

     print("This is a %s news" %DBpedia_label[predict(ex_text_str, model, vocab, 2)])
```

```
This is a NaturalPlace news
```

Second Prediction

```
[14] ex_text_str2 = "Cerithiella superba is a species of very small sea snail, a marine gastropod mollusk

     print("This text belongs to %s class" %DBpedia_label[predict(ex_text_str2, model, vocab, 2)])
```

```
This text belongs to Plant class
```

Third Prediction

```
[15] ex_text_str3 = "Nithari is a village in the western part of the state of Uttar Pradesh India borderin

     print("This text belongs to %s class" %DBpedia_label[predict(ex_text_str3, model, vocab, 2)])
```

```
This text belongs to Animal class
```