Καμάρης Άγγελος
sdi1900070

README Εργασία 4 άσκηση 1:

Preprocessing:

Χρησιμοποίησα τον κώδικα των προηγούμενων εργασιών για να προεπεξεργαστώ τα δεδομένα . Συγκεκριμένα κάνω κάθε πρόταση lowercase , δεν κρατάω τις λέξεις που περιέχουν κάτω από δύο γράμματα, δεν κρατάω λέξεις που εμφανίζονται εξαιρετικά σπάνια και τις κάνω lemmatize.

Resource problem:

Το google collab παρέχει μόνο 12gb ram για χρήση το οποίο είχε ως αποτέλεσμα να έχω κάποιους περιορισμούς στην δημιουργία του μοντέλου μου . Δηλαδή χρησιμοποιώ μόνο 4500 reviews για training, 90 λέξεις για max length και batches από 45 reviews .
Για να δείτε τις πλήρεις δυνατότητες του μοντέλου (εάν έχετε διαθέσιμη ram) αφαιρέστε την εντολή:
`smaller_df=df.iloc[:4500, :]`
Και ορίστε max length = 150 και στους 2 tokenizers.

Tokenizing phase:

Έχω επιλέξει ο tokenizer να κάνει αυτός το padding ενεργοποιώντας την μεταβλητή pad_to_max length . Επίσης για λόγους πόρων έχω διαλέξει Max length 90 συνεπώς λαμβάνω υπόψη μόνο 90 λέξεις του review(ιδανικά θα επέλεγα το μέγεθος του μεγαλύτερου review).

Fine tuning phase:

Η μέση λύση δεδομένου του resource problem ήταν να γίνεται training και testing για Batch size 48 και 2 epoch και τέλος εμφανίζουμε την αξιολόγηση του μοντέλου για τις μετρικές accuracy precision recall και f1_score.

Results:

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
print(accuracy_score(val_labels, val_preds))
print(precision_score(val_labels, val_preds))
print(recall_score(val_labels, val_preds))
print(f1_score(val_labels, val_preds))
```

```
0.825
0.8954489544895449
0.7599164926931107
0.8221343873517787
```

Παραδοχές:

Δεδομένου ότι το μοντέλο ήθελε αρκετές ώρες να τρέξει και δεν ήταν δυνατόν να δοκιμαστούν μεγάλα νούμερα σε κάποιους παραμέτρους οι πειραματισμοί ήταν περιορισμένοι. Παρόλα αυτά παρακάτω παραθέτω κάποιους πειραματισμούς που έγιναν πάνω στο μοντέλο μου

Max length:

Γιά max length=20 είχαμε λίγο παραπάνω από 50% accuracy καθώς δεν ήταν εφικτό να αξιολογείται κάποιο review με βάση μόνο 20 λέξεις οπότε η απόφαση του μοντέλου ήταν ουσιαστικά τυχαία.
Γιά max length=40 είχαμε περίπου 60% accuracy καθώς κάποιες κριτικές με λιγότερες λέξεις αξιολογόντουσαν σωστά
Γιά max length=70 είχαμε περίπου 75% accuracy καθώς για τις περισσότερες κριτικές 70 λέξεις ήταν αρκετές για την σωστή αξιολόγηση τους.
Γιά max length=90 έχουμε 82% accuracy καθώς καλύπτουμε όλο και περισσότερες κριτικές
Ιδανικά θα χρειαζόμασταν πόρους max lengh=μέγεθος μεγαλύτερου review στην βάση το οποίο θα μας έδινε περίπου
87-88% accuracy

Epochs
Οι μόνες χρονικά λογικές επιλογές ήταν 1 εως 5 epoch . Κατέληξα στα 2 γιατί αν και θέλει σχεδόν διπλάσιο χρόνο σε σχέση με το 1 βελτιώνει αρκετά την αποτελεσματικότητα του μοντέλου. Φυσικά αν κάποιος είχε αρκετό χρόνο θα μπορούσε να βάλει και 4 ή 5 αλλά εκεί υπάρχει ρίσκο για overfitting.

Batch size
Το batch size δεν άλλαζε ιδιαίτερα τα αποτελέσματα όταν κυμαίνεται ανάμεσα στο 32-64. Αποφάσισα να χρησιμοποιήσω το 48 που είναι ενδιάμεσα στο διάστημα και είναι και σχεδόν το μέγιστο που επιτρέπουν οι πόροι του συστήματος. Δοκίμασα και μικρότερα όπως 2 ,5 ,10 αλλά δεν είχαν θετική επιρροή στα αποτελέσματα μπορούσαν όμως να αυξήσουν τον χρόνο.