Αϊδίνης Γιώργος el16031
Τσιλιβής Θοδωρής el16032

Επεξεργασία Φωνής και Φυσικής Γλώσσας

Προπαρασκευή 3ου Εργαστηρίου

1) Προεπεξεργασία Δεδομένων

Ζητούμενο 1:

Για το συγκεκριμένο στάδιο έγινε χρήση της βιβλιοθήκης SciKitLearn, και συγκεκριμένα της μεθόδου LabelEncoder(). Με τον τρόπο αυτό μετατρέψαμε τις ετικέτες (labels) των δεδομένων από strings (positive, neutral, negative) σε αριθμούς κλάσεων (0,1,2).

1)Για το αρχείο **MR** τυπώνουμε τις πρώτες 10 ετικέτες, και τις αντιστοιχίες τους σε αριθμούς:

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
QUESTION 1 - convert data labels from strings to integers

1    positive
1    positive
1    positive
1    positive
1    positive
1    positive
1    positive
1    positive
1    positive
1    positive
```

2)Για το αρχείο **Semeval2017A** τυπώνουμε τις πρώτες 10 ετικέτες, και τις αντιστοιχίες τους σε αριθμούς:

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
QUESTION 1 - convert data labels from strings to integers

1    neutral
2    positive
1    neutral
2    positive
2    positive
2    positive
1    neutral
2    positive
0    negative
1    neutral
```

Ζητούμενο 2:

Ύστερα, θα κάνουμε το tokenization, που είναι ουσιαστικά η μετατροπή των δεδομένων εκπαίδευσης και testing από την μορφή τους (προτάσεις) σε μια λίστα λέξεων που θα μπορούμε πιο εύκολα να διαχειριστούμε.

Το πρόβλημα λοιπόν είναι το πως θα “σπάσουμε” τις προτάσεις σε λέξεις, χωρίς να χάνεται η τυπική ορθογραφία των λέξεων, ενώ παράλληλα να μπορεί να διατηρηθεί το νόημα του κειμένου, και όλα αυτά διατηρώντας το μικρότερο δυνατό αριθμό δεδομένων.

Και για τα δύο datasets χρησιμοποιήσαμε την συνάρτηση TweetTokenizer() της βιβλιοθήκης NLTK, που αποτελεί ταυτόχρονα αποδοτικό αλλά και αξιόπιστο τρόπο για να “σπάσουμε” τις προτάσεις που έχουμε.

Συγκεκριμένα, επιλέξαμε συνειδητά την μέθοδο αυτή, γιατί είναι μια μέθοδος που λαμβάνει υπόψιν το γεγονός ότι οι προτάσεις προέρχονται από το ίντερνετ, που από μόνο του δεν εγγυάται για την ορθογραφία και χρήση λέξεων του αγγλικού λεξιλογίου μόνο (με άλλα λόγια, μπορεί οι λέξεις στις οποίες θα σπάσει μια πρόταση να μην είναι σωστά γραμμένες, ή να μην αποτελούν καν λέξεις του αγγλικού λεξιλογίου, αλλά του “ιντερνετικού”).

Η συνάρτηση αυτή αφαιρεί επιπλέον γράμματα (“waaaaaay” -> “waaay”), αφαιρεί twitter handles (“@DonaldTrump”), και διατηρεί τυχόν “λέξεις” που ίσως να είναι σημαντικές (“:D” μπορεί να είναι κατά την άποψή μας σημαντική, καθώς δηλώνει χαρά, που έχει να κάνει με τον σκοπό μας, και δεν είναι λέξη του αγγλικού λεξιλογίου).

Το αποτέλεσμα της συνάρτησης αυτής είναι μια λίστα λέξεων.

3)Για το αρχείο **MR** τυπώνουμε τις πρώτες 10 tokenized προτάσεις:

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
QUESTION 2 - Define our PyTorch-based Dataset

sentence no. 1 :
the rock is destined to be the 21st century's new " conan " and that he's going to make a splash even greater than arnold schwarzenegger , jean-claud van damme or steven segal .

this is returned by the class as:

['the', 'rock', 'is', 'destined', 'to', 'be', 'the', '21st', "century's", 'new', '"', 'conan', '"', 'and', 'that', "he's", 'going', 'to', 'make', 'a', 'splash', 'even', 'greater', 'than', 'arnold', 'schwarzenegger', ',', 'jean-claud', 'van', 'damme', 'or', 'steven', 'segal', '.']

sentence no. 2 :
the gorgeously elaborate continuation of " the lord of the rings " trilogy is so huge that a column of words cannot adequately describe co-writer / director peter jackson's expanded vision of j . r . r . tolkien's middle-earth .

this is returned by the class as:

['the', 'gorgeously', 'elaborate', 'continuation', 'of', '"', 'the', 'lord', 'of', 'the', 'rings', '"', 'trilogy', 'is', 'so', 'huge', 'that', 'a', 'column', 'of', 'words', 'cannot', 'adequately', 'describe', 'co-writer', '/', 'director', 'peter', "jackson's", 'expanded', 'vision', 'of', 'j', '.', 'r', '.', 'r', '.', "tolkien's", 'middle-earth', '.']

sentence no. 3 :
effective but too-tepid biopic

this is returned by the class as:

['effective', 'but', 'too-tepid', 'biopic']

sentence no. 4 :
if you sometimes like to go to the movies to have fun , wasabi is a good place to start .
```

this is returned by the class as:

['if', 'you', 'sometimes', 'like', 'to', 'go', 'to', 'the', 'movies', 'to', 'have', 'fun', ',', 'wasabi', 'is', 'a', 'good', 'place', 'to', 'start', '.']

sentence no. 5 :
emerges as something rare , an issue movie that's so honest and keenly observed that it doesn't feel like one .

this is returned by the class as:

['emerges', 'as', 'something', 'rare', ',', 'an', 'issue', 'movie', "that's", 'so', 'honest', 'and', 'keenly', 'observed', 'that', 'it', "doesn't", 'feel', 'like', 'one', '.']

sentence no. 6 :
the film provides some great insight into the neurotic mindset of all comics - - even those who have reached the absolute top of the game .

this is returned by the class as:

['the', 'film', 'provides', 'some', 'great', 'insight', 'into', 'the', 'neurotic', 'mindset', 'of', 'all', 'comics', '-', '-', 'even', 'those', 'who', 'have', 'reached', 'the', 'absolute', 'top', 'of', 'the', 'game', '.']

sentence no. 7 :
offers that rare combination of entertainment and education .

this is returned by the class as:

['offers', 'that', 'rare', 'combination', 'of', 'entertainment', 'and', 'education', '.']

sentence no. 8 :
perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions .

this is returned by the class as:

['perhaps', 'no', 'picture', 'ever', 'made', 'has', 'more', 'literally', 'showed', 'that', 'the', 'road', 'to', 'hell', 'is', 'paved', 'with', 'good', 'intentions', '.']

sentence no. 9 :
steers turns in a snappy screenplay that curls at the edges ; it's so clever you want to hate it . but he somehow pulls it off .

this is returned by the class as:

['steers', 'turns', 'in', 'a', 'snappy', 'screenplay', 'that', 'curls', 'at', 'the', 'edges', ';', "it's", 'so', 'clever', 'you', 'want', 'to', 'hate', 'it', '.', 'but', 'he', 'somehow', 'pulls', 'it', 'off', '.']

sentence no. 10 :
take care of my cat offers a refreshingly different slice of asian cinema .

this is returned by the class as:

['take', 'care', 'of', 'my', 'cat', 'offers', 'a', 'refreshingly', 'different', 'slice', 'of', 'asian', 'cinema', '.']

4)Για το αρχείο **Semeval2017A** τυπώνουμε τις πρώτες 10 tokenized προτάσεις:

~~~~~~

QUESTION 2 - Define our PyTorch-based Dataset

sentence no. 1 :
05 Beat it - Michael Jackson - Thriller (25th Anniversary Edition) [HD] http://t.co/A4K2B86PBv

this is returned by the class as:
~~~~~~

['05', 'Beat', 'it', '-', 'Michael', 'Jackson', '-', 'Thriller', '(', '25th', 'Anniversary', 'Edition', ')', '[', 'HD', ']', 'http://t.co/A4K2B86PBv']

sentence no. 2 :
Jay Z joins Instagram with nostalgic tribute to Michael Jackson : Jay Z apparently joined Instagram on Saturday and .. http://t.co/Qj9I4eCvXy

this is returned by the class as:

['Jay', 'Z', 'joins', 'Instagram', 'with', 'nostalgic', 'tribute', 'to', 'Michael', 'Jackson', ':', 'Jay', 'Z', 'apparently', 'joined', 'Instagram', 'on', 'Saturday', 'and', '..', 'http://t.co/Qj9I4eCvXy']

sentence no. 3 :
Michael Jackson : Bad 25th Anniversary Edition (Picture Vinyl): This unique picture disc vinyl includes the original 1 http://t.co/fKXhToAAuW

this is returned by the class as:

['Michael', 'Jackson', ':', 'Bad', '25th', 'Anniversary', 'Edition', '(', 'Picture', 'Vinyl', '):', 'This', 'unique', 'picture', 'disc', 'vinyl', 'includes', 'the', 'original', '1', 'http://t.co/fKXhToAAuW']

sentence no. 4 :
I liked a video http://t.co/AaR3pjp2PI One Direction singing " Man in the Mirror " by Michael Jackson in Atlanta , GA [June 26 ,

this is returned by the class as:

['I', 'liked', 'a', 'video', 'http://t.co/AaR3pjp2PI', 'One', 'Direction', 'singing', '"', 'Man', 'in', 'the', 'Mirror', '"', 'by', 'Michael', 'Jackson', 'in', 'Atlanta', ',', 'GA', '[', 'June', '26', ',']

sentence no. 5 :
18th anniv of Princess Diana's death . I still want to believe she is living on a private island away from the public . With Michael Jackson .

this is returned by the class as:

['18th', 'anniv', 'of', 'Princess', "Diana's", 'death', '.', 'I', 'still', 'want', 'to', 'believe', 'she', 'is', 'living', 'on', 'a', 'private', 'island', 'away', 'from', 'the', 'public', '.', 'With', 'Michael', 'Jackson', '.']

sentence no. 6 :
The 1st time I heard Michael Jackson sing was in Honolulu , Hawaii @ a restaurant on radio . It was A . B . C . I was 13 . I loved it !

this is returned by the class as:

['The', '1st', 'time', 'I', 'heard', 'Michael', 'Jackson', 'sing', 'was', 'in', 'Honolulu', ',', 'Hawaii', '@', 'a', 'restaurant', 'on', 'radio', '.', 'It', 'was', 'A', '.', 'B', '.', 'C', '.', 'I', 'was', '13', '.', 'I', 'loved', 'it', '!']

sentence no. 7 :
' Michael Jackson ' appeared on Saturday 29 at the 9th place in the Top 20 of Miami's Trends : http://t.co/dXN2FWgUhb #trndnl

this is returned by the class as:

['"', 'Michael', 'Jackson', '"', 'appeared', 'on', 'Saturday', '29', 'at', 'the', '9th', 'place', 'in', 'the', 'Top', '20', 'of', "Miami's", 'Trends', ':', 'http://t.co/dXN2FWgUhb', '#trndnl']

sentence no. 8 :
Are you old enough to remember Michael Jackson attending the Grammys with Brooke Shields and Webster sat on his lap during the show ?

this is returned by the class as:

['Are', 'you', 'old', 'enough', 'to', 'remember', 'Michael', 'Jackson', 'attending', 'the', 'Grammys', 'with', 'Brooke', 'Shields', 'and', 'Webster', 'sat', 'on', 'his', 'lap', 'during', 'the', 'show', '?']

sentence no. 9 :
do u enjoy his 2nd rate Michael Jackson bit ? Honest ques . Like the can't feel face song but god it's so obvious they want MJ 2.0

this is returned by the class as:

['do', 'u', 'enjoy', 'his', '2nd', 'rate', 'Michael', 'Jackson', 'bit', '?', 'Honest', 'ques', '.', 'Like', 'the', "can't", 'feel', 'face', 'song', 'but', 'god', "it's", 'so', 'obvious', 'they', 'want', 'MJ', '2.0']

sentence no. 10 :
The Weeknd is the closest thing we may get to Michael Jackson for a long time ... especially since he damn near mimics everything

this is returned by the class as:

['The', 'Weeknd', 'is', 'the', 'closest', 'thing', 'we', 'may', 'get', 'to', 'Michael', 'Jackson', 'for', 'a', 'long', 'time', '...', 'especially', 'since', 'he', 'damn', 'near', 'mimics', 'everything']

Ζητούμενο 3:

Έπειτα, θα πρέπει να υλοποιήσουμε την κωδικοποίηση του κάθε token (κάθε λέξης της "σπασμένης" πρότασης), δηλαδή να αντιστοιχίζουμε το κάθε token με την λέξη στο embedding που έχουμε (pretrained embeddings).

Η δουλειά αυτή δεν είναι ιδιαίτερα απαιτητική, πρέπει όμως να γίνουν κάποιες επιλογές, συγκεκριμένα για το αν θα επιλέξουμε να "πετάξουμε" κάποιες προτάσεις που είναι ιδιαίτερα μεγάλες.

Επιλέξαμε να μην αγνοήσουμε τους outliers (δηλαδή προτάσεις που έχουν πολύ μεγάλο μέγεθος), καθώς οι προτάσεις που έχουμε δεν είναι πολλές στον αριθμό, και ακόμα δεν είδαμε κάποια πρόταση στα datasets που να έχει υπερβολικό μέγεθος. Συνεπώς, κάθε φορά, το μέγιστο μέγεθος που επιλέγουμε είναι το μέγιστο μέγεθος οποιασδήποτε πρότασης στο dataset.

Για αυτόν τον λόγο, θα πρέπει να κάνουμε zero padding, δηλαδή πλέον όλες οι προτάσεις θα έχουν το μέγεθος της μεγαλύτερης, απλά οι προτάσεις που έχουν μικρότερο μέγεθος θα έχουν τόσα μηδενικά στο τέλος τους, όσα χρειάζονται για να φτάσουν το μέγεθος της μεγαλύτερης πρότασης.

Ακόμα, θα πρέπει να κωδικοποιήσουμε κάθε λέξη, που σημαίνει να "κοιτάζουμε" σε ποιόν αριθμό αντιστοιχεί η κάθε λέξη από το pretrained embedding που έχουμε, και να αντικαθιστούμε την λέξη με τον αριθμό αυτόν. Αν μια λέξη δεν είναι εντός του λεξιλογίου, τότε την αντιστοιχίζουμε με τον αριθμό της λέξης '<unk>' (αυτή η διαδικασία λέγεται Out Of Vocabulary word handling).

5)Για το αρχείο **MR** τυπώνουμε τις πρώτες 5 κωδικοποιημένες προτάσεις:

~~~~~~

QUESTION 3 - Calculating the max length of training strings, to configure zero-padding

The average length of each sentiment is: 21
The maximum length of each sentiment is: 59

Sentence embedding 1
sentence: ['the', 'rock', 'is', 'destined', 'to', 'be', 'the', '21st', "century's", 'new', '"', 'conan', '"', 'and', 'that', "he's", 'going', 'to', 'make', 'a', 'splash', 'even', 'greater', 'than', 'arnold', 'schwarzenegger', ',', 'jean-claud', 'van', 'damme', 'or', 'steven', 'segal', '.'] , target: positive
sentence's word embedding: [1   1138   15 10454   5   31   1   5034 400001   51
     9 18513   9   6   13 400001   223   5   160   8
  16807   152   1414   74   5819   6681   2 400001   1462 43708
~~~~~~

47 4412 26985 3 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] , label: 1

Sentence embedding 2
sentence: ['the', 'gorgeously', 'elaborate', 'continuation', 'of', '"', 'the', 'lord', 'of', 'the', 'rings', '"', 'trilogy', 'is', 'so', 'huge', 'that', 'a', 'column', 'of', 'words', 'cannot', 'adequately', 'describe', 'co-writer', '/', 'director', 'peter', "jackson's", 'expanded', 'vision', 'of', 'j', '.', 'r', '.', 'r', '.', "tolkien's", 'middle-earth', '.'] , target: positive
sentence's word embedding: [1 78616 5135 10117 4 9 1 2371 4 1
6820 9 12305 15 101 1325 13 8 3236 4
1375 1120 12424 4467 47768 275 370 1295 400001 2853
3139 4 6892 3 1912 3 1912 3 400001 55754
3 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0] , label: 1

Sentence embedding 3
sentence: ['effective', 'but', 'too-tepid', 'biopic'] , target: positive
sentence's word embedding: [2038 35 400001 34277 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 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0] , label: 1

Sentence embedding 4
sentence: ['if', 'you', 'sometimes', 'like', 'to', 'go', 'to', 'the', 'movies', 'to', 'have', 'fun', ',', 'wasabi', 'is', 'a', 'good', 'place', 'to', 'start', '.'] , target: positive
sentence's word embedding: [84 82 1072 118 5 243 5 1 2460 5 34 2906
2 66408 15 8 220 242 5 466 3 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 0 0 0] , label: 1

Sentence embedding 5
sentence: ['emerges', 'as', 'something', 'rare', ',', 'an', 'issue', 'movie', "that's", 'so', 'honest', 'and', 'keenly', 'observed', 'that', 'it', "doesn't", 'feel', 'like', 'one', '.'] , target: positive
sentence's word embedding: [12398 20 646 2349 2 30 496 1006 400001 101
6082 6 23499 4583 13 21 400001 999 118 49
3 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 0 0 0] , label: 1

6)Για το αρχείο **Semeval2017A** τυπώνουμε τις πρώτες 5 κωδικοποιημένες προτάσεις:

~~~~~~

QUESTION 3 - Calculating the max length of training strings, to configure zero-padding

The average length of each sentiment is: 22
The maximum length of each sentiment is: 61

Sentence embedding 1
sentence: ['05', 'Beat', 'it', '-', 'Michael', 'Jackson', '-', 'Thriller', '(', '25th', 'Anniversary', 'Edition', ')', '[', 'HD', ']', 'http://t.co/A4K2B86PBv'] , target: neutral
sentence's word embedding: [17261 400001 21 12 400001 400001 12 400001 24 8962
400001 400001 25 2824 400001 5281 400001 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 0 0 0 0 0 0 0 0
0] , label: 1

Sentence embedding 2
sentence: ['Jay', 'Z', 'joins', 'Instagram', 'with', 'nostalgic', 'tribute', 'to', 'Michael', 'Jackson', ':', 'Jay', 'Z', 'apparently', 'joined', 'Instagram', 'on', 'Saturday', 'and', '..', 'http://t.co/Qj9I4eCvXy'] , target: positive
sentence's word embedding: [400001 400001 7698 400001 18 20557 5079 5 400001 400001
46 400001 400001 1897 1031 400001 14 400001 6 400001
400001 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 0 0 0 0
0] , label: 2

Sentence embedding 3
sentence: ['Michael', 'Jackson', ':', 'Bad', '25th', 'Anniversary', 'Edition', '(', 'Picture', 'Vinyl', '):', 'This', 'unique', 'picture', 'disc', 'vinyl', 'includes', 'the', 'original', '1', 'http://t.co/fKXhToAAuW'] , target: neutral
sentence's word embedding: [400001 400001 46 400001 8962 400001 400001 24 400001 400001
400001 400001 3007 1836 5977 11193 1013 1 930 177
400001 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 0 0 0 0
0] , label: 1

Sentence embedding 4
sentence: ['I', 'liked', 'a', 'video', 'http://t.co/AaR3pjp2PI', 'One', 'Direction', 'singing', '"', 'Man', 'in', 'the', 'Mirror', '"', 'by', 'Michael', 'Jackson', 'in', 'Atlanta', ',', 'GA', '[', 'June', '26', ','] , target: positive
sentence's word embedding: [400001 5573 8 975 400001 400001 400001 4100 9 400001
7 1 400001 9 22 400001 400001 7 400001 2
400001 2824 400001 1077 2 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
0] , label: 2

Sentence embedding 5
sentence: ['18th', 'anniv', 'of', 'Princess', "Diana's", 'death', '.', 'I', 'still', 'want', 'to', 'believe', 'she', 'is', 'living', 'on', 'a', 'private', 'island', 'away', 'from', 'the', 'public', '.', 'With', 'Michael', 'Jackson', '.'] , target: positive
sentence's word embedding: [4014 400001 4 400001 400001 337 3 400001 150 304
5 734 68 15 757 14 8 673 584 421
26 1 199 3 400001 400001 400001 3 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] , label: 2

2) Μοντέλο:

Ζητούμενο 4:

Αρχικά σε αυτό το στάδιο θα πρέπει να δημιουργήσουμε το μοντέλο που θα εκπαιδεύσουμε. Για να λειτουργήσει σωστά το νευρωνικό δίκτυο, θα πρέπει να έχει ένα αρχικό layer που θα αντιστοιχίζει τις λέξεις εισόδου σε μια διάσταση που μπορεί το embedding που διαθέτουμε να καταλάβει.

Το layer αυτό θα λαμβάνει ως όρισμα προτάσεις του κειμένου, που είναι σε μορφή DataLoader(), και θα τους κάνει προβολή σε συνεχή χώρο, έτσι ώστε οι κοντινά σημασιολογικές λέξεις να βρίσκονται σχετικά κοντά.

Στην συγκεκριμένη περίπτωση, τα βάρη αυτού του layer θα πρέπει να είναι αρχικοποιημένα (και παγωμένα) στις τιμές που έχουμε λάβει από το αρχείο glove.6B.50d.txt.

Θα πρέπει λοιπόν να ορίσουμε μέγεθος αυτού του layer να είναι όσο και το μέγεθος των διανυσμάτων που βρίσκονται μέσα σε αυτό το αρχείο, δηλαδή 50, και τον αριθμό των διανυσμάτων να είναι όσα είναι τα διανύσματα που βρίσκονται μέσα σε αυτό το αρχείο.

7) Χρησιμοποιούμε προεκπαιδευμένα διανύσματα για δύο λόγους. Ο πρώτος λόγος είναι για να ελαφρύνουμε τον φόρτο της δουλειάς της εκπαίδευσης του δικτύου, καθώς το να έχουμε ήδη έτοιμες τις τιμές σημαίνει ότι ο αλγόριθμος εκπαίδευσης έχει να βελτιστοποιήσει ένα λιγότερο layer, και άρα ταχύτερη εκπαίδευση. Ο δεύτερος λόγος είναι ότι αυτά τα embeddings είναι εξαχθέντα από μια πολύ μεγαλύτερη συλλογή δεδομένων, και αφενός μπορεί να μας βοηθήσουν περισσότερο στην επίτευξη της καλύτερης δυνατής ακρίβειας, και αφετέρου οι τυχαίες τιμές θα καθυστερούσαν να λάβουν βέλτιστες τιμές (αν τις λάμβαναν κιόλας, αφού μπορεί η εκπαίδευση να έφτανε κάποιο πλάτωμα).

8) Ο λόγος που κρατάμε τις τιμές των βαρών του embedding layer παγωμένες είναι διττός. Αρχικά, το κάνουμε αυτό για να αποφύγουμε αρκετούς υπολογισμούς που δεν είναι απαραίτητοι, καθώς το να επαναϋπολογίζουμε τις τιμές (προσπαθώντας να τις βελτιώσουμε) έχει αρκετά μεγάλο “χτύπημα” στην επίδοση του δικτύου, αφού είναι μια διαδικασία αρκετά χρονοβόρα. Ακόμα, είναι πιθανό να κάνουμε overfit αφήνωντας τις τιμές αυτές να εξελιχθούν, που μπορεί να μας δώσει πολύ καλό training accuracy, αλλά χειρότερο testing accuracy, αφού το μοντέλο θα μάθει πολύ καλά τα δεδομένα που του δώσαμε για να εκπαιδευθεί.

Ζητούμενο 5:

Τώρα πλέον πρέπει να διαμορφώσουμε το υπόλοιπο network, δηλαδή το output layer, την μη γραμμική συνάρτηση μεταφοράς, έτσι ώστε να γίνεται αναπαράσταση στον χώρο εξόδου του προβλήματος (2 η 3ων διαστάσεων στην περίπτωση μας).

Επιλέξαμε ως συνάρτηση μεταφοράς την ReLU(), ενώ δοκιμάσαμε και την Tanh(), αλλά η ReLU είχε λίγο καλύτερα αποτελέσματα (αλλά ήταν αρκετά κοντά).

9) Στο ερώτημα του γιατί βάζουμε μια μη γραμμική συνάρτηση ως συνάρτηση μεταφοράς στο τελευταίο layer, η απάντηση βρίσκεται στην δυνατότητά τους να αναπαραστήσουν συναρτήσεις που δεν είναι γραμμικές. Αυτό είναι ιδιαίτερα χρήσιμο, καθώς πολλές φορές τα δεδομένα που έχουμε δεν έχουν αυστηρά γραμμική συμπεριφορά, αλλά παρουσιάζουν κάποια κυρτότητα στον χώρο που βρίσκονται.

Επιπλέον, δεν θα είχε κάποια διαφορά αν χρησιμοποιούσαμε 2 ή περισσότερους γραμμικούς μετασχηματισμούς, καθώς δεν έχει σημασία πόσα γραμμικά layers θα χρησιμοποιήσουμε, αφού το

δίκτυο μπορεί να προσεγγίσει μόνο γραμμικές συμπεριφορές. Πρέπει λοιπόν κάπως να παρουσιάσουμε μια μη γραμμικότητα στο μοντέλο.

Ζητούμενο 6:

Τώρα θα υλοποιήσουμε την διαδικασία του forward pass, δηλαδή την μεταφορά των δεδομένων σειριακά στα layers του δικτύου.

Ουσιαστικά, αφού εισαχθούν οι λέξεις στο embedding layer, ώστε να γίνει η αντιστοίχιση κάθε όρου σε ένα διάνυσμα, θα εξέλθουν οι πληροφορίες με διαστάσεις (batch size, max_length, embedding_dimension).

Στην συνέχεια, θα πρέπει να δημιουργηθεί μια ενιαία, απλή αναπαράσταση για όλες τις λέξεις, που θα την υπολογίζουμε βρίσκοντας τον μέσο όρο των embeddings σε μια πρόταση. Ο μέσος όρος αυτός είναι ακριβώς αυτός που αναφέρουμε ως “κέντρο βάρους”, καθώς το κέντρο βάρους είναι ουσιαστικά ο μέσος όρος των διανυσμάτων (στον πολυδιάστατο χώρο).

Έπειτα, πρέπει να εφαρμοστεί η συνάρτηση που εισάγει την μη γραμμικότητα, και τέλος να γίνει η τελική αναπαράσταση στον χώρο των κλάσεων.

10)Ουσιαστικά, η όλη χρησιμότητα των embeddings είναι να απεικονίσουμε αυτό που οι άνθρωποι καταλαβαίνουμε διαισθητικά, δηλαδή την εγγύτητα εννοιών. Οι άνθρωποι καταλαβαίνουμε εύκολα από τα συμφραζόμενα ότι οι λέξεις “κατοικίδιο” και “γάτα” είναι σχετικά κοντά (καθώς η γάτα αποτελεί υποσύνολο του κατοικιδίου), αλλά ο υπολογιστής δεν έχει κάποιο τρόπο για να το καταλάβει. Για να λύσουμε το πρόβλημα αυτό, χρησιμοποιούμε τα embeddings, τα οποία δημιουργούνται κοιτάζοντας από πολλές προτάσεις τα συμφραζόμενα και βρίσκουν ομοιότητες μεταξύ λέξεων, έτσι ώστε να μπορούμε να βρούμε πότε δύο λέξεις μπορούν να χρησιμοποιηθούν η μια στην θέση της άλλης. Αυτό, αν και δεν είναι κατανόηση της εγγύτητας των εννοιών, είναι αρκετά κοντά, και στην πράξη δουλεύει. Στην διαδικασία αυτή, λέξεις που μπορούν να χρησιμοποιηθούν η μια στην θέση της άλλης (άρα και είναι “κοντά” κατά προσέγγιση εννοιολογικά) καταλήγουν να έχουν περίπου ίδιες τιμές στα διανύσματα στο ολικό embedding. Άρα στον ολικό χώρο, τα διανύσματα τους θα βρίσκονται κοντά. Συνεπώς, ο μέσος όρος των διανυσμάτων μιας πρότασης, απεικονίζει ένα διάνυσμα, που (θεωρητικά) αντιπροσωπεύει την κεντρική έννοια της πρότασης. Αυτό στην πράξη δεν είναι πάντα αληθές βέβαια.

11) Μια προφανής αδυναμία της προσέγγισης αυτής είναι όταν λέξεις έχουν την ίδια ορθογραφία, αλλά διαφορετικό νόημα, όπως book (όπως βιβλίο), και book (κλείνω κάποιο τραπέζι/ εισιτήρια κτλ). Ακόμα, η σειρά εμφάνισης των λέξεων πολλές φορές παίζει σημαντικό νόημα στην μεταφορά ενός μηνύματος, και άρα πρέπει να δίνεται αρκετή σημασία, που όμως με την προσέγγιση αυτήν δεν γίνεται. Τέλος, τα σημεία στίξης είναι και αυτά σημαντικά στην κατανόηση του κειμένου, και δεν πρέπει να αφαιρούνται από την πρόταση (στην δικιά μας περίπτωση δεν αφαιρούνται, αλλά γενικά είναι μια τακτική που ακολουθείται).

Ζητούμενο 7:

Για να φορτώσουμε τα δεδομένα στο μοντέλο χρησιμοποιήσαμε την κλάση DataLoader(), έτσι ώστε να σπάσουμε το dataset σε μικρότερα κομάτια, τα batches.

12) Όσο μεγαλύτερο είναι το batch size, τόσο μικρότερο πρέπει να είναι το learning rate προκειμένου να εκπαιδεύσουμε με ακρίβεια. Το μέγεθος του batch επηρεάζει την εκπαίδευση του μοντέλου με δύο τρόπους. Όταν έχουμε μικρό μέγεθος, αλλάζει πολύ γρήγορα το βάρος σε κάθε εποχή, αλλά αν μικρύνει πολύ, τότε έχουμε αρκετό θόρυβο στο σύστημα. Αν αντίθετα έχουμε μεγάλο μέγεθος, τόσο λιγότερο θόρυβο έχουμε στο σύστημα, αλλά είναι πιθανό να συγκλίνει το

νευρωνικό σε κάποιο τοπικό ελάχιστο και όχι ολικό, και βέβαια το σύστημα εκπαιδεύεται πιο αργά. Πρέπει να υπάρχει λοιπόν μια ισορροπία στο μέγεθος των batches.

13) Το ανακάτεμα των δεδομένων (shuffling) χρειάζεται για δύο λόγους. Ο προφανής λόγος είναι για να αποφύγουμε να μάθει μια προφανή κατανομή των δεδομένων (πχ όλα τα θετικά πρώτα και μετά όλα τα αρνητικά), δηλαδή δεν θα μάθει σαν παράγοντα την σειρά με την οποία δέχτηκε τα δεδομένα. Ο δεύτερος, και ίσως λιγότερο προφανής λόγος είναι για να αποφύγουμε να μείνει το νευρωνικό σε κάποιο τοπικό ελάχιστο, που ίσως να βρεθεί λόγω της σειράς των δεδομένων. Η τυχαιότητα με την οποία εξασφαλίζονται τα δεδομένα στο νευρωνικό είναι καίριας σημασίας στον στοχαστικό χαρακτήρα με τον οποίο δουλεύει το Stochastic Gradient Descent. Σαν μέθοδος, το shuffling είναι σημαντική και αναγκαία, παρόλα αυτά πρέπει να χρησιμοποιείται όταν οι καταστάσεις το απαιτούν, και όχι σε κάθε βήμα (και με προσοχή στο shuffling data – labels).

Ζητούμενο 8:

Στο κεφάλαιο αυτό προσπαθούμε να θέσουμε τις παραμέτρους για την βελτιστοποίηση του μοντέλου. Συγκεκριμένα, χρησιμοποιήσαμε ως κριτήριο για το dataset MR το BCEWithLogitsLoss, ενώ για το Semeval2017A το CrossEntropyLoss. Γενικά βελτιστοποιήσαμε κάθε παράμετρο στο model.parameter, που βέβαια έχει requires_grad = True, έτσι ώστε να επιτρέπεται ο υπολογισμός στο backpropagation. Για μέθοδο optimizer επιλέξαμε την μέθοδο Adam.

Ζητούμενο 9:

Όπως ζητήθηκε από την άσκηση, υλοποιήθηκαν οι μέθοδοι train_dataset() και eval_dataset(). Η train_dataset() ουσιαστικά κάνει ότι θα περίμενε κάποιος, εκτελεί μια εποχή, δηλαδή περνάει τα δεδομένα από το μοντέλο, βρίσκει το error, και μέσω του backpropagation ενημερώνει τα κατάλληλα βάρη του δικτύου. Αντίστοιχα, η eval_dataset() καλείται στο τέλος μιας εποχής για να εξετάσει την ακρίβεια του μοντέλου στα δεδομένα του test set.

Ζητούμενο 10:

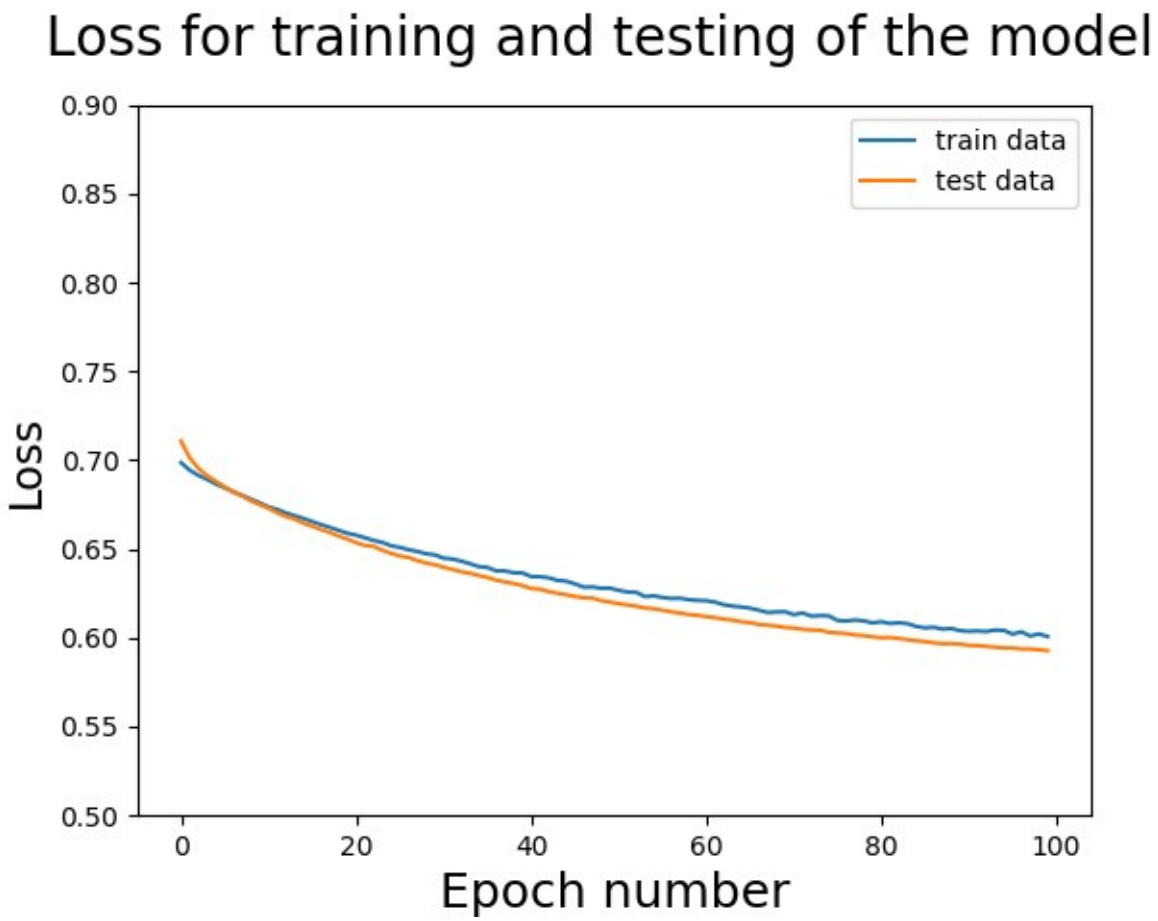
Αρχικά αποφασίσαμε οι εποχές να είναι 100, και όχι οι default 50, για να δούμε (σε καμία περίπτωση όχι αρκετά) λίγο περισσότερο την εξέλιξη του μοντέλου μας με τις παραμέτρους που επιλέξαμε. Αποφασίσαμε να πάρουμε σαν pretrained word embeddings το αρχείο glove.6B.50d.txt, καθώς είδαμε ότι ο υπολογιστής που τρέχαμε τους κώδικες “ζοριζόταν” με το dataset Semeval2017A (είναι παλιό laptop χωρίς gpu acceleration). Παρακάτω παραθέτουμε τα μετρικά για κάθε dataset, καθώς και γραφήματα για τα losses για κάθε dataset.

Για το dataset MR:

[==========] ...Epoch 100, Loss: 0.6383

Metric analysis for the model in this Epoch:

F1 metric for training: 0.6880996226005435
Accuracy for training: 0.6881
Recall metric for training: 0.6881009104084064
F1 metric for testing: 0.6767253575764214
Accuracy for testing: 0.676737160120846
Recall metric for testing: 0.6767629741202246



Για το dataset Semeval2017A:
[==========] ...Epoch 100, Loss: 1.0589

Metric analysis for the model in this Epoch:

F1 metric for training: 0.47673528136342364
Accuracy for training: 0.5660278394190035
Recall metric for training: 0.5511877032818387
F1 metric for testing: 0.5043253181625365
Accuracy for testing: 0.5389938130901987
Recall metric for testing: 0.5352462488368742