**Α1. Προεπεξεργασία και Προετοιμασία δεδομένων:**

**α)** Η διαδικασία της προεπεξεργασίας των δεδομένων μπορεί να βρεθεί στο αρχείο **preprocessing.py**. Αρχικά, φορτώνουμε το σύνολο δεδομένων σε ένα pandas dataframe, χρησιμοποιώντας ως delimiter το tab. Έπειτα, πραγματοποιούμε κάποιους ελέγχους στο σύνολο δεδομένων, εκμεταλλευόμενοι τις δυνατότητες της βιβλιοθήκης pandas. Ελέγχουμε για τυχόν μη-υπάρχουσες ή μη ορισμένες τιμές οπουδήποτε μέσα στο dataset. Αυτός ο έλεγχος γίνεται προκειμένου να εξασφαλίσουμε την ακεραιότητα του συνόλου δεδομένων, καθώς απούσες ή μη καλά ορισμένες τιμές στα γεγονότα, μπορεί να οδηγήσουν σε μεγάλα σφάλματα κατά την εκπαίδευση του νευρωνικού μας δικτύου. Έπειτα, με την εντολή **df.describe()**, μπορούμε να δούμε χρήσιμες στατιστικές ιδιότητες των στηλών του dataframe που περιέχουν αριθμητικά δεδομένα, όπως την μέση τιμή της εκάστοτε στήλης, την τυπική της απόκλιση, την ελάχιστη και την μέγιστη τιμή της κ.α. Έτσι μπορούμε να καταλάβουμε περισσότερα πράγματα για το σύνολο δεδομένων με το οποίο εργαζόμαστε. Τα αποτελέσματα της εντολής φαίνονται παρακάτω:

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, ασπρόμαυρο

Περιγραφή που δημιουργήθηκε αυτόματα

Σειρά έχει η υλοποίηση του **Bag-of-Words (BoW)** Μοντέλου. Το μοντέλο αυτό είναι ιδανικό για την διαχείριση text-based συνόλου δεδομένων, δημιουργεί ένα λεξικό από όλες τις διαφορετικές λέξεις που υπάρχουν σε αυτό. Εδώ οι λέξεις του λεξικού, τα λεγόμενα **tokens**, είναι **unigrams**, δηλαδή κάθε token είναι μία μοναδική λέξη. Έτσι στην περίπτωση μας, το μοντέλο δεν καταμετρά την θέση, την γραμματική ή την δομή των προτάσεων που απαρτίζουν το σύνολο δεδομένων μας, παρά καταμετρά μόνο συχνότητες εμφανίσεων λέξεων, χρησιμοποιώντας μία συγκεκριμένη τεχνική που ονομάζεται **tf-idf vectorization (Term frequency — Inverse document frequency)**. H τεχνική αυτή μας επιτρέπει να ταυτοποιήσουμε τις πιο σημαντικές λέξεις στο κείμενο. Αποτελείται από δύο ποσότητες:

1. **Term Frequency (Tf):** Μας δίνει το πλήθος των εμφανίσεων ενός token που περιλαμβάνεται σε ένα document και υπολογίζεται ως εξής:

**Εικόνα που περιέχει κείμενο, γραμματοσειρά, γραμμή, λευκό

Περιγραφή που δημιουργήθηκε αυτόματα**

1. **Inverse Document Frequency (Idf):** Μας δίνει την μοναδικότητα ενός token, δεδομένου του συνόλου των documents. Η κεντρική ιδέα με το συγκεκριμένο index είναι ότι μία λέξη η οποία εμφανίζεται αρκετά συχνά στα περισσότερα documents του συνόλου δεδομένων, δεν αποδίδει ιδιαίτερη πληροφορία σχετικά με το εκάστοτε document. Η συγκεκριμένη ποσότητα υπολογίζεται ως εξής:

**Εικόνα που περιέχει γραμματοσειρά, κείμενο, γραμμή, λευκό

Περιγραφή που δημιουργήθηκε αυτόματα**

Η τελική βαθμολογία tf-idf προκύπτει πολλαπλασιάζοντας τις δύο παραπάνω ποσότητες, προσδίδοντας μία τιμή σε κάθε token του λεξικού, ευνοώντας τα πιο σπάνια tokens.

Για την υλοποίηση του συγκεκριμένου μοντέλου χρησιμοποιήθηκε η βιβλιοθήκη **scikit-learn** και συγκεκριμένα η κλάση **TfidfVectorizer** της. Προτιμήθηκε να χρησιμοποιηθεί λίστα με συγκεκριμένα stopwords, προκειμένου να φιλτραριστούν από το λεξικό λέξεις οι οποίες είναι συχνά επαναλαμβανόμενες, ωστόσο δεν προσδίδουν κάποια χρήσιμη πληροφορία για το εκάστοτε κείμενο, όπως οι συνδετικές λέξεις. Συγκεκριμένα με τις παρακάτω εντολές, δημιουργούμε ένα αντικείμενο της κλάσης TfidfVectorizer και βρίσκουμε το tfidf μητρώο του λεξικού από τα κείμενα το συνόλου δεδομένων:



Το αντικείμενο της κλάσης TfidfVectorizer, παίρνει ως ορίσματα:

1. **Stop\_words = stopwords:** Μία λίστα από stopwords, την οποία πήραμε από την βιβλιοθήκη **nltk.** Η λίστα αυτή έχει 256 διαφορετικές λέξεις, πολλές από τις οποίες είναι αρχαίες ελληνικές, πράγμα το οποίο είναι αρκετά χρήσιμο για το συγκεκριμένο σύνολο δεδομένων.
2. **Max\_features = 1000:** Καθορίζουμε τον αριθμό των unigrams που θέλουμε να έχει το λεξικό μας. Οι λέξεις καθορίζονται με βάση το tf (term frequency), δηλαδή την συχνότητα εμφάνισης τους σε κάθε κείμενο. Έτσι, έχοντας αποβάλλει λέξεις άχρηστες για την ανάλυση που θέλουμε να πραγματοποιήσουμε, μας μένουν λέξεις οι οποίες είναι συχνά εμφανιζόμενες, ωστόσο έχουν χρήσιμη πληροφορία σχετικά με το εκάστοτε κείμενο στο οποίο βρίσκονται. Το ποσό καθορίζεται από την εκφώνηση γι’ αρχή.

Έτσι προκύπτει ένα csr() μητρώο, του οποίου οι γραμμές είναι τα διάφορα documents του συνόλου δεδομένων και οι στήλες οι λέξεις του λεξικού. Στην i-οστή γραμμή αντιστοιχεί Η i-οστή επιγραφή του συνόλου δεδομένων, ενώ στην i-οστή στήλη αντιστοιχεί η i-ος΄τη λέξη του λεξικού. Τιμή στο αντίστοιχο κελί υπάρχει όταν η λέξη σε εκείνη την στήλη βρίσκεται στην αντίστοιχη επιγραφή και η τιμή της αντιστοιχεί στην tf-idf τιμή της λέξης αυτής. Πράγματι, κρίνοντας από το σχήμα του μητρώου, βλέπουμε ότι τα αποτελέσματά μας είναι ακριβή:



Έπειτα, βρίσκουμε τις idf τιμές καθενός από τις λέξεις που υπάρχουν στο λεξικό μας, καθώς αυτές οι τιμές μας δίνουν μία εικόνα για την υπόθεση που κάναμε παραπάνω. Δεν θέλουμε ούτε η τιμή να είναι υψηλή, συνεπώς η λέξη να συναντάται σπάνια στο σύνολο των επιγραφών, ούτε πάρα πολύ χαμηλή, πράγμα που θα δήλωνε ότι συναντάται σε όλα τα κείμενα, οπότε δεν θα μας έδινε καμία χρήσιμη πληροφορία για την εκάστοτε περιγραφή. Παρακάτω, φαίνονται οι πρώτες και οι τελευταίες 10 τιμές του πίνακα idf\_values:

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματα

Έτσι καταφέρνουμε να βρούμε λέξεις με σπουδαιότητα, χωρίς να είναι ούτε πολύ σπάνιες, ούτε πάρα πολύ συχνές.

**β)** Παρ’ ότι οι τιμές των vectors βρίσκονται σχετικά κοντά μεταξύ τους, επιλέγουμε να κάνουμε κανονικοποίηση στο εύρος [0,1], τόσο στα vectors, δηλαδή τα δεδομένα εισόδου του μετέπειτα νευρωνικού που θα αναπτύξουμε, όσο και στα δεδομένα εξόδου, δηλαδή στις τιμές των στηλών date\_min, date\_max του dataframe που δημιουργήσαμε στην αρχή το τμήματος της προεπεξεργασίας. Η κανονικοποίηση επιλέγεται να εφαρμοστεί για πολλούς διαφορετικούς λόγους, κάποιοι εκ των οποίων είναι:

1. Αποφεύγεται η κυριαρχία κάποιων χαρακτηριστικών (τιμών), έναντι κάποιων άλλων, επηρεάζοντας έτσι την διαδικασία της μάθησης και οδηγώντας σε biased αποτελέσματα.
2. Οι μικρές τιμές των δεδομένων εισόδου, επιταχύνουν την σύγκλιση της διαδικασίας εισόδου, οδηγώντας στην κατασκευή ενός σταθερού νευρωνικού δικτύου.
3. Τα δεδομένα οπτικοποιούνται ευκολότερα, καθώς οι μεταξύ τους αποστάσεις μειώνονται.

Στην παρούσα εργασία, χρησιμοποιείται, όπως προειπώθηκε, min-max scaling στο πεδίο [0,1] τόσο για τα δεδομένα εισόδου, όσο και εξόδου. Η φόρμουλα η οποία χρησιμοποιείται είναι η εξής:

Εικόνα που περιέχει κείμενο, γραμματοσειρά, λευκό, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Χρησιμοποιείται ακόμα μία φορά η βιβλιοθήκη **sklearn,** από την οποία χρησιμοποιούμε την κλάση **MinMaxScaler,** για την κανονικοποίηση των δεδομένων μας. Αυτή η διαδικασία γίνεται με τις παρακάτω εντολές:

Εικόνα που περιέχει κείμενο, γραμματοσειρά, στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα

,όπου texts τα δεδομένα εισόδου του νευρωνικού και dates τα δεδομένα εξόδου του. Τα αποτελέσματα που προκύπτουν μετά την κανονικοποίηση (ενδεικτικά), φαίνονται παρακάτω:

**Δεδομένα εισόδου:**

Εικόνα που περιέχει στιγμιότυπο οθόνης, πληκτρολόγιο, υπολογιστής

Περιγραφή που δημιουργήθηκε αυτόματα

,το οποίο όπως εξηγήθηκε και παραπάνω είναι ένα sparse μητρώο με τις κανονικοποιημένες tf-idf τιμές των λέξεων που επιλέχθηκαν.

**Δεδομένα εξόδου:**

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά

Περιγραφή που δημιουργήθηκε αυτόματα

, όπου παρατηρούμε ότι οι τιμές είναι καταρχάς όλες θετικές και πολύ πιο κοντά μεταξύ τους από τις αρχικές, που βρίσκονταν στο εύρος [-720, 1453].

**γ)** Το cross validation, και συγκεκριμένα το **k-fold cross validation**, είναι μία τεχνική που αξιολογεί την συμπεριφορά ενός νευρωνικού δικτύου σε ένα ανεξάρτητο σύνολο δεδομένων. Βασική ιδέα της συγκεκριμένης τεχνικής, είναι ο διαχωρισμός του συνόλου δεδομένων σε k πολλαπλά διακριτά υποσύνολα (folds), ένα εκ των οποίων ορίζεται ως σύνολο επικύρωσης (validation set). Η διαδικασία αυτή επαναλαμβάνεται k φορές, με διαφορετικό σύνολο επικύρωσης κάθε φορά και η τελική απόδοση του μοντέλου εκτιμάται ως ο μέσος όρος των αποδόσεων των k επαναλήψεων. Η συγκεκριμένη τεχνική χρησιμοποιείται ως μία ικανοποιητική λύση στα προβλήματα της υπερπροσαρμογής όσο και της υποπροσαρμογής. Χρησιμοποιείται για ακόμα μία φορά η βιβλιοθήκη **sci-kit learn** Και συγκεκριμένα η **KFold** κλάση της, καθώς ζητάται να γίνει **5-fold Cross Validation**. Η αρχικοποίηση και παραμετροποίηση φαίνεται από τη παρακάτω εντολή:



Χρησιμοποιώντας την εντολή kf.split(X) δημιουργούμε τα 5 folds και έπειτα για κάθε ένα fold παίρνουμε τους δείκτες που χωρίζουν τα δεδομένα σε train και test sets. Συνεπώς στο τέλος δημιουργούμε μία λίστα από λεξικά, όπου το κάθε λεξικό περιέχει πληροφορία για το ποιο fold είναι, τα training και test sets των εισόδων και των targets, καθώς και τις μέσες τιμές κάθε γραμμής των y\_train και y\_test (των δύο ημερομηνιών). Έτσι έχουμε έτοιμα τα δεδομένα που θα εισάγουμε στο νευρωνικό για κάθε fold.

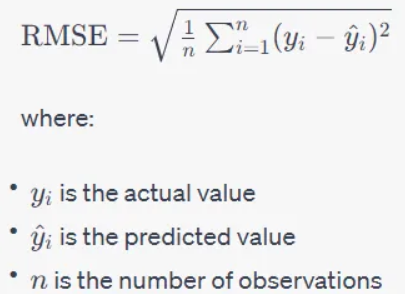
**Α2. Επιλογή Αρχιτεκτονικής**

Για την κατασκευή του νευρωνικού δικτύου χρησιμοποιήθηκε η βιβλιοθήκη **tensorflow** και συγκεκριμένα το API της, το **Keras**. Θεωρήθηκε η καλύτερη επιλογή, δεδομένου ότι προσφέρει ικανοποιητική λειτουργικότητα, χωρίς να χρειάζεται πολύπλοκη υλοποίηση. Υλοποιήθηκαν custom συναρτήσεις για:

1. Την παραμετροποίηση του νευρωνικού δικτύου, όπως τον καθορισμό του αριθμού των κρυφών επιπέδων, του αριθμού των κόμβων ανά κρυφό επίπεδο και του ρυθμού μάθησης.
2. Την μέτρηση σφάλματος μεταξύ των προβλέψεων και των targets στην έξοδο του νευρωνικού δικτύου, με την τεχνική του RMSE (Root Mean Square Error).
3. Την εκπαίδευση του νευρωνικού δικτύου, με τον καθορισμό του αριθμού των batches και των εποχών της εκπαίδευσης.
4. Τον έλεγχο της μάθησης του νευρωνικού δικτύου, μέσω της ερμηνείας του σφάλματος του σε ξεχωριστό κομμάτι των δεδομένων.

Χρησιμοποιείται o optimizer adam για την ελαχιστοποίηση της συνάρτησης σφάλματος, καθώς είναι αρκετά καλύτερος από άλλε εναλλακτικές, λόγω της γρήγορης σύγκλισής του ιδιαίτερα σε υψηλής διαστατικότητας διανυσματικούς χώρους. Ο Adam συνδυάζει τις τεχνικές του Momentum και του RMSprop(Root Mean Square Propagation), προκειμένου να υπολογίσει προσαρμοστικά το ρυθμό μάθησης για κάθε μία παράμετρο βασιζόμενος στη κλίμακα των προηγούμενων gradients.

**a)** Θεωρήσαμε την μέση τιμή των δύο ημερομηνιών ως το target value της εξόδου τοθ νευρωνικού δικτύου, και υλοποιήσαμε την συνάρτηση **custom\_RMSE,** η οποία υπολογίζει το root mean square error των δύο διανυσμάτων, του y\_pred, δηλαδή της εξόδου του νευρωνικού και του y\_true δηλαδή των target values, σύμφωνα με τον τύπο:



Η υλοποίηση της συνάρτησης φαίνεται παρακάτω:

Εικόνα που περιέχει κείμενο, γραμματοσειρά, στιγμιότυπο οθόνης, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα