# Πρόγραμμα forecast

## Σκοπός

Το πρόγραμμα αυτό έχει σκοπό να δημιουργήσει ένα μοντέλο πρόβλεψης για τις τιμές μετοχών. Το μοντέλο λαμβάνει σαν είσοδο μία χρονοσειρά N τιμών της μετοχής κατά το παρελθόν και δημιουργεί ένα μοντέλο πρόβλεψης βασισμένο σε κατάλληλη δομή νευρωνικού δικτύου.

## Διαδικασία

### Έλεγχος Παραμέτρων

Σε πρώτη φάση γίνεται ο έλεγχος των παραμέτρων εισόδου. Αυτές θα πρέπει να είναι ένα αρχείο χρονοσειρών σε μορφή csv και ο αριθμός των μετοχών που θα εξεταστούν. Στο csv αρχείο, η πρώτη στήλη αφορά την ονομασία της μετοχής και οι υπόλοιπες (έστω d) τις τιμές των μετοχών (από την παλαιότερη στην νεότερη). Αν οι παράμετροι δοθούν σωστά το πρόγραμμα συνεχίζει στην επόμενη λειτουργία, διαφορετικά τερματίζει με κατάλληλο μήνυμα σφάλματος.

### Προετοιμασία Δεδομένων

Τα δεδομένα που λαμβάνονται από το αρχείο εισόδου μετασχηματίζονται σε dataframe. Στη συνέχεια καθορίζεται το βάθος ημερών στο οποίο είναι επιθυμητό να βασιστεί η πρόβλεψη (μεταβλητή look\_back – default τιμή 4). Από το dataframe επιλέγονται οι πρώτες n γραμμές που καθορίζονται από την αντίστοιχη μεταβλητή εισόδου. Τα δεδομένα στη συνέχεια κανονικοποιούνται (διαιρείται κάθε τιμή των δεδομένων με την μέγιστη τιμή τους). Με την συνάρτηση create\_dataset προσαρμόζεται η μορφή κάθε χρονοσειράς για την εκπαίδευση μοντέλου με βάση look\_back δείγματα.

### Κατασκευή μοντέλου πρόβλεψης

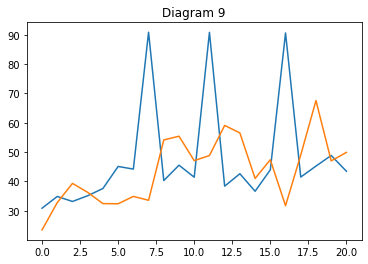
Η κατασκευή του μοντέλου βασίστηκε σε LSTM νευρωνικό δίκτυο με τρία επίπεδα LSTM. Η είσοδος του προσαρμόστηκε στα δεδομένα των μετοχών με βάθος χρόνου look back. Από τα πειράματα που προηγήθηκαν προέκυψε το συμπέρασμα ότι τρία LSTM επίπεδα μπορούν να υποστηρίξουν ένα μοντέλο πρόβλεψης με ικανοποιητική απόδοση. Προστέθηκε και ένα Dropout επίπεδο για την αποφυγή του overfitting.

### Εκπαίδευση του μοντέλου πρόβλεψης

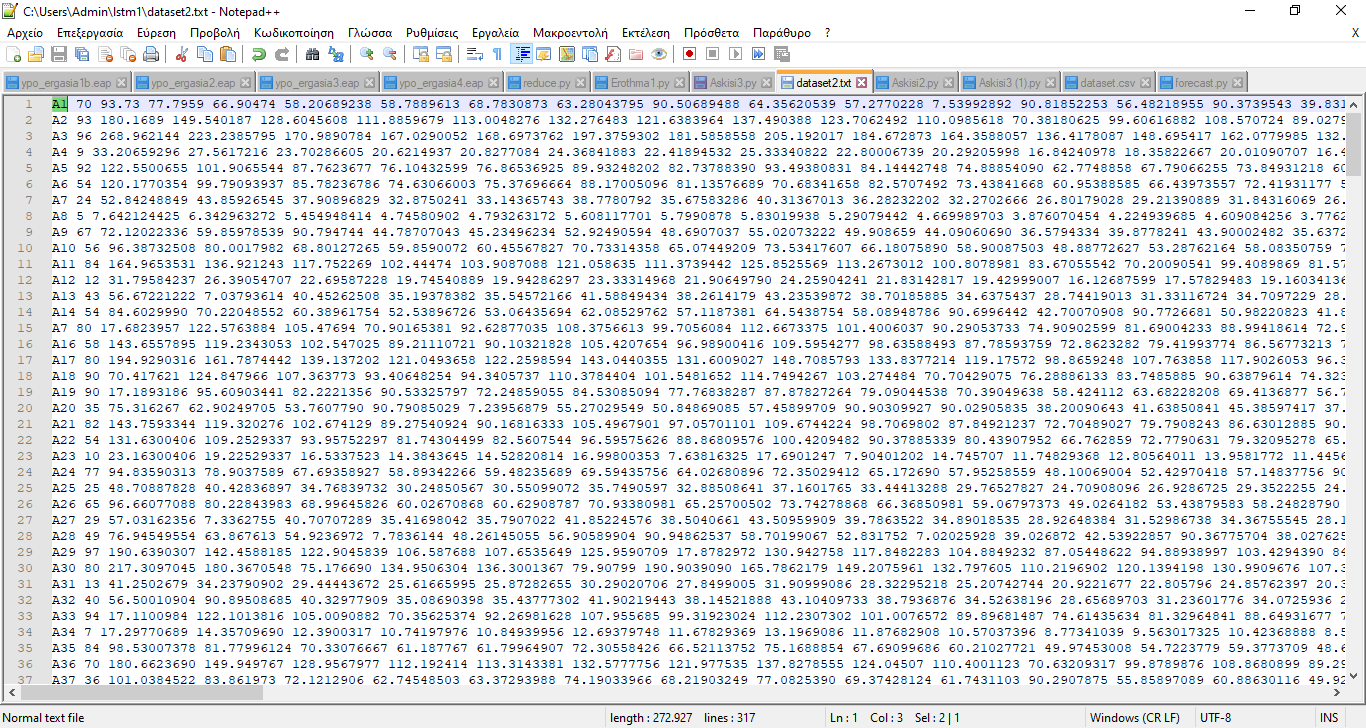
Για την εκπαίδευση του μοντέλου πρόβλεψης εξετάστηκαν διαφορετικές τιμές για epochs, batches και μέγεθος επιπέδων LSTM. Οι τιμές που έδωσαν ικανοποιητικές τιμές για το μέσο απόλυτο τετράγωνο του σφάλματος ήταν epochs=100, batch=3 και μέγεθος επιπέδου LSTM 50. P

### Πρόβλεψη

Στο τελευταίο μέρος του προγράμματος γίνεται πρόβλεψη για μία σειρά από χρονοσειρές. Για κάθε πρόβλεψη παρέχεται ένα διάγραμμα για τις τιμές που προβλέφθηκαν και τις πραγματικές. Ένα παράδειγμα τέτοιου διαγράμματος φαίνεται παρακάτω.



Δείγμα του περιεχομένου του αρχείου που χρησιμοποιήθηκε για τις δοκιμές φαίνεται παρακάτω.



# Πρόγραμμα forecast

## Σκοπός

Το πρόγραμμα αυτό έχει σκοπό στις χρονοσειρές των τιμών των μετοχών να ανακαλύπτει τις χρονικές στιγμές των εξάρσεων σε σχέση με την «ομαλή ροή» τους.

## Διαδικασία

### Έλεγχος Παραμέτρων

Σε πρώτη φάση γίνεται ο έλεγχος των παραμέτρων εισόδου. Αυτές θα πρέπει να είναι ένα αρχείο χρονοσειρών σε μορφή csv και ο αριθμός των μετοχών που θα εξεταστούν. Στο csv αρχείο, η πρώτη στήλη αφορά την ονομασία της μετοχής και οι υπόλοιπες (έστω d) τις τιμές των μετοχών (από την παλαιότερη στην νεότερη). Αν οι παράμετροι δοθούν σωστά το πρόγραμμα συνεχίζει στην επόμενη λειτουργία, διαφορετικά τερματίζει με κατάλληλο μήνυμα σφάλματος.

### Προετοιμασία Δεδομένων

Τα δεδομένα που λαμβάνονται από το αρχείο εισόδου μετασχηματίζονται σε dataframe. Στη συνέχεια καθορίζεται το βάθος ημερών στο οποίο είναι επιθυμητό να βασιστεί η πρόβλεψη (μεταβλητή look\_back – default τιμή 4). Από το dataframe επιλέγονται οι πρώτες n γραμμές που καθορίζονται από την αντίστοιχη μεταβλητή εισόδου. Τα δεδομένα στη συνέχεια κανονικοποιούνται (διαιρείται κάθε τιμή των δεδομένων με την μέγιστη τιμή τους). Με την συνάρτηση create\_dataset προσαρμόζεται η μορφή κάθε χρονοσειράς για την εκπαίδευση μοντέλου με βάση look\_back δείγματα.

### Κατασκευή του μοντέλου

Το μοντέλο που κατασκευάζεται είναι ένας αυτοκωδικοποιητής ο οποίος αποτελείται από μία σειρά από LSTM επίπεδα με ένα επίπεδο Dropout για την αποφυγή του overfitting. Από τα πειράματα προέκυψε ότι οι καταλληλότερες τιμές των παραμέτρων για την καλύτερη αποδοτικότητα του μοντέλου είναι: Επίπεδα LSTM:2, batch:20, epochs:100.

### Εκπαίδευση του μοντέλου πρόβλεψης

Το μοντέλο πρόβλεψης εκπαιδεύεται με το train μέρος των δεδομένων με ετικέτες το ίδιο.

Πρόβλεψη

Χρησιμοποιούνται ένα πλήθος χρονοσειρών (δίνονται από την γραμμή εντολών). Παράγονται οι τιμές από τον μοντέλο πρόβλεψης (δηλαδή η χρονοσειρά μεταβάλλεται στην «ομαλή» χρονοσειρά που αντιστοιχεί σε αυτή). Δημιουργείται ένα διάγραμμα το οποίο απεικονίζει την πραγματική και τη «ομαλή» χρονοσειρά. Στο διάγραμμα αυτό επισημαίνονται τα σημεία όπου η διαφορά των δύο χρονοσειρών ξεπερνάει το κατώφλι που τοποθετείται στην γραμμή εισόδου. Ένα παράδειγμα τέτοιου διαγράμματος φαίνεται παρακάτω.

Το αρχείο εισόδου που χρησιμοποιήθηκε για τις δοκιμές είναι το ίδιο που χρησιμοποιήθηκε στο πρόγραμμα forecast.

# Πρόγραμμα reduce

## Σκοπός

Το πρόγραμμα αυτό έχει σκοπό την λήψη διανυσμάτων μέγεθος d και την συμπίεση τους σε διανύσματα μήκους δ<d. Η είσοδος είναι δύο αρχεία csv, κάθε ένα εκ των οποίων σε κάθε του γραμμή περιλαμβάνει τις συντεταμένες ενός διανύσματος χωρισμένα με κόμμα.

## Διαδικασία

### Έλεγχος Παραμέτρων

Σε πρώτη φάση γίνεται ο έλεγχος των παραμέτρων εισόδου. Αυτές θα πρέπει να είναι δύο αρχεία χρονοσειρών σε μορφή csv. Σε κάθε αρχείο csv αρχείο, η πρώτη στήλη αφορά την ονομασία της μετοχής και οι υπόλοιπες (έστω d) τις τιμές των μετοχών (από την παλαιότερη στην νεότερη). Επιπλέον δίνονται και δύο ονόματα αρχείου στα οποία θα τοποθετηθούν οι αντίστοιχες παραστάσεις των περιεχομένων στα δύο αρχεία εισόδου, σε διανύσματα μικρότερων διαστάσεων. Αν οι παράμετροι δοθούν σωστά το πρόγραμμα συνεχίζει στην επόμενη λειτουργία, διαφορετικά τερματίζει με κατάλληλο μήνυμα σφάλματος.

### Προετοιμασία Δεδομένων

Τα δεδομένα που λαμβάνονται από το αρχείο εισόδου μετασχηματίζονται σε dataframe. Στη συνέχεια καθορίζεται το βάθος ημερών στο οποίο είναι επιθυμητό να βασιστεί η πρόβλεψη (μεταβλητή look\_back – default τιμή 4). Από το dataframe επιλέγονται οι πρώτες n γραμμές που καθορίζονται από την αντίστοιχη μεταβλητή εισόδου. Τα δεδομένα στη συνέχεια κανονικοποιούνται (διαιρείται κάθε τιμή των δεδομένων με την μέγιστη τιμή τους). Με την συνάρτηση create\_dataset προσαρμόζεται η μορφή κάθε χρονοσειράς για την εκπαίδευση μοντέλου με βάση look\_back δείγματα.

### Κατασκευή του μοντέλου

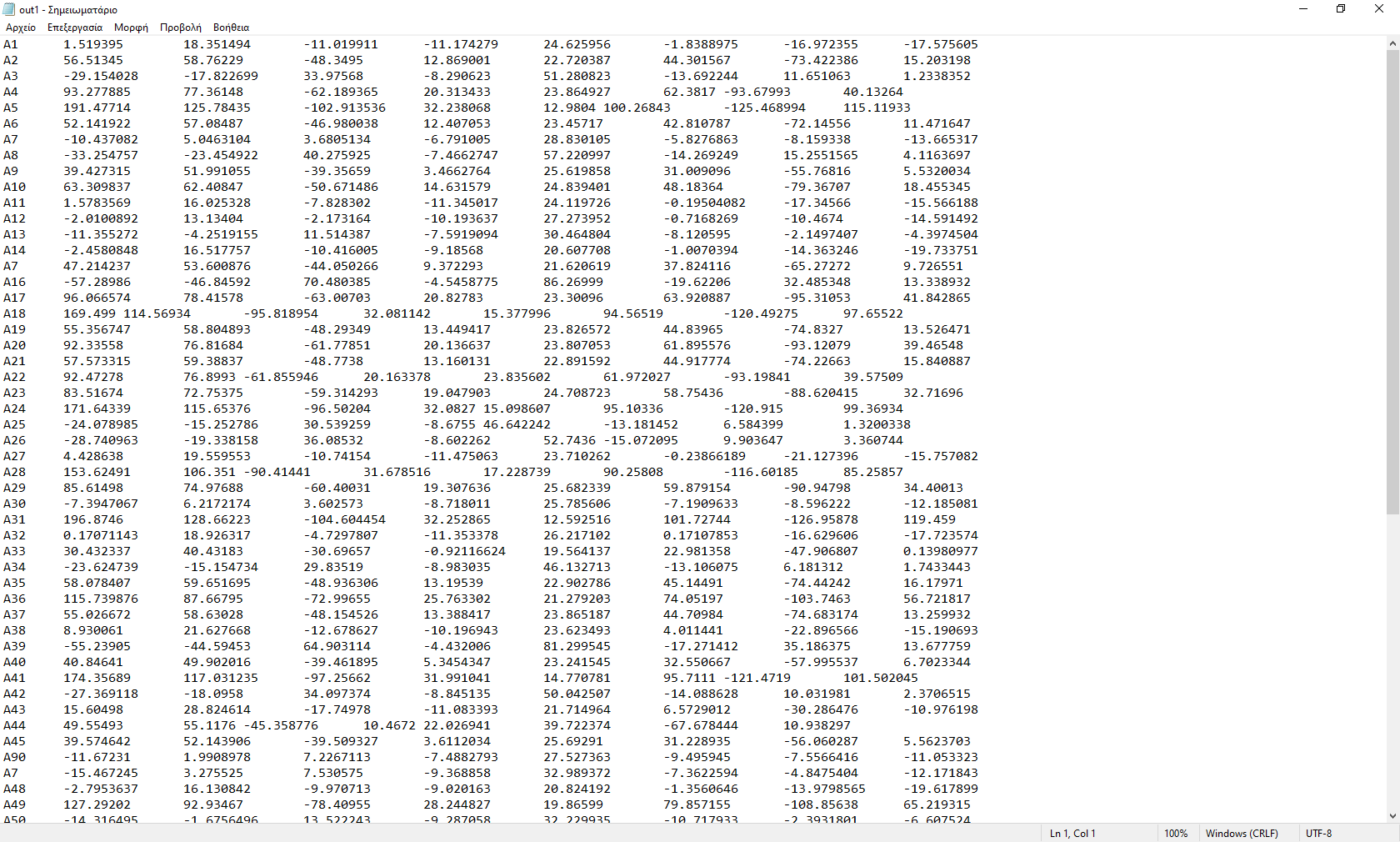
Το μοντέλο περιλαμβάνει συνελικτικά επίπεδα, καθώς και επίπεδα συμπίεσης και αποσυμπίεσης – συμμετρικά τοποθετημένα. Επιπλέον περιλαμβάνει και επίπεδο Dropout για την αποφυγή του overfitting. Από τα πειράματα που έγιναν προέκυψε το συμπέρασμα ότι οι καλύτερες παράμετροι για την απόδοση του μοντέλου είναι τέσσερα συνελικτικά επίπεδα (δύο στον κωδικοποιητή και δύο στον αποκωδικοποιητή). Κάθε συνελικτικό επίπεδο ακολουθείται από ένα MaxPooling (τμήμα κωδικοποιητή) ή ένα Upsampling (τμήμα αποκωδικοιητή). Στο τέλος του κωδικοποιητή υπάρχει ένα επίπεδο Dense και ένα επίπεδο Flatten για την προσαρμογή της εξόδου σε ένα διάνυσμα μίας διάστασης (με το επιθυμητό μέγεθος παραμέτρων).

### Εκπαίδευση του μοντέλου

Το μοντέλο εκπαιδεύτηκε με κοινή είσοδο και ετικέτες σε όλη του την έκταση. Οι τιμές των παραμέτρων που έδειξαν καλύτερη απόδοση στις δοκιμές είναι batch 16 και epochs 100.

### Χρήση του μοντέλου

Από το μοντέλο χρησιμοποιείται μόνο ο κωδικοποιητής ο οποίος συμπιέζει το διάνυσμα εισόδου και αποδίδει την αντιστοιχία του σε διάνυσμα μικρότερου μεγέθους. Από το μοντέλο περνάνε τα διανύσματα των αρχείων εισόδου και τα αποτελέσματα εισέρχονται στα αντίστοιχα αρχεία εξόδου. Στην εικόνα φαίνεται ένα δείγμα του αρχείου εξόδου.



Με τις τρέχουσες ρυθμίσεις το πρόγραμμα δημιουργεί διανύσματα μεγέθους 16 παραμέτρων από διανύσματα των 32Χ32 παραμέτρων. Σας είσοδο μπορεί να λάβει μία χρονοσειρά τουλάχιστον 64 τιμών. Από αυτές κρατά για επεξεργασία τις 64 πρώτες και εξετάζει τις τιμές με βάση τις προηγούμενες 32 ημέρες (δημιουργεί για κάθε μετοχή έναν πίνακα 32Χ32). Αυτά τα διανύσματα τα μετασχηματίζει σε διανύσματα 16 παραμέτρων.

# Δοκιμή της search για τα κωδικοποιημένα αρχεία

Εκτελέστηκε το πρόγραμμα search το κανονικό query και για το αντίστοιχο με περιεχόμενα τα συμπιεσμένα διανύσματα. Τα αποτελέσματα που προέκυψαν ήταν διαφορετικά ως προς το πιο κοντινό διάνυσμα για κάθε ένα από αυτά. Σε καμία περίπτωση δεν συνέπεσαν τα πιο κοντινά διανύσματα για τις κανονικές αναπαραστάσεις να είναι ίδιες με τις αντίστοιχες συμπιεσμένες. Στις επόμενες εικόνες φαίνονται δύο στιγμιότυπα για εκτιμήσεις για το ίδιο αρχείο.

