Ανάλυση Χρονοσειρών

Πρόβλεψη Ερευνητικών Τάσεων από Βιβλιογραφικά Δεδομένα

Δημήτριος Τσέλιος

ΠΜΣ Προηγμένα Πληροφοριακά Συστήματα, ΑΜ ΜΕ2059

1 ΕΙΣΑΓΩΓΗ

Μία από τις χρήσιμες εφαρμογές της ανάλυσης χρονοσειρών είναι η παρατήρηση της εξέλιξης των μεγεθών που περιγράφουν και η εκτίμηση της μελλοντικής πορείας της ακολουθίας των παρατηρήσεων. Η διαδικασία αυτή έχει τυποποιηθεί με επιτυχία μέσω μαθηματικών μοντέλων, τα οποία είναι γνωστά ως μοντέλα πρόβλεψης. Χωρίς αμφιβολία, το ενδιαφέρον και η σημασία της πρόβλεψης έχει αυξηθεί ραγδαία τα τελευταία χρόνια. Στόχος της παρούσας εργασίας είναι η ανάπτυξη ενός μοντέλου πρόβλεψης της ερευνητικής τάσης για τα επόμενα χρόνια από τα βιβλιογραφικά δεδομένα του DBLP ιστότοπου που διατηρεί βιβλιογραφικά στοιχεία δημοσιεύσεων, κυρίως στον χώρο της επιστήμης των υπολογιστών. Ειδικότερα, αναπτύχθηκε ένα μοντέλο πρόβλεψης της ερευνητικής τάσης των δημοσιεύσεων για τη θεματική περιοχή της υπολογιστικής γλωσσολογίας (computational linguistics) για τα επόμενα δέκα έτη. Για τη πρόβλεψη της χρονοσειράς, εφαρμόστηκε η διαδικασία εξαγωγής των δεδομένων, της προ-επεξεργασίας και προετοιμασίας των δεδομένων, της ανάλυσης των δεδομένων, της πρόβλεψης και της αποτίμησης της πρόβλεψης. Για την υλοποίηση των ανωτέρω, έγινε χρήση της γλώσσας προγραμματισμού python σε περιβάλλον spider και Jupiter, καθώς επίσης και χρήση των βιβλιοθηκών Pandas, Numpy, Sklearn Matplotlib, Statsmodel, Math, Seaborn και Pylab.

2 ΠΡΟΒΛΕΨΗ ΕΡΕΥΝΗΤΙΚΩΝ ΤΑΣΕΩΝ

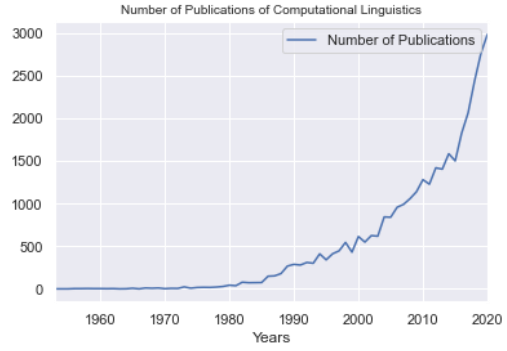
2.1 Εξαγωγή των Δεδομένων

Για την εξαγωγή των δεδομένων, χρησιμοποιήθηκε το αρχείο “dblp-2021-02-01.xml.gz” του ιστότοπου http://dblp.org/xml/release/ κατασκευάστηκε ένας text parser με τη γλώσσα προγραμματισμού Python. Ο συγκεκριμένος text parser, ανοίγει το αρχείο και δημιουργεί ένα λεξικό (count) με τη μέθοδο dict(), όπου αποθηκεύει ως κλειδιά (keys) τα έτη που έχουν δημοσιευθεί θέματα σχετικά με την υπολογιστική γλωσσολογία και ως τιμές (values) το πλήθος τους. Για την επίτευξη των παραπάνω, κατασκευάστηκε μια λίστα (words) με τις λέξεις οι οποίες θεωρήθηκαν απαραίτητες για την αναζήτηση, με στόχο την εύρεση των περισσότερων δυνατών αποτελεσμάτων. Οι λέξεις που χρησιμοποιήθηκαν είναι οι ακόλουθες: linguistic, computational linguistic, speech processing, machine translation, natural language process, nlp, computer based translation, translating machine, lexical resources, speech recognition, translation memory, computational lexicology, speech synthesis, social media mining, text editor, grammar correction, word processing, spelling correction. Ο parser διαβάζει το αρχείο ανά γραμμή και εντοπίζει πρώτα τους τίτλους που περιέχουν τις παραπάνω λέξεις. Πιο συγκεκριμένα, μέσω δομών ελέγχου, αναζητούνται οι γραμμές που περιέχουν τους τίτλους κάθε δημοσίευσης (‘<title>’) και ελέγχεται εάν υπάρχει κάποια από τις παραπάνω λέξεις στο περιεχόμενό τους. Εάν υπάρχει, αυξάνεται κατά ένα ο μετρητής (total) και γίνεται αναζήτηση στις επόμενες γραμμές για το έτος το οποίο αναφέρεται η εκάστοτε δημοσίευση (‘<year>’) . Σημειώνεται, ότι για να μην υπάρξει απώλεια πληροφορίας λόγω πεζών και κεφαλαίων γραμμάτων, κατά την αναζήτηση, γίνεται μετατροπή των γραμμάτων κάθε λέξης των τίτλων σε πεζά. Τα δεδομένα ταξινομούνται σε δύο στήλες σε έτη (Years) και αριθμό δημοσιεύσεων (Number of Publications) και αποθηκεύονται σε ένα αρχείο κειμένου (txt).

2.2 Προ-επεξεργασία και Προετοιμασία των Δεδομένων

Με χρήση της βιβλιοθήκης Pandas γίνεται εισαγωγή του dataset στο περιβάλλον Jupyter, δηλώνεται η συχνότητα των περιόδων ανά έτος και εντοπίζονται οι ελλειπείς τιμές. Το σύνολο δεδομένων περιέχει δημοσιεύσεις που ξεκινούν από το 1953 έως το 2020, καθώς το 2021 είναι τρέχον έτος και δε μπορεί να συμπεριληφθεί στην ανάλυση. Παρατηρήθηκε ότι εμφανίζονται ελλειπείς τιμές κατά τα πρώτα χρόνια του συνόλου δεδομένων, κάτι που είναι λογικό καθώς οι τιμές των προηγούμενων και των επόμενων ετών εμφανίζουν εξαιρετικά χαμηλές τιμές εκείνη την περίοδο, καθώς επίσης και λόγω του πρώιμου σταδίου που βρισκόταν ερευνητικά το επιστημονικό πεδίο της υπολογιστικής γλωσσολογίας τη δεκαετία του 1950. Για την αντιμετώπιση του προβλήματος των ελλειπών τιμών εκχωρήθηκαν στα έτη που δεν περιείχαν τιμές, οι τιμές του ακριβώς προηγούμενου έτους.

Στο διάγραμμα που ακολουθεί απεικονίζονται το σύνολο των δημοσιεύσεων που αφορούν την υπολογιστική γλωσσολογία, έπειτα από την διαδικασία προ-επεξεργασίας και επεξεργασίας του συνόλου δεδομένων.



Εικόνα 1: Ο αριθμός των δημοσιεύσεων ανά έτος που αφορούν την υπολογιστική γλωσσολογία. Ανάλυση Δεδομένων

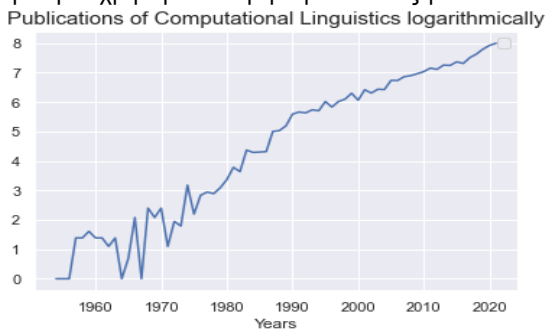
Σύμφωνα με το παραπάνω διάγραμμα, μια πρώτη εικόνα που μπορούμε να συμπεράνουμε για τα δεδομένα, είναι ότι μέχρι τα πρώτα χρόνια της δεκαετίας του 1980 ο αριθμός των δημοσιεύσεων κυμαίνεται σε εξαιρετικά χαμηλά επίπεδα, ενώ κατά τη διάρκεια της δεκαετίας του 1980 έως το 2020, παρατηρείται εκθετική αύξηση του αριθμού των δημοσιεύσεων Μπορούμε έχουμε μια πρώτη εκτίμηση για τα

δεδομένα της χρονοσειράς οτι εμφανίζουν τάση, ενώ δεν έχει τα χαρακτηριστικά της εποχικότητας, της κυκλικότητας και των ακραίων τιμών. Για να διαπιστώσουμε τις παραπάνω υποθέσεις, θα πρέπει να γίνει έλεγχος στασιμότητας χρονοσειράς. Μια χρονοσειρά είναι στάσιμη αν θεωρήσουμε ότι οι στατιστικές της ιδιότητες παραμένουν σταθερές στο χρόνο, δηλαδή όταν δεν υπάρχει συστηματική αλλαγή του μέσου όρου και της τυπικής απόκλισης της στο χρόνο. Το Dickey-Fuller test είναι μία μέθοδος που παράγει στατιστικά αποτελέσματα για την εκτίμηση της στασιμότητας. Για τον Dickey-Fuller κάνουμε την υπόθεση μηδέν (Ho) ότι η χρονοσειρά δεν είναι στάσιμη. Από αποτελέσματα που λαμβάνουμε θα κάνουμε έλεγχο σημαντικότητας μεταξύ του test_statistic και των critical_values των διαφόρων επιπέδων σημαντικότητας. Όπως παρατηρείται, ο στατιστικός έλεγχος είναι μεγαλύτερος από τις κριτικές τιμές των επιπέδων σημαντικότητας οπότε δεν απορρίπτουμε την υπόθεση μηδέν και θεωρούμε ότι η χρονοσειρά δεν είναι στάσιμη .

Πίνακας 1: Dickey-Fuller test

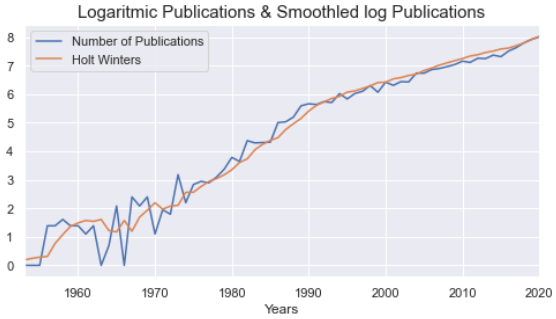
Test Statistic	4.56448454145
p-value	1.0
lags	11
Critical values:	
1%	-3.5529282035
5%	-2.9147306260
10%	-2.5951371556

Για να προχωρήσουμε στην κατασκευή του μοντέλου, θα πρέπει να μετατρέψουμε τη χρονοσειρά σε στάσιμη, δηλαδή να απαλλάξουμε τη χρονοσειρά από τάση και εποχικότητα. Ένας τρόπος για να μειωθεί η τάση είναι ο μετασχηματισμός της χρονοσειράς με τρόπο που να περιορίζει τις ψηλές τιμές σε σχέση με τις χαμηλές. Ο μετασχηματισμός μπορεί να γίνει με διάφορους τρόπους (λογάριθμο, τετραγωνική ή κυβική ρίζα). Στη συγκεκριμένη χρονοσειρά έγινε χρήση του λογάριθμου καθώς φαίνεται να ταιριάζει.



Εικόνα 2: Μετατροπή της χρονοσειράς σε λογαριθμική

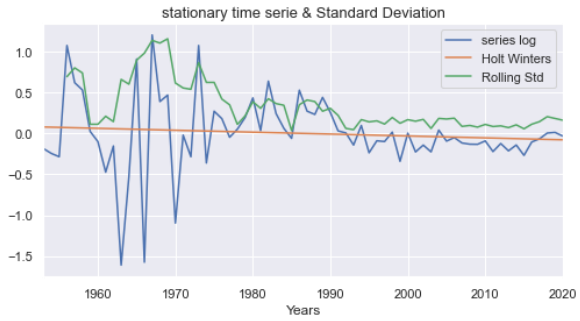
Από το διάγραμμα παρατηρούμε ότι το εύρος των τιμών των δημοσιεύσεων έχει μειωθεί και η τάση δείχνει να αυξάνεται πλέον γραμμικά. Υπάρχουν αρκετοί τρόποι μετασχηματισμού των χρονοσειρών ώστε να μετατραπούν σε στάσιμες, όπως αφαιρώντας τον κινητό μέσο ή την εκθετική εξομάλυνση. Στη λογαριθμική χρονοσειρά εφαρμόστηκε διπλή εκθετική εξομάλυνση, καθώς έπειτα από αποσύνθεση (decomposition) της χρονοσειράς, φάνηκε να υπάρχει μόνο τάση.



Εικόνα 3: Εξομάλυνση χρονοσειράς με τη μέθοδο Holt Winters

Προκειμένου να μετατραπεί η χρονοσειρά σε στάσιμη, θα πρέπει να αφαιρέσουμε από την λογαριθμική χρονοσειρά, την εξομαλυμένη χρονοσειρά.

Test Statistic	-8.6889830373
p-value	1.4148323e-14
Lags	0
Critical values	
1%	-3.5319549603



Εικόνα 4: Μετατροπή της χρονοσειράς σε στάσιμη.

5%	-2.9057551285
10%	-2.9057551285

Πίνακας2:Dickey-Fuller test

Για να επιβεβαιώσουμε τη στασιμότητα πραγματοποιήσαμε έλεγχο στασιμότητας εφαρμόζοντας Dickey-Fuller test, παρατηρούμε ότι στον στατιστικό έλεγχο ο test_statistic έχει τιμή μικρότερη από όλα τα επίπεδα σημαντικότητας , άρα μπορούμε να πούμε ότι με 99% σιγουριά η χρονοσειρά είναι στάσιμη.

2.3 Πρόβλεψη και Αποτίμηση της Πρόβλεψης

Για την ακρίβεια της πρόβλεψης γίνεται διάσπαση των δεδομένων σε train και test. Τα δεδομένα ως το 2015 θα αποτελέσουν το σύνολο εκπαίδευσης, ενώ από το 2016 έως το 2020 το σύνολο ελέγχου. Επιλέχθηκαν μεταξύ άλλων τα μοντέλα Holt Winters και ARIMA για να γίνει η πρόβλεψη της χρονοσειράς καθώς είναι αξιόπιστα και περισσότερο αυτοματοποιημένα, με αποτέλεσμα εμφανίζουν μικρότερα σφάλματα από πιο απλά μοντέλα όπως του μέσου όρου, του AR, MA ή του ARMA.
Εφαρμόζοντας το μοντέλο πρόβλεψης Holt Winters στα δεδομένα εκπαίδευσης, λαμβάνουμε τα παρακάτω αποτελέσματα στα δεδομένα ελέγχου:

Πίνακας 3: Σύγκριση προβλεπόμενων Holt Winters και πραγματικών

	Original values	Predicted values
2016-12-31	7.509335	7.619768
2017-12-31	7.631432	7.737739
2018-12-31	7.798113	7.855710
2019-12-31	7.924434	7.973681
2020-12-31	8.000685	8.091651

Για την αξιολόγηση της επίδοσης του μοντέλου πρόβλεψης θα χρησιμοποιήσουμε τρία διαφορετικά μέτρα σφάλματος. Το μέσο τετραγωνικό σφάλμα (MSE), τη ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) και μέσο απόλυτο σφάλμα (MAE).

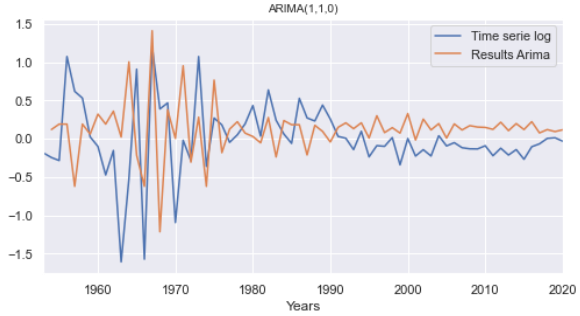
Πίνακας 4: Σφάλματα πρόβλεψης του μοντέλου Holt Winters

RMSE	0.086
MAE	0.083
MSE	0.007

Η μέθοδος ARIMA (Auto-Regressive Integrated Moving Averages) είναι μία γραμμική εξίσωση όπως είναι η γραμμική παλινδρόμηση. Οι βασικές παράμετροι του μοντέλου ARIMA που διαμορφώνουν την πρόβλεψη είναι οι (p, d, q):

1. Αυτοπαλινδρομούμενη χρονοσειρά (autoregressive time serie) τάξης p (AR(p)): Οι όροι του AR, είναι οι χρονικές υστερήσεις της εξαρτημένης μεταβλητής.
2. Χρονοσειρά κινητού μέσου τάξης q (MA(q)): Οι όροι του MA είναι οι χρονικές υστερήσεις των σφαλμάτων της πρόβλεψης.
3. Ο αριθμός των διαφορών (d) που παίρνουμε για μετατρέψουμε τη χρονοσειρά σε στάσιμη.

Οι παράμετροι p και q μπορούν να φανούν διαγραμματικά σε μια στάσιμη χρονοσειρά από τα διαγράμματα PACF και ACF αντίστοιχα, παρατηρώντας τους σημαντικούς συντελεστές των χρονικών υστερήσεων όπου φαίνονται να επηρεάζουν σημαντικά έως τη χρονική στιγμή 0. Όμως η οπτική παρατήρηση αυτών των χρονικών υστερήσεων δεν παρέχουν αξιοπιστία στις τιμές των παραμέτρων. Για να εντοπίσουμε τις βέλτιστες παραμέτρους του μοντέλου ARIMA χρησιμοποιήθηκε η μέθοδος της διασταυρωμένης επικύρωσης. Έτσι, θα λάβουμε πολλούς συνδυασμούς των παραμέτρων p,d,q και ταυτόχρονα τα λάβουμε το μέσο τετραγωνικό τους σφάλμα (MSE). Εφαρμόζοντας το μοντέλο πρόβλεψης ARIMA(1,1,0) στα δεδομένα εκπαίδευσης, λαμβάνουμε τα παρακάτω αποτελέσματα στα δεδομένα ελέγχου. Best ARIMA(1, 1, 0) MSE=0.002

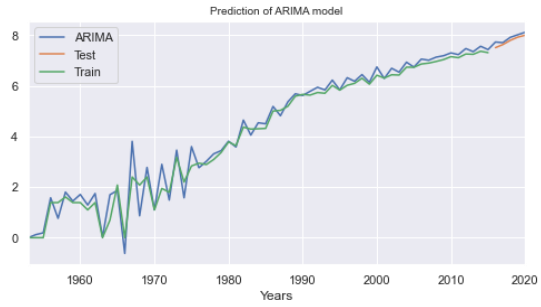


Εικόνα 5: Μοντέλο ARIMA(1,1,0)

	Original values	Predicted values
2016-12-31	7.509335	7.535034
2017-12-31	7.631432	7.583845
2018-12-31	7.798113	7.750078
2019-12-31	7.924434	7.891368
2020-12-31	8.000685	8.041857

Πίνακας 3: ARIMA(1,1,0) έναντι πραγματικών τιμών

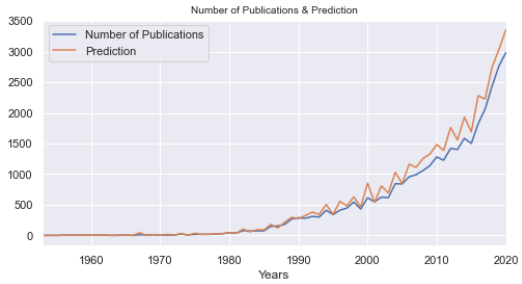
Για την αξιολόγηση της επίδοσης του μοντέλου πρόβλεψης θα χρησιμοποιήσουμε ξανά τα τρία στατιστικά μέτρα. Το μέσο τετραγωνικό σφάλμα (MSE), τη ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) και μέσο απόλυτο σφάλμα (MAE).
Συγκρίνοντας τα δεδομένα και οπτικά παρατηρείται ότι το μοντέλο που κατασκευάστηκε προβλέπει τόσο την ανοδική πορεία των πραγματικών δεδομένων όσο και την εκθετική τους αύξηση (Εικόνα 7).



Εικόνα 6: Δεδομένα ARIMA, εκπαίδευσης και ελέγχου.

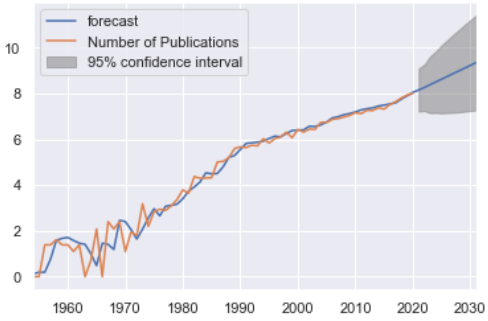
RMSE	0.040
MAE	0.039
MSE	0.002

Πίνακας 4: Σφάλματα πρόβλεψης του μοντέλου ARIMA(1,1,0)



Εικόνα 7: Σύγκριση δεδομένων πρόβλεψης και πραγματικού αριθμού δημοσιεύσεων.

Εφόσον επιλέχθηκε το καταλληλότερο μοντέλο, πραγματοποιήθηκε μια πρόβλεψη για την πορεία που αναμένεται να έχουν οι δημοσιεύσεις της υπολογιστικής γλωσσολογίας τα επόμενα δέκα έτη. Αυτό που παρατηρείται, είναι ότι στη λογαριθμική χρονοσειρά, φαίνεται να συνεχίζεται η γραμμική αύξηση που διατηρούσε, κάτι που σημαίνει ότι στην πραγματικότητα η τάση θα δείχνει να εμφανίζει εκθετική αύξηση αριθμού νέων δημοσιεύσεων έως το 2030.



Εικόνα 8: Πρόβλεψη πλήθους δημοσιεύσεων για τα επόμενα δέκα έτη.

3 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η συνεχής εξέλιξη της τεχνολογίας και το έντονο ερευνητικό ενδιαφέρον που υπάρχει γύρω από τη θεματική περιοχή της υπολογιστικής γλωσσολογίας (Computational Linguistics), φαίνεται να αυξάνεται εκθετικά στο χρόνο. Έτσι, όπως μπορεί να παρατηρηθεί διαγραμματικά, η συγκεκριμένη θεματική αναμένεται να σημειώσει ανοδική πορεία για τα επόμενα δέκα έτη.

BIBLIOGRAPHY

Brownlee, J. (2011). *Introduction to Time Series Forecasting With Python: How to Prepare Data and Develop Models to Predict the Future.* Machine Learning Mastery.

Dean, J. (2014). *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners.* Wiley.

Gebhard Kirchgässner, J. W. (2013). *Introduction to Modern Time Series Analysis.* Springer Texts in Business and Economics.

Janert, P. K. (2010). *Data Analysis with Open Source Tools: A hands-on guide for programmers and data scientists.* O'Reilly Media.

Søren Bisgaard, M. K. (2011). *Time Series Analysis and Forecasting by Example.* Wiley.