Πρόβλεψη ερευνητικών τάσεων για το Knowledge Representation

Ευαγγελία-Μαρία Σπύρου,

Φοιτήτρια Μεταπτυχιακού Προηγμένα Πληροφοριακά Συστήματα , AM:ME2056

1 ΕΙΣΑΓΩΓΗ

Με τον όρο χρονοσειρά εννοούμε μια ακολουθία παρατηρήσεων που λαμβάνεται σε ίσες χρονικές περιόδους. Σε αντίθεση με τα κανονικά δεδομένα ,οι χρονοσειρές απαιτούν χρονολογική σειρά από την άποψη της μηχανικής μάθησης. Στην παρούσα εργασία θα ασχοληθούμε με τεχνικές ανάλυσης χρονοσειρών με σκοπό την πρόβλεψη ερευνητικών τάσεων με βάση την προηγουμένη συμπεριφορά τους γύρω από την θεματική περιοχή του Knowledge Representation.Το Knowledge Representation είναι ένα μέρος της τεχνητής νοημοσύνης που περιγράφει πως μπορούμε να αντιπροσωπεύσουμε την γνώση στην τεχνητή νοημοσύνη ώστε μια έξυπνη μηχανή να μάθει από την γνώση και τις εμπειρίες ,ώστε να συμπεριφέρεται έξυπνα σαν άνθρωπος και να μπορεί να λύσει προβλήματα του πραγματικού κόσμου όπως η διάγνωση μιας ιατρικής κατάστασης ή επικοινωνία με τους ανθρώπους σε φυσική γλώσσα. Για την εκπόνηση της εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού python καθώς και οι βιβλιοθήκες Pandas, NumPy, Statsmodels, Matplotlib, Sklearn , Seaborn , Math.

2 ΔΙΑΧΕΙΡΙΣΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Για την παρούσα εργασία αντλήσαμε ένα σύνολο δεδομένων από τον ιστότοπο DBLP (https://dblp.uni-trier.de/) και συγκεκριμένα από το αρχείο (dblp-2021-02- 01.xml.gz).Το συγκεκριμένο αρχείο περιέχει ένα μεγάλο πλήθος δημοσιεύσεων γύρω από την επιστήμη των υπολογιστών. Για να αντλήσουμε τα δεδομένα ,δημιουργήσαμε σε python έναν parser , ο οποίος διαβάζει κάθε γραμμή αναζητώντας το tag <title>,ελέγχει αν βρίσκεται μέσα σε μια λίστα με λέξεις που αφορούν την θεματική περιοχή που μελετάμε και υστέρα αναζητάει στις επόμενες 4 γραμμές το tag <year> και προσθέτει στο data το title άμα δεν υπάρχει .Αυτή η διαδικασία γίνεται με τη χρήση της ιδιότητας του Dictionary της python που αποθηκεύει τα δεδομένα στη μορφή key(χρονιά) value(δημοσιεύσεις).Αφού εξάγουμε τα δεδομένα στη συνέχεια τα ταξινομούμε κατά αύξουσα χρονολογική σειρά και τα αποθηκεύουμε σε μορφή txt αρχείου. Οι λέξεις που συμπεριλήφθηκαν στη λίστα είναι η θεματική περιοχή, πεδία στα οποία χρησιμοποιείται (Semantic web,experts systems,frames,semantic nets,Knowledge-based Systems) τύπους knowledge (Meta-Knowledge,declarative Knowledge,heuristic Knowledge,procedural knowledge),καθώς και η θεματική περιοχή (automated reasoning) καθώς η συγκεκριμένη θεματική περιοχή στην ουσία βασίζεται στο Knowledge Representations.

Στην συνέχεια περάσαμε την χρονοσειρά σε dataframe και αφαιρέσαμε το 2021 διότι είναι η τρέχουσα χρονιά και θα χαλούσε την χρονοσειρά. Εφαρμόσαμε μεθόδους ώστε να εξασφαλίσουμε την σταθερή συχνότητα (ανά έτος) και στην συνέχεια αντιμετωπίσαμε το πρόβλημα των ελλείπων τιμών και την απεικονίσαμε διαγραμματικά. Από το διάγραμμα παρατηρούμε ότι η χρονοσειρά δεν εμφανίζει εποχικότητα, κυκλικότητα και ακραίες τιμές. Αλλά αυτό που μας φανερώνει το διάγραμμα της είναι ότι η χρονοσειρά δεν είναι απαλλαγμένη από τάση. Η τάση φαίνεται να αυξάνεται με εκθετικό ρυθμό. Το συμπέρασμα που θα μπορούσαμε να πούμε είναι ότι οι δημοσιεύσεις στην θεματική περιοχή που εξετάζουμε αυξάνονται με εκθετικό ρυθμό. Βέβαια δεν είναι εύκολο να βγάζουμε συμπεράσματα μόνο από τα διάγραμμα οπότε θα προχωρήσουμε και σε κάποια τεστ.

3 ΕΛΕΓΧΟΣ ΣΤΑΣΙΜΟΤΗΤΑΣ

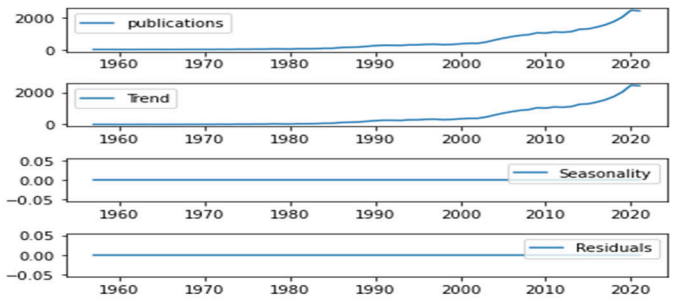
Λέμε ότι μια χρονοσειρά είναι στάσιμη όταν η μέση τιμή και η διακύμανση παραμένουν σταθερές στο χρόνο. Για να εξακριβώσουμε την στασιμότητα της χρονοσειράς θα χρησιμοποιήσουμε το Dickey-Fuller test της βιβλιοθήκης Statsmodels.

Dickey-Fuller test	
ADF Statistic:	5,140522
p-value:	1,000000
Critical Values:	
1%:	-3,558
5%:	-2,917
10%:	-2,596

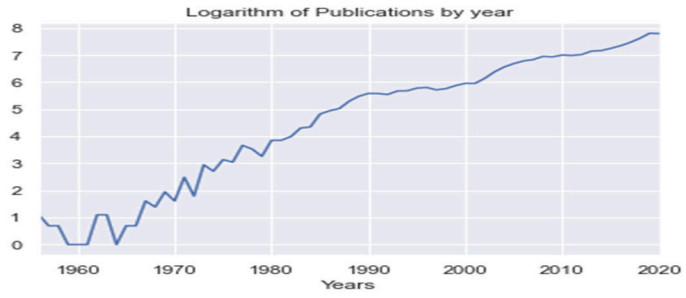
Η python μας παρέχει τις κρίσιμες τιμές 1%,5%,10% τις οποίες θα χρησιμοποιήσουμε σαν significant level.Επίσης κάνουμε την υπόθεση ότι για Ho:p>0,05 η χρονοσειρά δεν είναι στάσιμη έναντι του Ho:p<=0,05 ,οπού και μπορούμε να απορρίψουμε την μηδενική υπόθεση. Από ότι μπορούμε να δούμε από τα αποτελέσματα του Dickey fuller δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση διότι ο στατιστικός έλεγχος απέχει πολύ από τις κρίσιμες τιμές και η τιμή p είναι μεγαλύτερη από το (0,05) .Έτσι συμπεραίνουμε ότι η χρονοσειρά δεν είναι στάσιμη.

3.1 Μετατροπή χρονοσειράς σε Στάσιμη

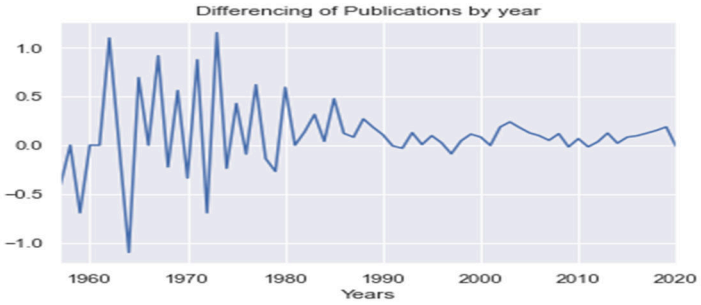
Μια χρονοσειρά δεν είναι στάσιμη γιατί μπορεί να αποτελείται από τάση και εποχικότητα για να διευκρινίσουμε τα χαρακτηριστικά της χρονοσειράς θα μοντελοποιήσουμε ξεχωριστά την τάση ,την εποχικότητα και τα υπόλοιπα με την μέθοδο της αποσύνδεσης και θα τα απεικονίσουμε διαγραμματικά .



Όπως μπορούμε να δούμε από την απεικόνιση η χρονοσειρά περιέχει μόνο τάση και δεν περιέχει εποχικότητα ή υπόλοιπα .



Ένας τρόπος για να σταθεροποιήσουμε την μη σταθερή διακύμανση της χρονοσειράς και να μειώσουμε την τάση είναι ο λογάριθμος.
Αφού εφαρμόσουμε τον λογάριθμο στην χρονοσειρά θα την απεικονίσουμε και τώρα μπορούμε να παρατηρήσουμε ότι η τάση φαίνεται να αυξάνεται γραμμικά.



Στη συνέχεια θα χρησιμοποιούμε τον διαφορικό μετασχηματισμό για να καταστήσουμε στάσιμη τη χρονοσειρά.
Ο διαφορικός μετασχηματισμός (Differencing) μπορεί να βοηθήσει στην σταθεροποίηση της χρονοσειράς εξαλείφοντας ή μειώνοντας την τάση. Ο μετασχηματισμός αυτός λειτουργεί αφαιρώντας την προηγουμένη παρατηρητή από την τρέχουσα. Κάνοντας ελέγχους στις περιόδους που χρησιμοποιεί ο διαφορικός μετασχηματισμός σε εύρος τιμών (0,10) τα καλυτέρα αποτελέσματα μας τα δίνει η περίοδος 1.

Dickey-Fuller test	
ADF Statistic:	-5,89297
p-value:	0,0000
Critical Values:	
1%:	-3,544
5%:	-2,911
10%:	-2,593

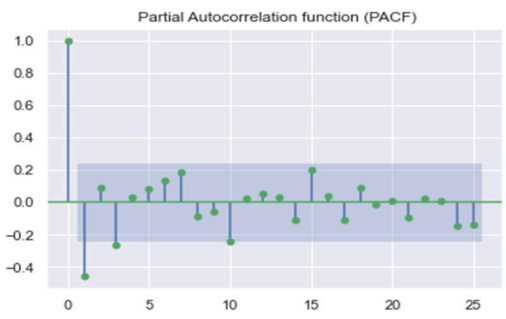
Από τα αποτελέσματα που μας δίνει το Dickey-Fuller test που το εφαρμόσαμε πάνω στο διαφορικό μετασχηματισμό συμπεραίνουμε ότι μπορούμε να απορρίψουμε την μηδενική υπόθεση με significant level μικρότερο από 1%.Αυτο σημαίνει ότι η χρονοσειρά είναι στάσιμη ή δεν έχει δομή που εξαρτάται από το χρόνο.

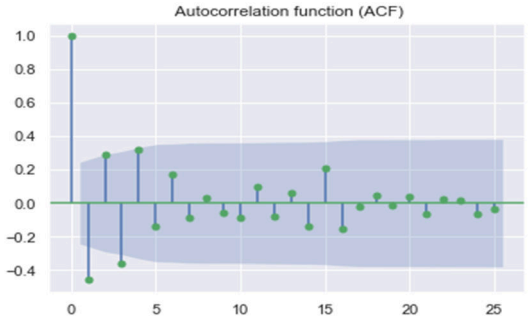
4 ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ ΧΡΟΝΟΣΕΙΡΑΣ

Οι χρονοσειρές μας δίνουν την δυνατότητα να μπορέσουμε να προβλέψουμε μελλοντικές τιμές με βάση τις προηγούμενες τιμές τους. Δυο από τις πιο γνωστές μεθόδους που χρησιμοποιούνται στις χρονοσειρές για πρόβλεψη είναι το ARIMA και ο Holt-Winters.θα κατασκευάσουμε αυτά τα δυο μοντέλα και στην συνέχεια θα συγκρίνουμε μεταξύ τους το μέσο τετραγωνικό σφάλμα(MSE) και το μέσο απολυτό σφάλμα(MAE) σαν μητρικές για να δούμε ποιο είναι καλύτερο για να κάνουμε πρόβλεψη.

4.1 Μέθοδος Πρόβλεψης ARIMA

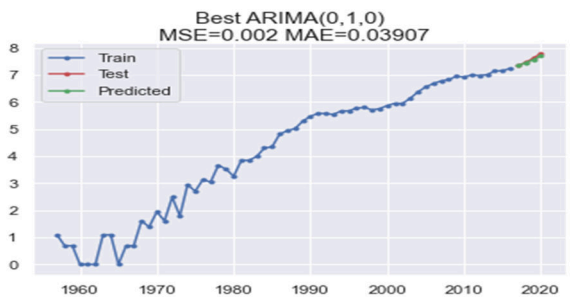
Το ARIMA σημαίνει **A**utoreg**r**essive **I**ntegrated **M**oving **A**verage.Το ARIMA είναι ένα μοντέλο που μπορεί να προσαρμοστεί σε δεδομένα χρονοσειρών προκειμένου να προβλέψει μελλοντικά σημεία στην σειρά. Υπάρχουν 3 παράμετροι (p,d,q) που χρησιμοποιούνται για της παραμετροποιήσεις του μοντέλου ARIMA. Για να μπορέσουμε να καθορίσουμε τις παραμέτρους θα χρησιμοποιήσουμε δυο τρόπους .Ο πρώτος είναι να απεικονίσουμε διαγραμματικά την αυτοσυσχέτιση (η χρονική υστέρηση που η ACF περνάει το διάστημα εμπιστοσύνης και προσεγγίζει το 0 είναι το q) και την μερική αυτοσυσχέτιση(η χρονική υστέρηση που η PACF περνάει το διάστημα εμπιστοσύνης και προσεγγίζει το 0 είναι το p) .Το d είναι η διαφορά που παίρνουμε για να μετατρέψουμε την χρονοσειρά σε στάσιμη οπότε θα πάρουμε την πρώτη διαφορά d=1 .

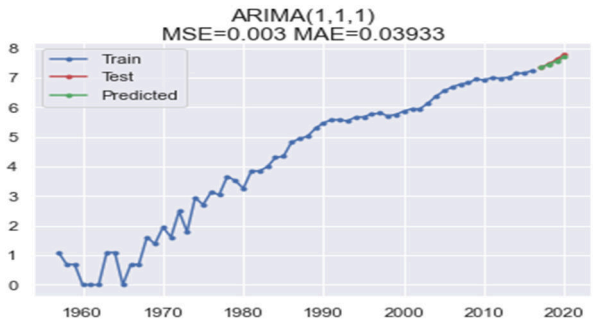




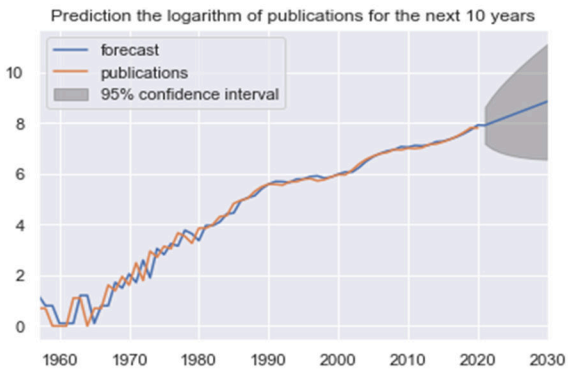
Αρά έχουμε από την απεικόνιση του ACF
και PCF ότι το q=1 και το p=1 για να κατασκευάσουμε το μοντέλο μας.

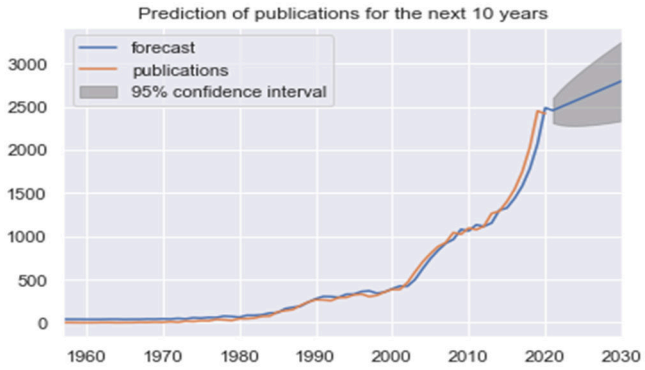
Ο δεύτερος τρόπος που αναφέραμε είναι να κατασκευάσουμε δυο συναρτήσεις οι οποίες παίρνουν ως είσοδο ένα σύνολο δεδομένων και ένα σύνολο παραμέτρων p,d,q .Δημιουργείται ένα μοντέλο για κάθε παράμετρο και αξιολογείται από το μέσο τετραγωνικό σφάλμα. Στο τέλος έχουμε μια λίστα από μοντέλα και τα MSE.Η πρόταση της συνάρτησης με το χαμηλότερο τετραγωνικό σφάλματα και καλύτερο μοντέλο σύμφωνα με την συνάρτηση είναι το ARIMA(0,1,0) με MSE=0,002.Αλλα μας έδωσε και αλλά μοντέλα με πολύ μικρά σφάλματα ένα από αυτά ήταν και το ARIMA(1,1,1) που υπολογίσαμε πριν από το ACF και PCF με MSE=0.003. Εφόσον τώρα πήραμε κάποια μοντέλα ώστε να μπορέσουμε να φτιάξουμε το μοντέλο πρόβλεψης .Θα πάμε να χωρίσουμε τα δεδομένά μας σε train μέχρι το «2015» και test από το «2016» έως το «2019» και θα πάμε να κάνουμε πρόβλεψη σε κάθε μοντέλο ξεχωριστά για να δούμε αν τα προβλεπόμενα αποτελέσματα ταιριάζουν με τα αναμενόμενα και θα εκτυπώσουμε το μέσο απολυτό σφάλμα για να δούμε αν το καλύτερο μοντέλο που μας προτείνει έχει και αυτό το χαμηλότερο απολυτό σφάλμα.





Από ότι μπορούμε να δούμε και τα δυο μοντέλα μας δίνουν πολύ κοντά αποτελέσματα σε σχέση με το test και το predicted και τα μέσα απόλυτα σφάλματα έχουν σχετικά αμελητέα διαφορά μεταξύ τους. Αλλά και πάλι το Best ARIMA που μας προτείνει η συνάρτηση που κατασκευάσαμε μας δίνει το μοντέλο με το μικρότερο απόλυτο σφάλμα και θα το προτιμήσουμε για να κάνουμε την πρόβλεψη για τα επόμενα δέκα χρόνια. Η χρονοσειρά ξεκινάει από το 1956 και φτάνει έως το 2020 (64 χρόνια) +10=74 .

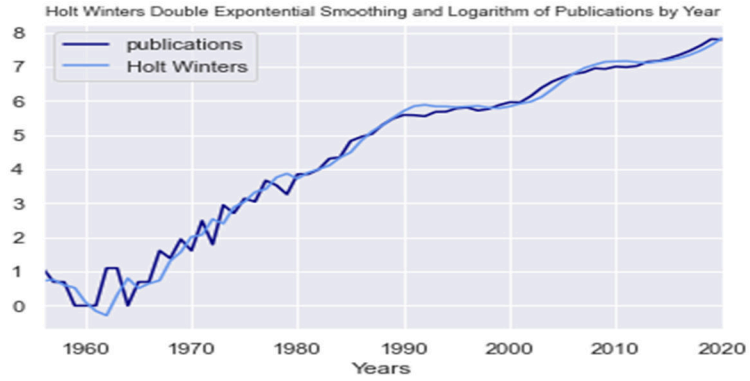




Εφαρμόζοντας το μοντέλο στα log δεδομένά αλλά και στα αρχικά παρατηρούμε ότι η πρόβλεψη είναι πολύ κοντά με τα δεδομένα και μάλιστα τα επόμενα 10 χρόνια βλέπουμε οι δημοσιεύσεις γύρω από Knowledge Representation να αυξάνονται. Βέβαια αυτό μπορεί να οφείλεται στο γεγονός ότι οι δημοσιεύσεις είχαν μια απότομη αύξηση στις προηγούμενες παρατηρήσεις τους και η πρόβλεψη τείνει να αυξάνεται με εκθετικούς ρυθμούς. Αυτό το συμπεραίνουμε στο γεγονός ότι το 2020 κάνει μια μικρή κλήση προς τα κάτω .Το forecast το αντιλαμβάνεται αλλά συνεχίζει ανοδικά.

4.2 Μέθοδος πρόβλεψης Holt-Winters

Ο Holt_Winters είναι μια μέθοδος που χρησιμοποιείται και αυτή για πρόβλεψη χρονοσειρών. Η χρήση της συγκεκριμένη μεθόδου μας επιτρέπει να εξομαλύνουμε τη χρονοσειρά και έπειτά να κάνουμε πρόβλεψη. Η εκθετική εξομάλυνσή βασίζεται στην εκθετική μείωση της βαρύτητας που δίνεται στα στοιχεία των προηγουμένων περιόδων. Δηλαδή όσο ποιο παλιά είναι τα στοιχεία τόσο πιο μικρή η βαρύτητα σε αντίθεση με τα πρόσφατα που θα έχουν μεγαλύτερη αξία από τα παλιότερά. Υπάρχουν 3 τύποι εξομάλυνσης η απλή(για δεδομένα χωρίς τάση και εποχικότητα),η διπλή(για δεδομένα με τάση) και η τριπλή εκθετική εξομάλυνση (για δεδομένα με τάση και εποχικότητα).Εμείς θα χρησιμοποιήσαμε την διπλή εκθετική εξομάλυνσή γιατί από ότι παρατηρήσαμε και παραπάνω από την αποσύνθεση της χρονοσειράς ,έχουμε μόνο τάση. Αφού εφαρμόζουμε την διπλή εκθετική εξομάλυνσή θα την απεικονίσουμε διαγραμματικά μαζί με τα Log δεδομένά μας .Όπως παρατηρούμε από το διάγραμμα τα log δεδομένά μας ταιριάζουν πολύ καλά με το holt.



Αφού χωρίσουμε τα δεδομένά σε train και test θα προσαρμόσουμε τα δεδομένά ώστε να προχωρήσουμε σε πρόβλεψη. Στη συνέχεια απεικονίζουμε τα train test και προβλεπόμενα μαζί στο ίδιο διάγραμμα. Παρατηρούμε ότι οι προβλέψεις απέχουν κάπως για να προσεγγίσουν τις κανονικές. Βέβαια και πάλι δεν μπορούμε να πούμε ότι είναι το καλύτερο ή το χειρότερο μοντέλο αν δεν υπολογίσουμε το μέσο τετραγωνικό σφάλμα και το απολυτό σφάλμα. Υπολογίζοντας τα σφάλματα έχουμε:

MSE: 0.07664899061170441

MAE:0.24547211108070432

Αρά τώρα ήμαστε σε θέση να πούμε ότι το καλύτερο μοντέλο είναι όντως το Best ARIMA αφού είχε τα χαμηλότερα σφάλματα.