Πρόγραμμα Μεταπτυχιακών Σπουδών:
«Διαχείριση Πληροφοριών σε Βιβλιοθήκες, Αρχεία, Μουσεία»

Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης
Σχολή Διοικητικών, Οικονομικών και Κοινωνικών Επιστημών

Αυτοματοποιημένη θεματική κατηγοριοποίηση στο ενεργό αρχείο του ΠΑΔΑ

Ioannis Triantafyllou

Digitally signed by Ioannis Triantafyllou
Date: 2023.11.13 16:40:48 +02'00'

Φοιτητής: Βασίλειος Βαλλιάνος

Επιβλέπων: Ιωάννης Τριανταφύλλου

Διαύγεια – Μεταδεδομένα Πράξεων

Πρόγραμμα Διαύγεια:

δημοσίευση στο διαδίκτυο των αποφάσεων:

- των κυβερνητικών οργάνων
- των φορέων του στενού και του ευρύτερου δημόσιου τομέα
- των Ανεξάρτητων Αρχών
- οργανισμών τοπικής αυτοδιοίκησης

Κάθε απόφαση ή αλλιώς **πράξη** πλαισιώνεται από ένα σύνολο μεταδεδομένων τα οποία την περιγράφουν:

- ΑΔΑ (Αριθμός Διαδικτυακής Ανάρτησης)
- **Θέμα** πράξης
- Ημερομηνία έκδοσης (Unix timestamp)
- **Είδος πράξης**
- **Θεματική (-ές) κατηγορία (-ες) πράξης**
- Το έγγραφο της πράξης σε μορφή **PDF** (URL)

Κατηγοριοποίηση Πράξεων του ΠΑΔΑ στη Διαύγεια

Το ΠΑΔΑ κάνει χρήση:

- 14 ετικετών «Θεματική κατηγορία» από τις 25 διαθέσιμες
- 19 ετικετών «Είδος πράξης» από τις 35 διαθέσιμες

ΘΕΜΑΤΙΚΕΣ ΚΑΤΗΓΟΡΙΕΣ ΠΡΑΞΕΩΝ - ΠΑΔΑ	
ΠΑΡΑΓΩΓΗ, ΤΕΧΝΟΛΟΓΙΑ ΚΑΙ ΕΡΕΥΝΑ	ΟΙΚΟΝΟΜΙΚΗ ΖΩΗ
ΕΠΙΣΤΗΜΕΣ	ΕΝΕΡΓΕΙΑ
ΕΠΙΚΟΙΝΩΝΙΑ ΚΑΙ ΜΟΡΦΩΣΗ	ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΑΝΤΑΓΩΝΙΣΜΟΣ
ΟΙΚΟΝΟΜΙΚΕΣ ΚΑΙ ΕΜΠΟΡΙΚΕΣ ΣΥΝΑΛΛΑΓΕΣ	ΔΑΠΑΝΕΣ ΕΠΙΧ/ΝΩΝ ΦΟΡΕΩΝ ΑΡΘΡΟΥ 10Β Ν 3861/10
ΔΗΜΟΣΙΟΝΟΜΙΚΑ	ΑΠΟΦΑΣΗ ΔΙΑΘΕΣΗΣ ΑΝΟΙΚΤΩΝ ΔΕΔΟΜΕΝΩΝ
ΑΠΑΣΧΟΛΗΣΗ ΚΑΙ ΕΡΓΑΣΙΑ	ΕΥΡΩΠΑΪΚΗ ΈΝΩΣΗ
ΔΗΜΟΣΙΑ ΔΙΟΙΚΗΣΗ	ΥΓΕΙΑ

ΕΙΔΗ ΠΡΑΞΕΩΝ - ΠΑΔΑ	
ΕΓΚΡΙΣΗ ΔΑΠΑΝΗΣ	ΥΠΗΡΕΣΙΑΚΗ ΜΕΤΑΒΟΛΗ
ΑΝΑΛΗΨΗ ΥΠΟΧΡΕΩΣΗΣ	ΠΡΟΚΗΡΥΞΗ ΠΛΗΡΩΣΗΣ ΘΕΣΕΩΝ
ΛΟΙΠΕΣ ΑΤΟΜΙΚΕΣ ΔΙΟΙΚΗΤΙΚΕΣ ΠΡΑΞΕΙΣ	ΙΣΟΛΟΓΙΣΜΟΣ – ΑΠΟΛΟΓΙΣΜΟΣ
ΕΓΚΡΙΣΗ ΠΡΟΥΠΟΛΟΓΙΣΜΟΥ	ΔΙΟΡΙΣΜΟΣ
ΑΝΑΘΕΣΗ ΕΡΓΩΝ/ΠΡΟΜΗΘΕΙΩΝ/ΥΠΗΡΕΣΙΩΝ/ΜΕΛΕΤΩΝ	ΕΓΚΥΚΛΙΟΣ
ΚΑΝΟΝΙΣΤΙΚΗ ΠΡΑΞΗ	ΚΑΤΑΚΥΡΩΣΗ
ΟΡΙΣΤΙΚΟΠΟΙΗΣΗ ΠΛΗΡΩΜΗΣ	ΠΡΑΞΕΙΣ ΧΩΡΟΤΑΞΙΚΟΥ - ΠΟΛΕΟΔΟΜΙΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ
ΠΡΑΞΗ ΠΟΥ ΑΦΟΡΑ ΣΕ ΣΥΛΛΟΓΙΚΟ ΟΡΓΑΝΟ – ΕΠΙΤΡΟΠΗ - ΟΜΑΔΑ ΕΡΓΑΣΙΑΣ - ΟΜΑΔΑ ΕΡΓΟΥ - ΜΕΛΗ ΣΥΛΛΟΓΙΚΟΥ ΟΡΓΑΝΟΥ	ΠΡΑΞΗ ΠΟΥ ΑΦΟΡΑ ΣΕ ΘΕΣΗ ΓΕΝΙΚΟΥ - ΕΙΔΙΚΟΥ ΓΡΑΜΜΑΤΕΑ - ΜΟΝΟΜΕΛΕΣ ΟΡΓΑΝΟ
ΣΥΜΒΑΣΗ	ΔΩΡΕΑ - ΕΠΙΧΟΡΗΓΗΣΗ
ΠΕΡΙΛΗΨΗ ΔΙΑΚΗΡΥΞΗΣ	

Δημιουργία Datasets – Αρχικά Αποτελέσματα

- Δημιουργήθηκαν:

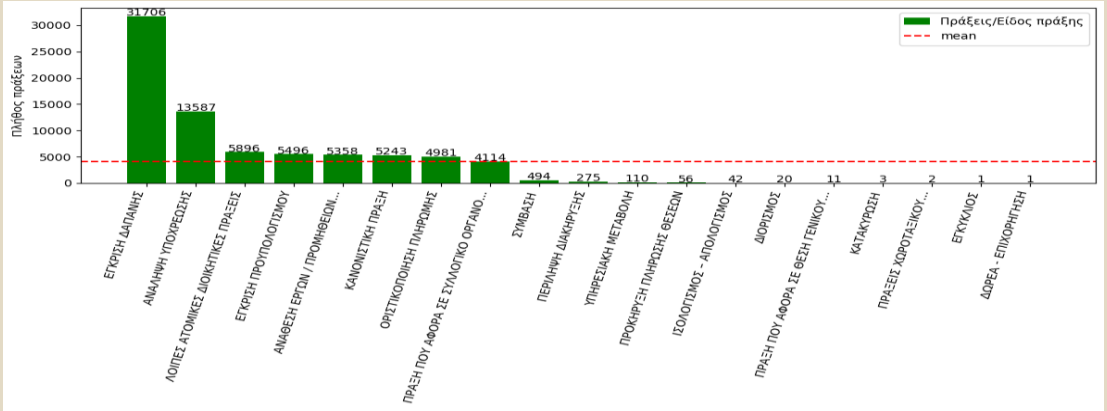
- **Training Dataset** που κάλυπτε την περίοδο: **22-03-2018** έως **21-03-2023**
- **Test Dataset** που κάλυπτε την περίοδο: **22-03-2023** έως **23-06-2023**

- **Εξαιρετικά** καλά αποτελέσματα τόσο στη φάση αξιολόγησης (evaluation) όσο και στη φάση επικύρωσης (validation) της κατηγοριοποίησης της **«Θεματικής κατηγορίας»**

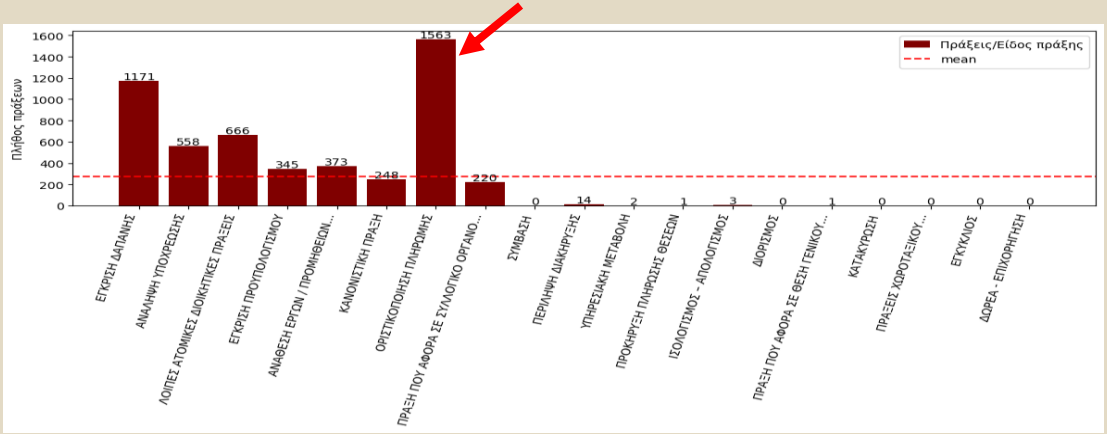
- Πολύ **χαμηλά** αποτελέσματα στη φάση επικύρωσης της κατηγοριοποίησης του **«Είδους πράξης»** σε αντίθεση με τα αποτελέσματα της φάσης αξιολόγησης.
- Αποδόθηκαν στην **ασυμφωνία των στατιστικών** μεταξύ των δομών των Datasets
- Test Dataset->Διάγραμμα κατανομής κλάσεων: κορυφή στην κλάση «Οριστικοποίηση πληρωμής».
- Μάιος και Ιούνιος: το μεγαλύτερο πλήθος πράξεων της συγκεκριμένης κλάσης

Ασυμφωνία Στατιστικών των Datasets – Είδος πράξης

Training Dataset



Test Dataset



Διερευνητικές Ενέργειες

Σκοπός: να διαπιστωθεί αν τα αποτελέσματα επηρεάζονται από τη στατιστική κατανομή των Datasets ή αν δεν αποδίδουν επαρκώς οι μέθοδοι μηχανικής μάθησης

❑ 1η ενέργεια

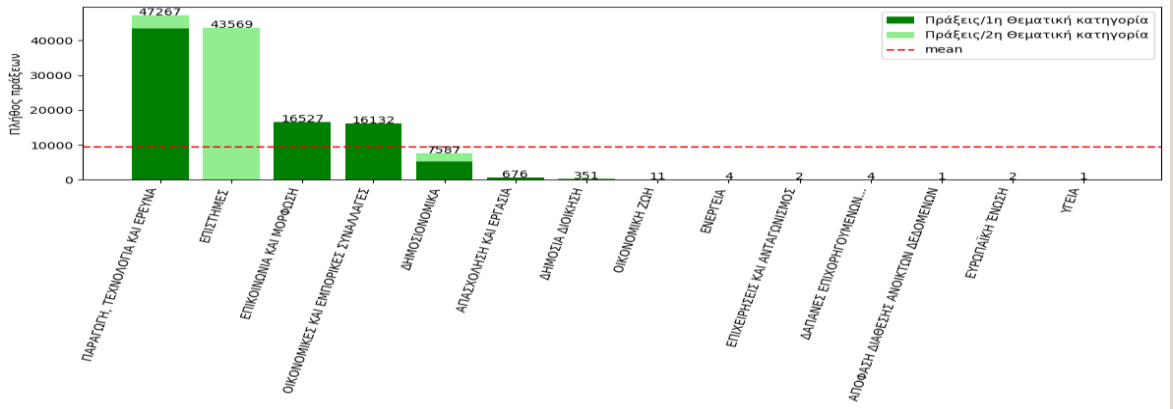
- Δημιουργήθηκαν **δυο νέα** Test Datasets που κάλυπταν περιόδους **μειωμένες** κατά **1 και 2 μήνες** σε σχέση με το αρχικό
- **Διατηρήθηκαν** τα **υψηλά** αποτελέσματα της φάσης επικύρωσης στην **«Θεματική κατηγορία»**
- Παρατηρήθηκε **βελτίωση** των αποτελεσμάτων της φάσης επικύρωσης στο **«Είδος πράξης»**

❑ 2η ενέργεια

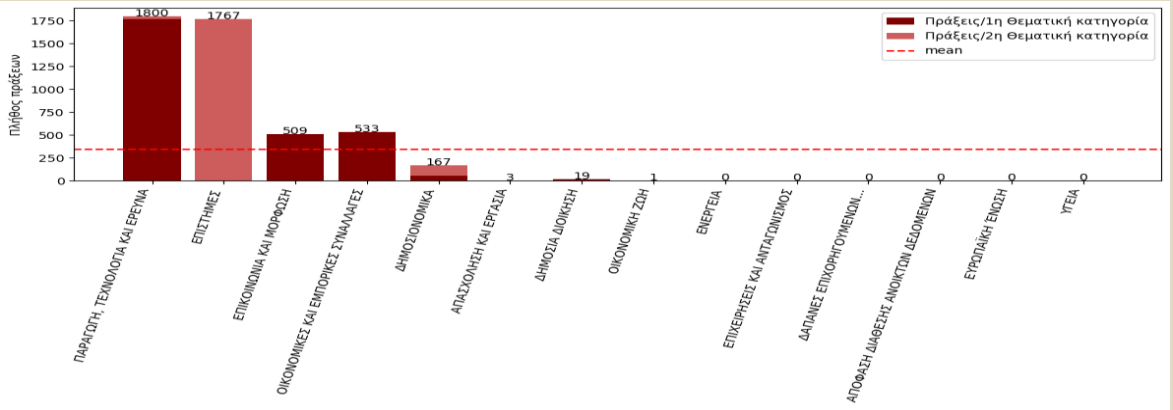
- Δημιουργήθηκε **νέο** Training Dataset από την **συνένωση** των αρχικών Training και Test Datasets
- Δημιουργήθηκε **νέο** Test Dataset που κάλυπτε την περίοδο: **24-06-2023** έως **23-08-2023**

Θεματική Κατηγορία – Training και Test Datasets

Training Dataset
- 82561 πράξεις
- 14 θεματικές κατηγορίες
- το μέγιστο 2 θεματικές κατηγορίες σε κάθε πράξη (multilabel)
- 6 θεματικές < 10 πράξεις
- mean = 9438

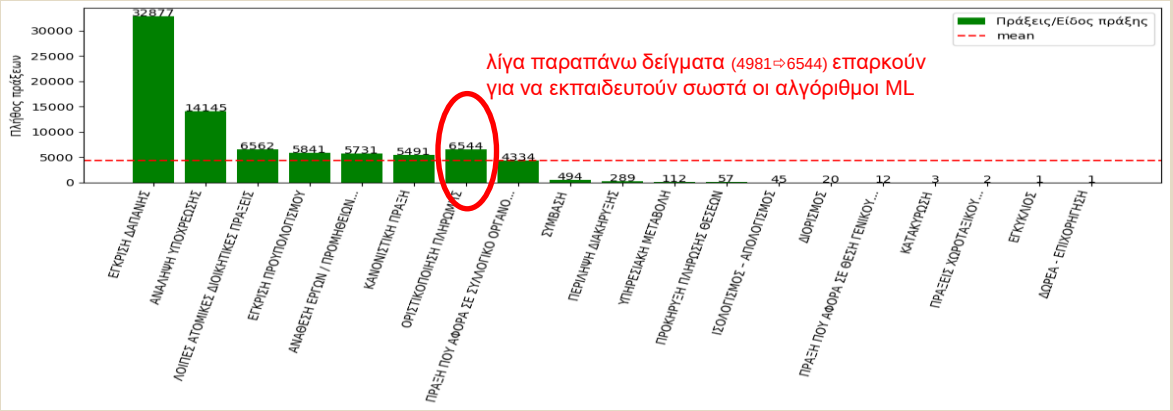


Test Dataset
- 2890 πράξεις
- 6 θεματικές → 0 πράξεις
- mean = 343

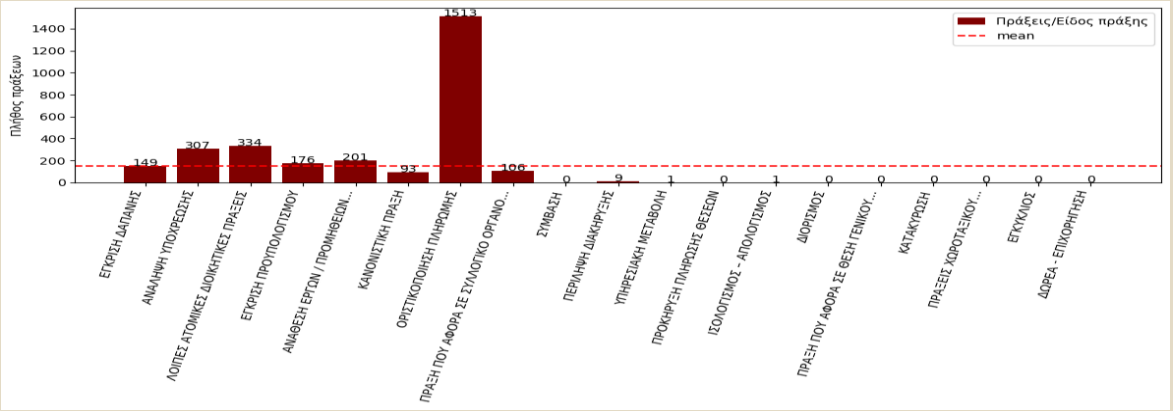


Είδος Πράξης – Training και Test Datasets

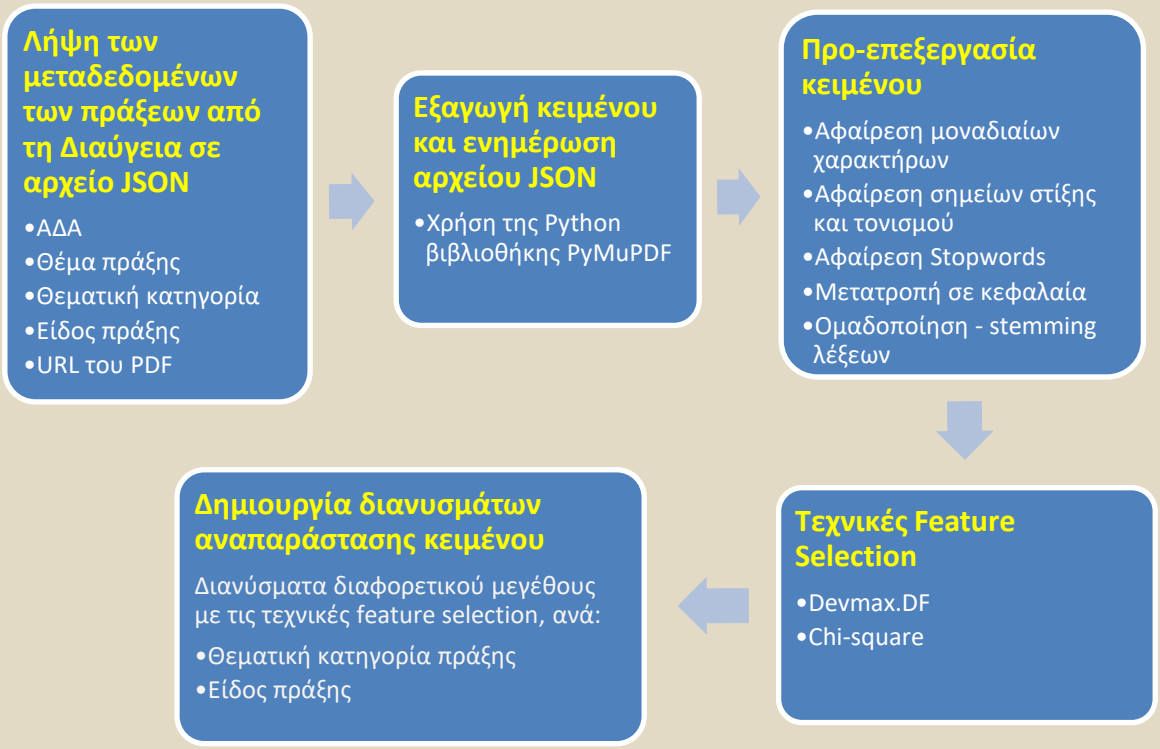
Training Dataset
- 82561 πράξεις
- 19 είδη πράξης
- 1 είδος ανά πράξη (multiclass)
- 4 είδη < 10 πράξεις
- mean = 4345



Test Dataset
- 2890 πράξεις
- 8 είδη → 0 πράξεις
- mean = 152



Δημιουργία Διανυσμάτων Αναπαράστασης Κειμένου



Λήψη Μεταδεδομένων – Εξαγωγή και Προεπεξεργασία Κειμένου

- Η «Διαύγεια» παρέχει **API** για την υποστήριξη των παρεχόμενων λειτουργιών, Βασίζεται στην python βιβλιοθήκη **requests** με κύριο format το **JSON**
- Για την εξαγωγή του κειμένου από το pdf χρησιμοποιήθηκε η βιβλιοθήκη **PyMuPDF**
- Η βιβλιοθήκη **requests** χρησιμοποιήθηκε για την λήψη των pdf τοπικά
- Χρησιμοποιήθηκε η βιβλιοθήκη **Natural Language Toolkit (NLTK)**
- Έγινε χρήση των stop-words της βιβλιοθήκης **spaCy**

Προεπεξεργασία Κειμένου

28,7 εκατομμύρια

λέξεις (tokens)

Αφαίρεση:
- stop-words
- σημείων στίξης
- μοναδιαίων χαρακτήρων, αριθμών

262 χιλιάδες

μοναδικές λέξεις

Grouping

68 χιλιάδες

ομάδες λέξεων

Grouping - Feature Selection

ΒΑΣΙΚΗ ΛΕΞΗ	ΠΑΡΟΜΟΙΕΣ ΛΕΞΕΙΣ
ΤΡΑΠΕΖΑ	ΤΡΑΠΕΖΑ ΤΡΑΠΕΖΙΚΟΣ ΤΡΑΠΕΖΙΚΑ ΤΡΑΠΕΖΑΣ ΤΡΑΠΕΖΙΚΩΝ ΤΡΑΠΕΖΙ ΤΡΑΠΕΖΑΡΙΑ ΤΡΑΠΕΖΙΩΝ
ΣΚΡΕΚΟΥ	ΣΚΡΕΚΟΥ ΣΚΡΕΚΟΥΣ ΣΚΡΕΚΑΣ ΣΚΡΕΚΑ ΣΚΡΕΚΗΣ
ΟΡΙΣΤΙΚΟΣ	ΟΡΙΣΤΙΚΟΣ ΟΡΙΣΤΙΚΟΠΟΙΗΣΗ ΟΡΙΣΤΙΚΑ ΟΡΙΣΤΙΚΩΣ ΟΡΙΣΤΙΚΟΠΟΙΩ ΟΡΙΣΤΙΚΗΣ
ΕΛΕΓΧΘΗΚΕ	ΕΛΕΓΧΘΗΚΕ ΕΛΕΓΧΘΗΚΕΣ ΕΛΕΓΧΘΟΥΝ
ΕΝΤΑΛΜΑ	ΕΝΤΑΛΜΑ ΕΝΤΑΛΜΑΤΩΝ ΕΝΤΑΛΜΑΤΑ ΕΝΤΑΛΜΑΤΟΣ ΕΝΤΑΛΜΑΣ
ΕΠΩΝΥΜΙΑ	ΕΠΩΝΥΜΙΑ ΕΠΩΝΥΜΟΣ ΕΠΩΝΥΜΟ ΕΠΩΝΥΜΟΥ ΕΠΩΝΥ ΕΠΩΝΥΜ ΕΠΩΝΥΜΙ
ΚΑΤΑΧΩΡΗΘΗΚΕ	ΚΑΤΑΧΩΡΗΘΗΚΕ ΚΑΤΑΧΩΡΗΣΗ ΚΑΤΑΧΩΡΗΣΕΩΝ ΚΑΤΑΧΩΡΗΣΕΙΣ ΚΑΤΑΧΩΡΗΝΩ ΚΑΤΑΧΩΡΕΙ
ΕΚΚΑΘΑΡΙΣΗ	ΕΚΚΑΘΑΡΙΣΗ ΕΚΚΑΘΑΡΙΣΤΗ ΕΚΚΑΘΑΡΙΣΤΙΚΟΣ ΕΚΚΑΘΑΡΙΣΕΩΣ ΕΚΚΑΘΑΡΙΣΤΕΣ ΕΚΚΑΘΑΡΙΣΤΡΙΑ
ΕΥΔΟΚΙΜΙΔΗΣ	ΕΥΔΟΚΙΜΙΔΗΣ ΕΥΔΟΚΙΑ ΕΥΔΟΚΙΜΗΣΗ ΕΥΔΟΚΙΜΟΣ ΕΥΔΟΚΙΜΩ ΕΥΔΟΚΙΜΑ
ΕΠΑΓΓΕΛΜΑΤΙΚΗ	ΕΠΑΓΓΕΛΜΑΤΙΚΗ ΕΠΑΓΓΕΛΜΑΤΩΝ ΕΠΑΓΓΕΛΜΑΤΙΑΣ ΕΠΑΓΓΕΛΜΑΤΙΚΟΣ
ΧΡΗΜΑΤΙΚΟ	ΧΡΗΜΑΤΙΚΟ ΧΡΗΜΑΤΙΚΟΣ ΧΡΗΜΑΤΙΚΩΝ ΧΡΗΜΑ ΧΡΗΜΑΤΙΚΟΥ
ΠΑΡΑΣΤΑΤΙΚΟΣ	ΠΑΡΑΣΤΑΤΙΚΟΣ ΠΑΡΑΣΤΑΤΙΚΟ ΠΑΡΑΣΤΑΤΗΣ ΠΑΡΑΣΤΑΤΙΚΑ ΠΑΡΑΣΤΑΤΙΚΟΥ ΠΑΡΑΣΤΑΤΙΚΗΣ
ΑΙΤΙΟΛΟΓΙΑ	ΑΙΤΙΟΛΟΓΙΑ ΑΙΤΙΟΛΟΓΩ ΑΙΤΙΟΛΟΓΗΜΕΝΟΣ ΑΙΤΙΟΛΟΓΕΙΤΑΙ ΑΙΤΙΟ ΑΙΤΙΟΛΟΓΗΜΕΝΑ ΑΙΤΙΟΛΟΓΗΣΗ

Διανύσματα – Bag Of Words

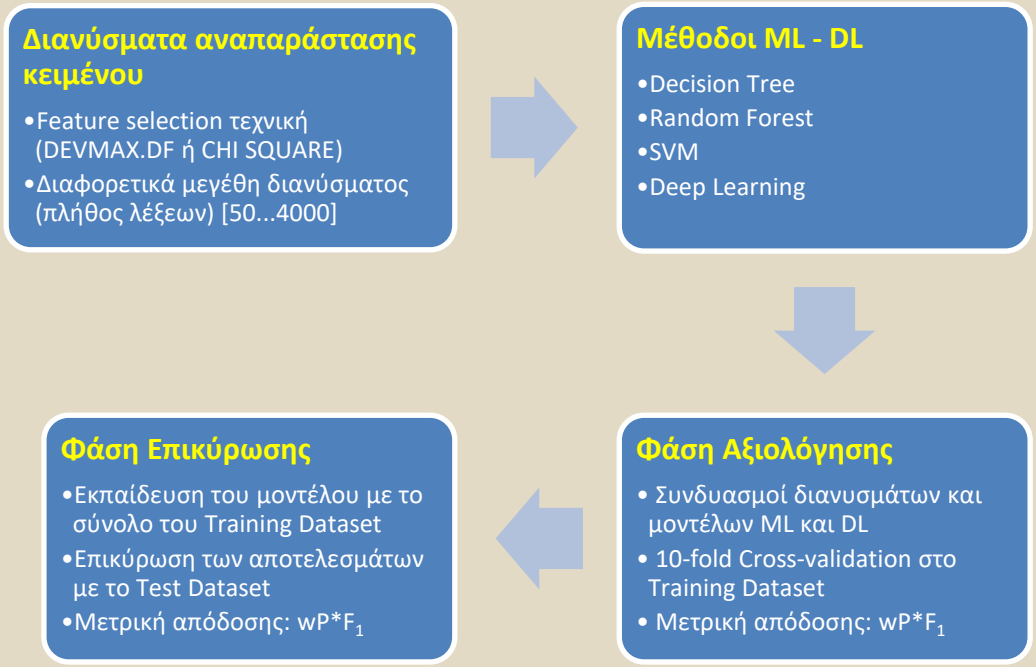
Τμήμα αρχείου διανύσματος αναπαράστασης κειμένου

```
In [8]: df.head(20)
```

Out[8]:

	feature1-ΕΛΚΕ	feature2-ΤΙΤΛΟ	feature3-ΚΩΔΙΚΟΣ	feature4-ΚΑΤΗΓΟΡΙΑ	feature5-ΜΟΔΥ	feature6-ΥΠΕΥΘΥΝΟΥ	feature7-ΧΡΗΜΑΤΟΔΟΤΗΣΗ	feature8-ΕΡΓΟ	feature9-ΕΥΡΩ	feature10-LOCATION
Ω8ΚΧ46Μ9ΞΗ-1ΗΜ	1	1	1	1	1	0	1	1	1	1
6Σ9246Μ9ΞΗ-6ΥΩ	1	1	1	1	1	0	1	1	1	1
ΩΦΝ646Μ9ΞΗ-6ΧΙ	1	1	1	1	1	0	1	1	1	1
Ω1ΞΙ46Μ9ΞΗ-9ΜΥ	0	0	0	0	0	0	0	0	0	1
69ΒΠ46Μ9ΞΗ-ΨΚ6	1	1	1	0	1	0	1	1	1	1
72ΙΨ46Μ9ΞΗ-ΡΙΩ	1	1	1	1	1	0	1	1	1	1
Ψ2ΟΗ46Μ9ΞΗ-ΓΓ7	1	1	1	0	1	0	1	1	1	1
ΨΜ6Μ46Μ9ΞΗ-ΜΟΚ	0	0	0	0	0	0	1	1	1	1
9ΣΓ946Μ9ΞΗ-ΒΣΤ	1	1	1	0	1	0	1	1	1	1
7ΝΑ746Μ9ΞΗ-Β1Μ	1	0	1	0	1	0	0	1	1	1
7Χ3Ι46Μ9ΞΗ-ΦΧ7	1	1	1	0	1	0	1	1	1	1
Ω28Χ46Μ9ΞΗ-606	1	1	1	0	1	0	1	1	1	1

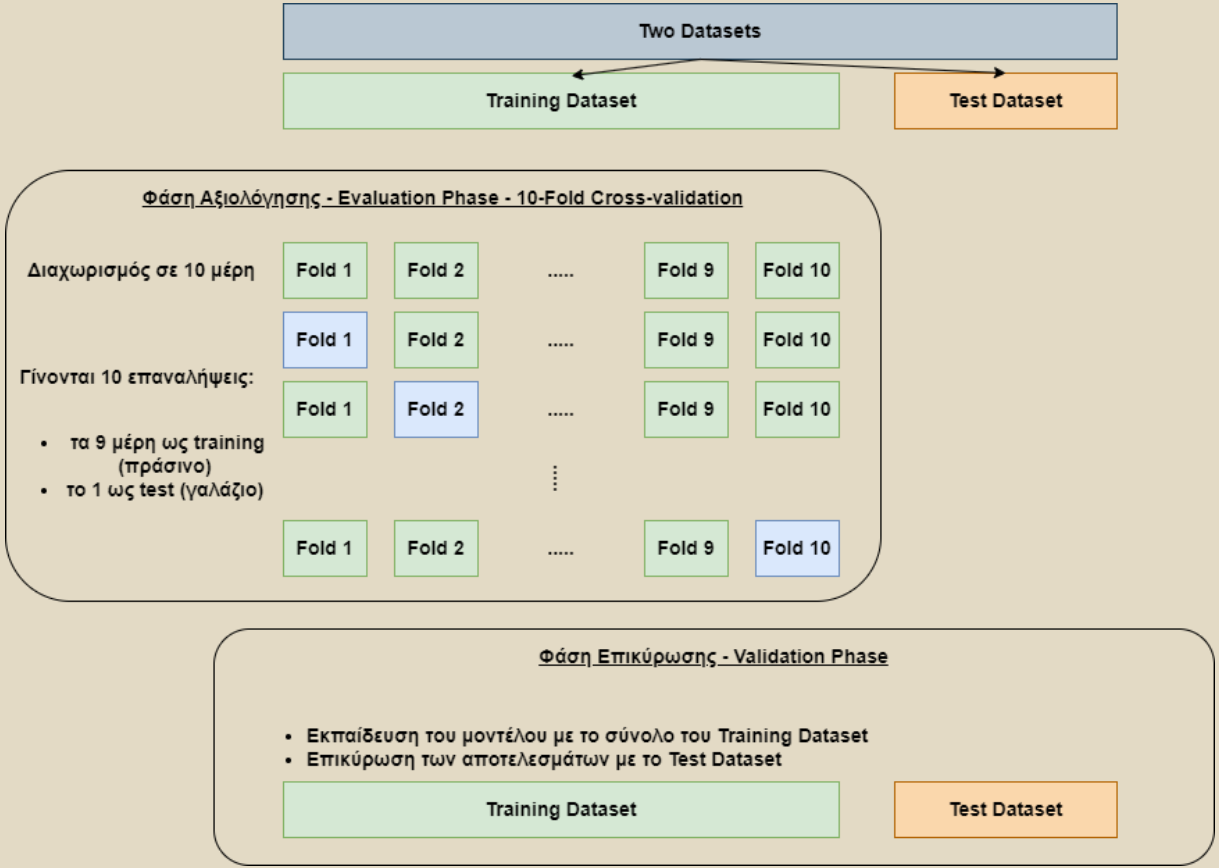
Εφαρμογή Μεθόδων ML



$wP*F_1$:

- Σταθμισμένο (ως προς το μέγεθος των κλάσεων) γινόμενο Precision*F_1 score
- Έξτρα P προσφέρει αυξημένη ευαισθησία στην σωστή εκτίμηση των ορθά ταξινομημένων δειγμάτων

2 Φάσεις: Αξιολόγησης - Επικύρωσης



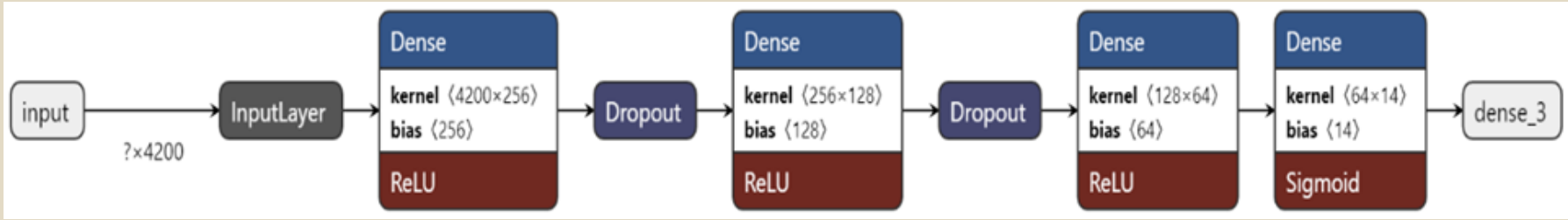
Μοντέλα Μηχανικής και Βαθιάς Μάθησης

- ❑ Decision Tree
- ❑ Random Forest
- ❑ SVM

} python βιβλιοθήκη **Scikit-Learn**

- ❑ Deep Learning
  Dense model
  - απλό feedforward μοντέλο
  - 3 κρυφά layers (ReLU)
  - dropout layer (0.5 rate)

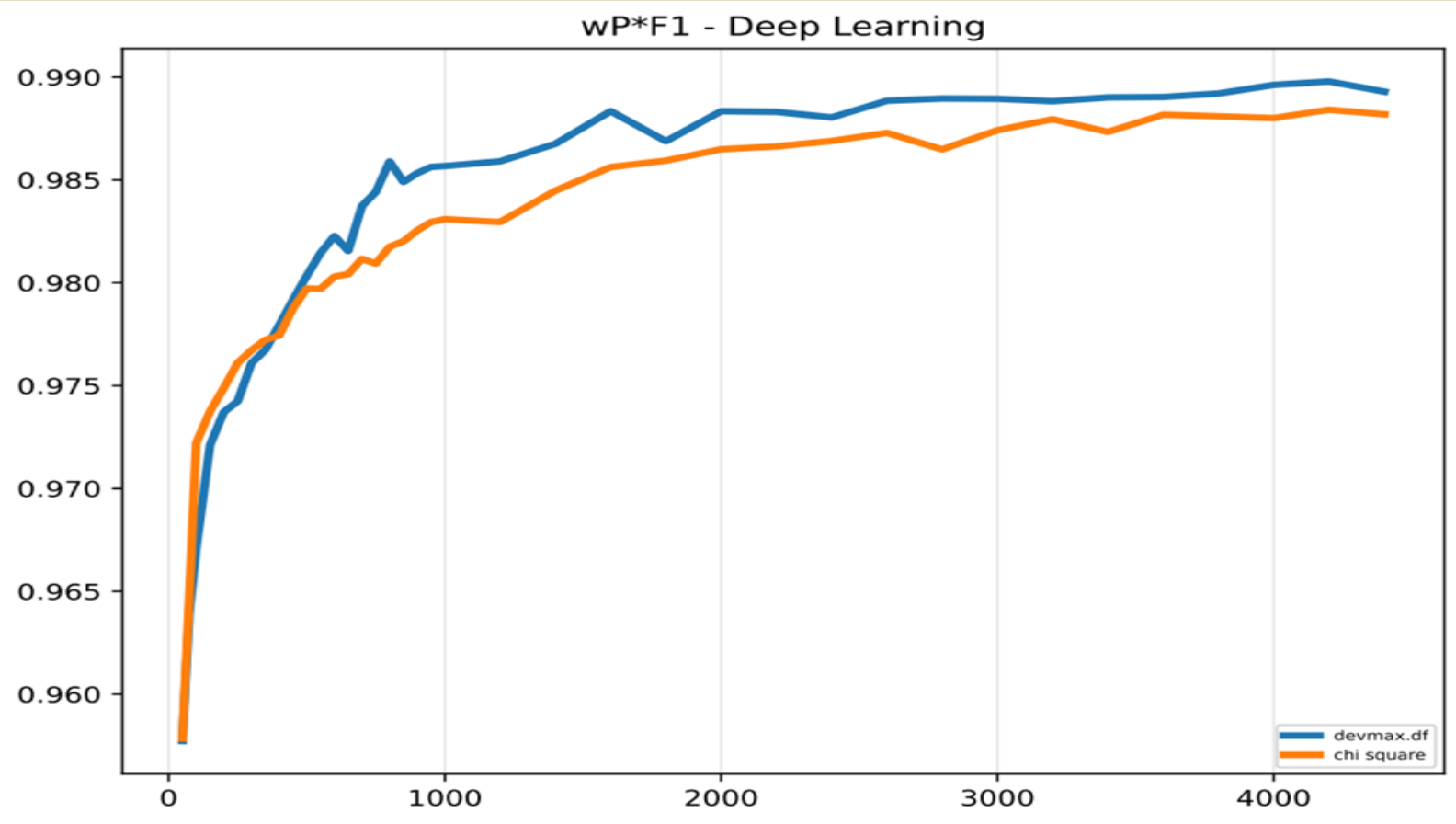
} python βιβλιοθήκη **Keras**



Αποτελέσματα Evaluation - Validation
Θεματική Κατηγορία

			Evaluation Phase Training Dataset (22/03/2018 – 23/06/2023)			Validation Phase Training => Test Dataset (24/06/2023 – 23/08/2023)		
Μοντέλο	**Τεχνική**	**Λέξεις (διάνυσμα)**	**wF_1**	**wP**	**$wP*F_1$**	**wF_1**	**wP**	**$wP*F_1$**
Decision Tree	chi square	3200	98,70	98,66	97,44	97,70	98,31	96,30
Random Forest	chi square	3000	99,02	99,11	98,18	97,17	99,21	96,42
SVM	devmax.df	1400	98,79	98,86	97,71	96,88	98,42	95,39
Deep Learning	devmax.df	4200	99,44	99,42	98,89	98,74	98,50	**97,32**

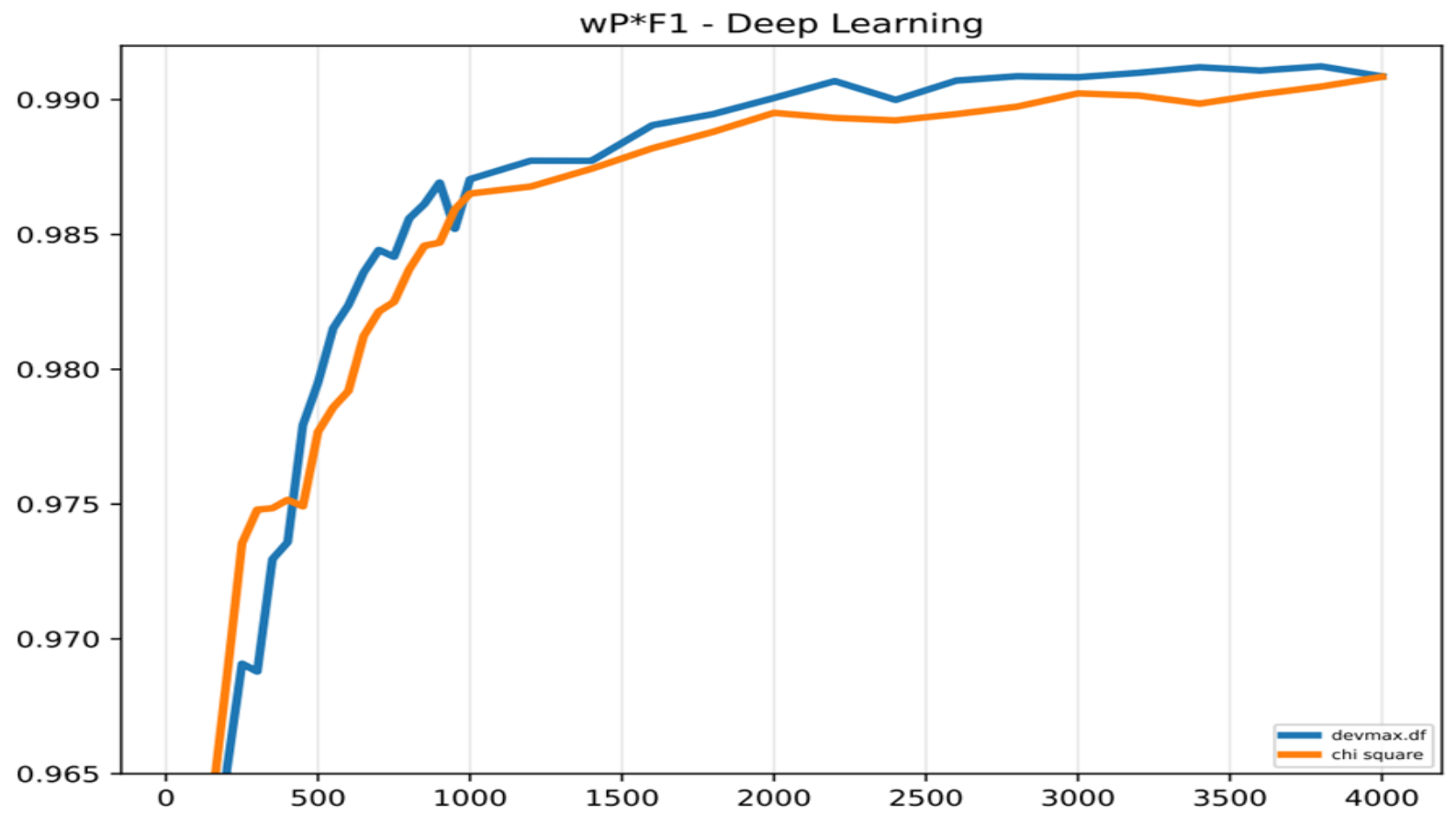
Διάγραμμα wP*F1 – Θεματική κατηγορία



Αποτελέσματα Evaluation - Validation
Είδος Πράξης

			Evaluation Phase Training Dataset (22/03/2018 – 23/06/2023)			Validation Phase Training => Test Dataset (24/06/2023 – 23/08/2023)		
Μοντέλο	**Τεχνική**	**Λέξεις (διάνυσμα)**	**wF_1**	**wP**	**$wP*F_1$**	**wF_1**	**wP**	**$wP*F_1$**
Decision Tree	chi square	1800	98,19	98,21	96,58	92,05	90,42	84,38
Random Forest	chi square	1600	98,87	99,08	98,02	96,18	95,13	92,40
SVM	devmax.df	2600	98,67	98,92	97,69	96,57	95,62	93,21
Deep Learning	devmax.df	3400	99,50	99,52	99,06	98,73	98,83	**97,68**

Διάγραμμα wP*F1 – Είδος πράξης



Πίνακες Ανάλυσης Αποτελεσμάτων ανά Κατηγορία

Θεματική κατηγορία

DEEP LEARNING - DEVMAX.DF - 4200W	TN	FN	TP	FP	Precision	F1	Εκχωρήσεις/ κλάση
ΠΑΡΑΓΩΓΗ, ΤΕΧΝΟΛΟΓΙΑ ΚΑΙ ΕΡΕΥΝΑ	1073	10	1790	17	0.9906	0.9925	1800
ΕΠΙΣΤΗΜΕΣ	1120	12	1755	3	0.9983	0.9957	1767
ΕΠΙΚΟΙΝΩΝΙΑ ΚΑΙ ΜΟΡΦΩΣΗ	2375	3	506	6	0.9883	0.9912	509
ΟΙΚΟΝΟΜΙΚΕΣ ΚΑΙ ΕΜΠΟΡΙΚΕΣ ΣΥΝΑΛΛΑΓΕΣ	2331	2	531	26	0.9533	0.9743	533
ΔΗΜΟΣΙΟΝΟΜΙΚΑ	2701	6	161	22	0.8798	0.9200	167
ΑΠΑΣΧΟΛΗΣΗ ΚΑΙ ΕΡΓΑΣΙΑ	2887	2	1	0	1.0000	0.5000	3
ΔΗΜΟΣΙΑ ΔΙΟΙΚΗΣΗ	2871	8	11	0	1.0000	0.7333	19
ΟΙΚΟΝΟΜΙΚΗ ΖΩΗ	2889	1	0	0	-	-	1
ΕΝΕΡΓΕΙΑ	2890	0	0	0	-	-	0
ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΑΝΤΑΓΩΝΙΣΜΟΣ	2890	0	0	0	-	-	0
ΔΑΠΑΝΕΣ ΕΠΙΧ/ΝΩΝ ΦΟΡΕΩΝ ΑΡΘΡΟΥ 10Β Ν 3861/10	2890	0	0	0	-	-	0
ΑΠΟΦΑΣΗ ΔΙΑΘΕΣΗΣ ΑΝΟΙΚΤΩΝ ΔΕΔΟΜΕΝΩΝ	2890	0	0	0	-	-	0
ΕΥΡΩΠΑΪΚΗ ΈΝΩΣΗ	2890	0	0	0	-	-	0
ΥΓΕΙΑ	2890	0	0	0	-	-	0
Σταθμισμένος ΜΟ:					wP 0.9850	wF1 0.9874	Σύνολο: 4799
					wP*F1 0.9732		

Είδος πράξης

DEEP LEARNING - DEVMAX.DF - 3400W	TN	FN	TP	FP	Precision	F1	Εκχωρήσεις/ κλάση
ΕΓΚΡΙΣΗ ΔΑΠΑΝΗΣ	2733	0	149	8	0.9490	0.9739	149
ΑΝΑΛΗΨΗ ΥΠΟΧΡΕΩΣΗΣ	2583	1	306	0	1.0000	0.9984	307
ΛΟΙΠΕΣ ΑΤΟΜΙΚΕΣ ΔΙΟΙΚΗΤΙΚΕΣ ΠΡΑΞΕΙΣ	2552	24	310	4	0.9873	0.9568	334
ΕΓΚΡΙΣΗ ΠΡΟΥΠΟΛΟΓΙΣΜΟΥ	2714	0	176	0	1.0000	1.0000	176
ΑΝΑΘΕΣΗ ΕΡΓΩΝ/ΠΡΟΜΗΘΕΙΩΝ/ΥΠΗΡΕΣΙΩΝ/ΜΕΛΕΤΩΝ	2689	0	201	0	1.0000	1.0000	201
ΚΑΝΟΝΙΣΤΙΚΗ ΠΡΑΞΗ	2773	4	89	24	0.7876	0.8641	93
ΟΡΙΣΤΙΚΟΠΟΙΗΣΗ ΠΛΗΡΩΜΗΣ	1377	7	1506	0	1.0000	0.9977	1513
ΠΡΑΞΗ ΠΟΥ ΑΦΟΡΑ ΣΕ ΣΥΛΛΟΓΙΚΟ ΟΡΓΑΝΟ – ΕΠΙΤΡΟΠΗ - ΟΜΑΔΑ ΕΡΓΑΣΙΑΣ - ΟΜΑΔΑ ΕΡΓΟΥ - ΜΕΛΗ ΣΥΛΛΟΓΙΚΟΥ ΟΡΓΑΝΟΥ	2782	0	106	2	0.9815	0.9907	106
ΣΥΜΒΑΣΗ	2890	0	0	0	-	-	0
ΠΕΡΙΛΗΨΗ ΔΙΑΚΗΡΥΞΗΣ	2881	1	8	0	1.0000	0.9412	9
ΥΠΗΡΕΣΙΑΚΗ ΜΕΤΑΒΟΛΗ	2889	0	1	0	1.0000	1.0000	1
ΠΡΟΚΗΡΥΞΗ ΠΛΗΡΩΣΗΣ ΘΕΣΕΩΝ	2890	0	0	0	-	-	0
ΙΣΟΛΟΓΙΣΜΟΣ – ΑΠΟΛΟΓΙΣΜΟΣ	2889	0	1	0	1.0000	1.0000	1
ΔΙΟΡΙΣΜΟΣ	2890	0	0	0	-	-	0
ΠΡΑΞΗ ΠΟΥ ΑΦΟΡΑ ΣΕ ΘΕΣΗ ΓΕΝΙΚΟΥ - ΕΙΔΙΚΟΥ ΓΡΑΜΜΑΤΕΑ - ΜΟΝΟΜΕΛΕΣ ΟΡΓΑΝΟ	2890	0	0	0	-	-	0
ΚΑΤΑΚΥΡΩΣΗ	2890	0	0	0	-	-	0
ΠΡΑΞΕΙΣ ΧΩΡΟΤΑΞΙΚΟΥ - ΠΟΛΕΟΔΟΜΙΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ	2890	0	0	0	-	-	0
ΕΓΚΥΚΛΙΟΣ	2890	0	0	0	-	-	0
ΔΩΡΕΑ - ΕΠΙΧΟΡΗΓΗΣΗ	2890	0	0	0	-	-	0
Σταθμισμένος ΜΟ:					wP 0.9883	wF1 0.9872	Σύνολο: 2890
					wP*F1 0.9768		

Συμπεράσματα

- Με τα αρχικά Datasets, οι αποδόσεις των μοντέλων στις φάσεις αξιολόγησης και επικύρωσης της κατηγοριοποίησης της «Θεματικής κατηγορίας» ήταν εξαιρετικά υψηλές.
- Αντίθετα η προσπάθεια κατηγοριοποίησης βάσει του «Είδους πράξης» πέτυχε πολύ χαμηλές αποδόσεις στη φάση της επικύρωσης παρά τις υψηλές αποδόσεις που είχαν επιτευχθεί στη φάση της αξιολόγησης.
- Ως αιτία των χαμηλών αποδόσεων θεωρήθηκε η στατιστική ασυνέπεια μεταξύ των δομών των Training και Test Dataset.
- Αρχικά έγιναν προσπάθειες βελτίωσης των αποτελεσμάτων με Test Dataset που κάλυπταν διαφορετικές χρονικές περιόδους με τα αποτελέσματα να βελτιώνονται αισθητά. Στη συνέχεια δημιουργήθηκε νέο Training Dataset, το οποίο προέκυψε από τη συγχώνευση των παλιών Training και Test Datasets, καθώς και νέο Test Dataset που κάλυπτε επόμενη χρονική περίοδο.
- Με την δημιουργία των νέων Datasets, έγινε εκ νέου αξιολόγηση και επικύρωση των αλγορίθμων τόσο για το Είδος πράξης, που παρουσίασε με τα αρχικά Datasets χαμηλές αποδόσεις, όσο και για την Θεματική κατηγορία.

Συμπεράσματα

- Για την κατηγοριοποίηση της Θεματικής κατηγορίας τα υψηλότερα αποτελέσματα στην φάση της επικύρωσης επιτεύχθηκαν με την μετρική Devmax.df και το Deep Learning μοντέλο με την μετρική wP*F1 να επιτυγχάνει ποσοστό 97,32%.
- Για την κατηγοριοποίηση του Είδους πράξης τα υψηλότερα αποτελέσματα στην φάση της επικύρωσης επιτεύχθηκαν με την μετρική Devmax.df και το Deep Learning μοντέλο με την μετρική wP*F1 να επιτυγχάνει ποσοστό 97,68%.
- Τα καλύτερα αποτελέσματα και στις δυο φάσεις αξιολόγησης και επικύρωσης επιτεύχθηκαν με την μετρική **Devmax.df** και το **Deep Learning** μοντέλο τόσο στην κατηγοριοποίηση της «Θεματικής κατηγορίας» όσο και στο «Είδος πράξης»

				Evaluation Phase			**Validation Phase**		
Κατηγορία	**Μοντέλο**	**Τεχνική**	**Λέξεις (διάνυσμα)**	**wF_1**	**wP**	**$wP*F_1$**	**wF_1**	**wP**	**$wP*F_1$**
Θεματική κατηγορία	Deep Learning	devmax.df	4200	99,44	99,42	**98,89**	98,74	98,50	**97,32**
Είδος πράξης	Deep Learning	devmax.df	3400	99,50	99,52	**99,06**	98,73	98,83	**97,68**

Συμπεράσματα

- Η υπολογιστική αρχειακή επιστήμη αντιπροσωπεύει ένα αναπτυσσόμενο επιστημονικό πεδίο που ενσωματώνει τις παραδοσιακές πρακτικές της αρχειακής επιστήμης με τις τεχνολογικές εξελίξεις στον τομέα των υπολογιστών και της ανάλυσης μεγάλων δεδομένων.
- Στο πλαίσιο της διαχείρισης πανεπιστημιακών αρχείων, αυτός ο αναδυόμενος κλάδος έχει τη δυνατότητα να φέρει επανάσταση στον τρόπο με τον οποίο τα εκπαιδευτικά ιδρύματα χειρίζονται, επεξεργάζονται και διατηρούν τις τεράστιες αποθήκες διοικητικών, ακαδημαϊκών και ιστορικών δεδομένων τους.
- Αξιοποιώντας τη δύναμη των υπολογιστικών μεθόδων επεξεργασίας φυσικής γλώσσας και μηχανικής μάθησης τα πανεπιστήμια μπορούν να εξορθολογήσουν τις διαδικασίες διαχείρισης αρχείων τους, να βελτιώσουν την ανάκτηση πληροφοριών και να διευκολύνουν την αποτελεσματική ανάλυση πολύπλοκων αρχειακών συλλογών.
- Η παρούσα διπλωματική ανέδειξε τις αξιοσημείωτες δυνατότητες αυτοματοποιημένης ταξινόμησης και κατηγοριοποίησης διαφορετικών τύπων αρχειακών εγγράφων του ΠΑΔΑ, λειτουργώντας πιλοτικά, με τελικό στόχο και όφελος την απλοποίηση του περίπλοκου έργου της οργάνωσης και ευρετηρίασης μεγάλων δεδομένων ψηφιακών εγγράφων.

Ευχαριστώ για την προσοχή σας!

-

Ερωτήσεις;