ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΟΛΟΓΙΑΣ ΑΥΤΟΜΑΤΗΣ ΑΝΑΓΝΩΡΙΣΗΣ ΓΕΩΓΡΑΦΙΚΟΥ ΙΔΙΩΜΑΤΙΣΜΟΥ ΤΟΥ ΣΥΓΓΡΑΦΕΑ ΣΕ ΣΥΛΛΟΓΗ ΚΕΙΜΕΝΩΝ ΑΠΟ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Σιμάκης Παναγιώτης

Πανεπιστήμιο Πατρών
Πολυτεχνική Σχολή
Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

simakis@ceid.upatras.gr

8 Μαρτίου 2021

Περιεχόμενα

Εισαγωγή

Η ραγδαία ανάπτυξη των μέσων κοινωνικής δικτύωσης και συνολικά του παγκόσμιου ιστού έχουν δημιουργήσει τεράστιο όγκο δεδομένων. Η ίδια του η διαχείριση και η εξαγωγή πληροφορίας αποτελεί καθημερινή πρόκληση. Η επιλογή και του γραπτού λόγου για την έκφραση των χρηστών δημιουργεί ακόμα περισσότερες προκλήσεις ως προς την εξαγωγή γνώσεις από το μεγάλο όγκο δεδομένων που αυξάνεται συνεχώς.

Αντικείμενο της παρούσας εργασίας είναι η ανάπτυξη μεθοδολογίας αυτόματης αναγνώρισης γεωγραφικού ιδιωματισμού του συγγραφέα μέσα από συλλογή κειμένων από μέσα κοινωνικής δικτύωσης. Δηλαδή έχοντας ως είσοδο ένα σύνολο κειμένων, το σύστημα αυτό θα είναι σε θέση να δώσει στην έξοδο την γεωγραφική προέλευση του συγγραφέα.

Βασικές Έννοιες

Εξόρυξης Γνώσης

- Συσταδοποίηση
- Κανόνες συσχέτισης
- Κατηγοριοποίηση

Κατηγοριοποίηση Κειμένου (Text Categorization)

- Ευρετηριοποίηση Κειμένου (Document Indexing)
- Εκπαίδευση κατηγοριοποιητή

Βασικές Έννοιες

- Αναγνώριση Συγγραφέα (Authorship Attribution)
- Αναγνώριση Γεωγραφικού Ιδιωματισμού

Εργαλεία Υλοποίησης

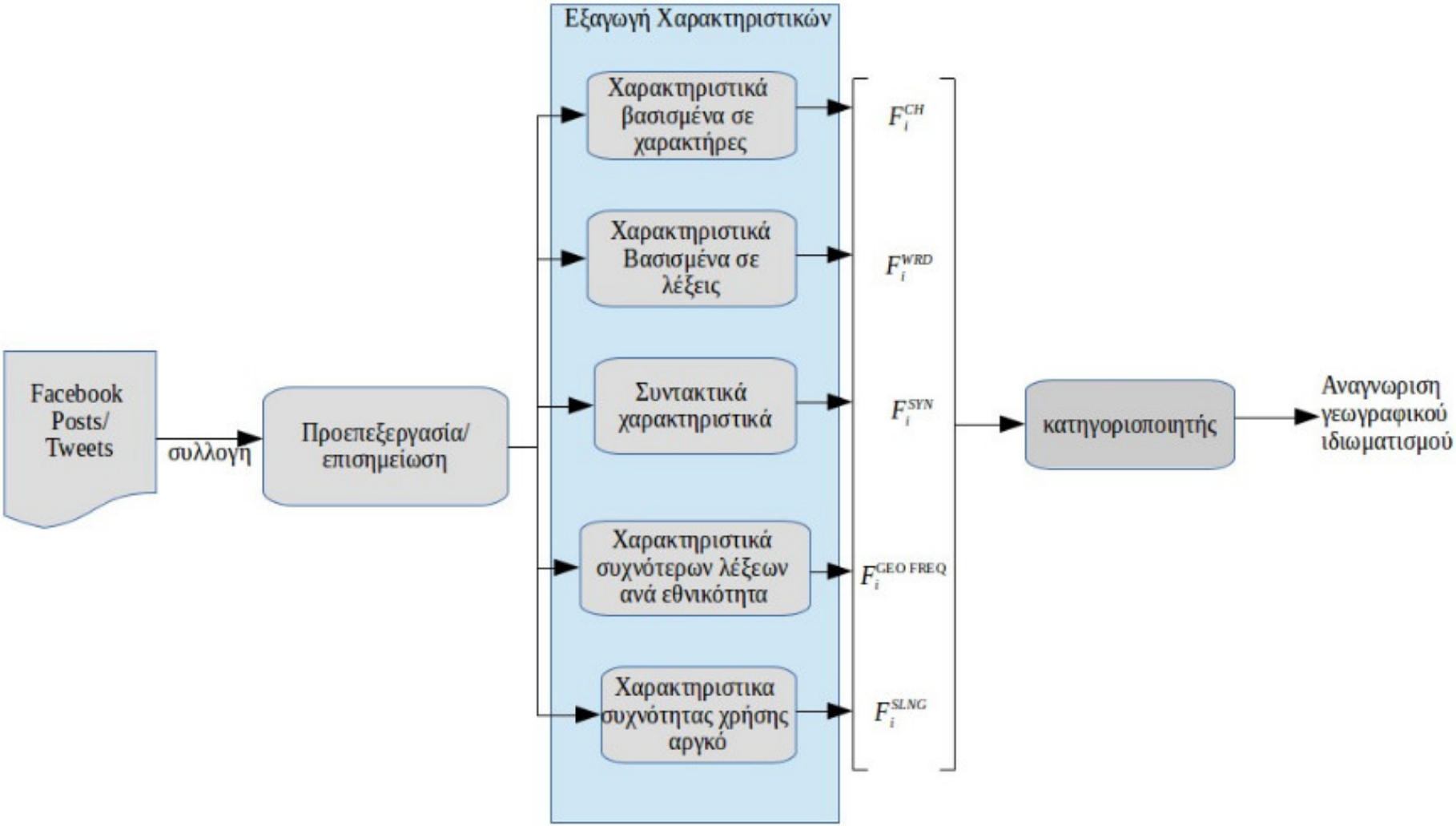
η υλοποίηση για την εξαγωγή των χαρακτηριστικών έγινε με χρήση της Python

- NLTK (Natural Language ToolKit)
- WEKA (Waikato Environment for Knowledge Analysis)

Μεθοδολογία

για την υλοποίηση της πειραματικής διαδικασίας ακολουθήθηκε η εξής μεθοδολογία:

1. Συλλογή Δεδομένων
2. Προεπεξεργασία/επισημείωση δεδομένων
3. Εξαγωγή Χαρακτηριστικών
4. Επιλογή Χαρακτηριστικών
5. Πειράματα κατηγοριοποίησης
6. Συμπεράσματα


Εξαγωγή Χαρακτηριστικών
Facebook Posts/ Tweets
συλλογη
Προεπεξεργασία/ επισημείωση
Χαρακτηριστικά βασισμένα σε χαρακτήρες
Χαρακτηριστικά Βασισμένα σε λέξεις
Συντακτικά χαρακτηριστικά
Χαρακτηριστικά συχνότερων λέξεων ανά εθνικότητα
Χαρακτηριστικα συχνότητας χρήσης αργκό
F_i^{CH}
F_i^{WRD}
F_i^{SYN}
$F_i^{GEO\ FREQ}$
F_i^{SLNG}
κατηγοριοποιητής
Αναγνωριση γεωγραφικού ιδιωματισμού

Συλλογή Δεδομένων

η συλλογή κειμένων έγινε μέσα από προσωπικούς λογαριασμούς χρηστών από το Facebook και το Twitter

- Facepager: το εργαλείο συλλογής των κειμένων
- 252.112 κείμενα, 357 διαφορετικοί συγγραφείς
- 34.969.115 χαρακτηρες και 5.372.512 λέξεις

Προεπεξεργασία/Επισημείωση Δεδομένων

ακριβές πρότυπο επισημειωμένου αρχείου

id	Κείμενο	Φύλο	Ηλικιακη κατηγορία	Ακριβής ηλικία	Κοινωνικό δίκτυο	Θεματική περιγραφή	εθνικότητα	Επιπλέον πληροφορίες
1		F/M	A/B/C/ D/E/F*	> 14	Facebook/ Twitter		US/CAN/ UK/AUS/ NNS	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Εξαγωγή Χαρακτηριστικών

- Βασισμένα σε χαρακτήρες (46 χαρακτηριστικά)
- Βασισμένα σε λέξεις (8 χαρακτηριστικά)
- Συντακτικά (13 χαρακτηριστικά)
- Βασισμένα στο περιεχόμενο (9 χαρακτηριστικά)
  - Χαρακτηριστικά συχνότερων λέξεων ανα εθνικότητα
  - Χαρακτηριστικά συχνότητας χρήσης αργκό ανα εθνικότητα

Επιλογή Χαρακτηριστικών (Feature Selection)

α/α	ReliefF score	Χαρακτηριστικό
1	0.00329917	Συνολικός αριθμός κεφαλαίων χαρακτήρων ανα αριθμό χαρακτήρων
2	0.00263219	Συνολικός αριθμός μικρών λέξεων ανα σύνολο λέξεων
3	0.00190155	Συνολικός αριθμός κενών χαρακτήρων
4	0.00181929	Μέσο μήκος λέξης
5	0.00175353	Συνολικός αριθμός αλφαβητικών χαρακτήρων ανα σύνολο χαρακτήρων
6	0.00149185	Συνολικός αριθμός σημείων στίξης ανα σύνολο χαρακτήρων
7	0.00132278	Συνολικός αριθμός χαρακτήρων στις λέξεις ανα σύνολο χαρακτήρων
8	0.00108097	Συνολικός αριθμός συμβόλων ανα σύνολο χαρακτήρων
9	0.00095305	Μέσο μήκος πρότασης ανα σύνολο χαρακτήρων
10	0.00086441	Συνολικός αριθμός ψηφίων ανα σύνολο χαρακτήρων

Πίνακας: Πίνακας κατάταξης των πρώτων 10 χαρακτηριστικών

Αποτελέσματα Κατηγοριοποίησης

	top10	top20	top30	top40	top50	top60	all
J48	52.28	**53.88**	52.24	51.23	50.31	51.26	51.30
MLP	47.17	51.16	**51.20**	51.11	50.07	49.36	45.01
RandomTree	**55.58**	55.20	50.66	46.01	41.53	40.62	40.52
REPTree	50.45	53.18	53.22	52.87	52.01	52.87	52.88
RBFNetwork	46.70	**49.30**	48.89	48.00	48.53	47.91	48.44
Bagging	57.96	**61.87**	59.61	58.34	57.50	58.34	58.35
Boosting	46.71	47.62	47.62	47.62	**47.78**	47.62	47.62
IBk	**56.34**	49.08	46.52	41.57	39.08	39.05	38.88

Πίνακας: Αποτελέσματα κατηγοριοποίησης

Συμπεράσματα/Προεκτάσεις

- Εξαγωγή εξειδικευμένων χαρακτηριστικών για αναγνώριση γεωγραφικού ιδιωματισμού
  - Βελτιστοποίηση/Εξέλιξη
  - Βελτίωση των ποσοστών κατηγοριοποίησης
- Περαιτέρο έρευνα στο πεδίο της αυτόματης αναγνώρισης συγγραφέα βάσει γεωγραφικού προσδιορισμού

Ερωτήσεις

Άδεια Χρήσης