Ανάκτηση Πληροφοριών και Εξόρυξη Δεδομένων

2η Εργασία

Ταξινόμηση και Συσταδοποίηση Κειμένων

Βογιατζή Αμαλία, 56542

Γκαντίδης Χρήστος, 56483

GitHub Project Link: https://github.com/christosg88/ml_geopardy

Ανάλυση και Περιγραφή του Dataset

Αρχικό Dataset

- Μορφή του dataset: το dataset που επιλέξαμε αποτελείται από ερωτήσεις του παιχνιδιού ερωτήσεων Jeopardy και βρίσκεται σε μορφή json αρχείου. Για κάθε ερώτηση έχουμε τα εξής πεδία:

    1. ημερομηνία εμφάνισης
    2. σωστή απάντηση
    3. κατηγορία
    4. ερώτηση
    5. γύρος
    6. αριθμός του τηλεοπτικού παιχνιδιού
    7. χρηματική αξία ερώτησης

- Πλήθος ερωτήσεων : το πλήθος των ερωτήσεων είναι ίσο με 216930 και χωρίζονται σε 27995 κατηγορίες.

    Επειδή το πλήθος των κατηγοριών ήταν πολύ μεγάλο για τον σκοπό αυτής της εργασίας και επειδή το πλήθος των ερωτήσεων σε ορισμένες κατηγορίες ήταν πολύ μικρό (κατηγορίες με μόνο 1 ερώτηση), επιλέξαμε ένα υποσύνολο των αρχικών κατηγοριών και συνενώσαμε παρόμοιες κατηγορίες μεταξύ τους.

Επεξεργασμένο Dataset

- Μορφή του dataset: το dataset μετά την επεξεργασία έχει όμοια μορφή με το αρχικό dataset αλλά αποτελείται από μικρότερο πλήθος ερωτήσεων.

- Πλήθος ερωτήσεων : το πλήθος των ερωτήσεων είναι ίσο με 17959 και χωρίζονται σε 10 κατηγορίες.

- Κατηγορίες ερωτήσεων και πλήθος ερωτήσεων ανά κατηγορία

  Οι κατηγορίες που περιλαμβάνει το επεξεργασμένο dataset είναι:

  1. Geography με 3206 ερωτήσεις
  2. Science & Nature με 2901 ερωτήσεις
  3. People με 2311 ερωτήσεις
  4. Literature με 2021 ερωτήσεις
  5. History με 1925 ερωτήσεις
  6. Grammar με 1859 ερωτήσεις
  7. Art με 1401 ερωτήσεις
  8. Music με 1379 ερωτήσεις
  9. Food με 614 ερωτήσεις
  10. Sports με 342 ερωτήσεις

- Πλήθος λέξεων και συχνότητα λέξεων

  Μετά από το stemming έχουμε συνολικά 19015 διαφορετικές λέξεις. Οι 10 πιο συχνά εμφανιζόμενες λέξεις είναι:

  1. 'name': 1709,
  2. 'citi': 1076,
  3. 'one': 1047,
  4. 'first': 964,
  5. 'countri': 778,
  6. 'capit': 750,
  7. 'call': 722,
  8. 'island': 627,
  9. 'state': 607,
  10. 'us': 530,

Ως χαρακτηριστικό για την ταξινόμηση και την συσταδοποίηση επιλέγεται η κάθε λέξη που προκύπτει από το stemming.

Ταξινόμηση (Classification)

Αρχικά, χωρίζουμε το dataset σε δύο μέρη με το training set να περιλαμβάνει το 80% των ερωτήσεων και το test set το υπόλοιπο 20%. Σε κάθε εκτέλεση, μπερδεύεται το dataset ώστε να επιλέγονται τυχαία δείγματα (ερωτήσεις) που θα περιλαμβάνονται στο κάθε σύνολο. Στη συνέχεια, περνάμε το training set από έναν CountVectorizer, ο οποίος δημιουργεί έναν πίνακα vocabulary με γραμμές όσες οι ερωτήσεις στο training set και στήλες όσες το πλήθος των λέξεων. Έτσι, στην θέση (i,j) του vocabulary έχουμε την απόλυτη συχνότητα εμφάνισης της λέξης j στην ερώτηση i. Έπειτα, περνάμε το vocabulary από τον TfidfTransformer και παίρνουμε την σχετική συχνότητα εμφάνισης των λέξεων. Κατόπιν, εκπαιδεύουμε δύο

ταξινομητές (classifiers): έναν που βασίζεται στον Naive Bayes, και συγκεκριμένα έναν πολυωνυμικό Naive Bayes (MultinomialNB) και έναν που χρησιμοποιεί Supprt Vector Machine (SVM), και πιο ειδικά Linear Support Vector Classifier (Linear SVC). Μετά, μετασχηματίζουμε με τον CountVectorizer και το TfidfTransformer το test set και προβλέπουμε με τους classifiers την κατηγορία στην οποία ανήκει η κάθε ερώτηση στο test set. Συγκρίνοντας τις τιμές που προβλέφθηκαν με τις πραγματικές κατηγορίες του test set προκύπτουν οι ακόλουθες μετρικές επιτυχίας.

Μετρική	***Naive Bayes %***	***SVM %***
Accuracy	64.5	69.7
Precision	74.2	69.8
Recall	53.5	68.7
F1 Score	56.1	69.1

Συσταδοποίηση (Clustering)

Αρχικά, εφαρμόζουμε K-Means Clustering σε κάθε κατηγορία ερωτήσεων ξεχωριστά, με αριθμό clusters 1, ώστε να βρούμε το κέντρο της κάθε κατηγορίας. ώστε να επιλέγονται τυχαία δείγματα (ερωτήσεις) που θα περιλαμβάνονται στο κάθε σύνολο. Στη συνέχεια, περνάμε το dataset από έναν CountVectorizer, ο οποίος δημιουργεί έναν πίνακα vocabulary με γραμμές όσες οι ερωτήσεις στο training set και στήλες όσες το πλήθος των λέξεων. Έτσι, στην θέση (i,j) του vocabulary έχουμε την απόλυτη συχνότητα εμφάνισης της λέξης j στην ερώτηση i. Έπειτα, περνάμε το vocabulary από τον TfidfTransformer και παίρνουμε την σχετική συχνότητα εμφάνισης των λέξεων.

Στη συνέχεια, εφαρμόζουμε τον K-Means στο dataset, με σκοπό την δημιουργία K = 10 clusters, χρησιμοποιώντας ως αρχικούς κεντροειδείς τα κέντρα των κλάσεων που υπολογίστηκαν στο προηγούμενο βήμα. Έπειτα, για κάθε ερώτηση προβλέπουμε την κατηγορία που την ταξινόμησε ο K-Means και τις συγκρίνουμε με τις πραγματικές κατηγορίες. Από την παραπάνω διαδικασία εξάγονται οι ακόλουθες μετρικές επιτυχίας.

Μετρική	***K-Means %***
F1 Score	37.4
Adjusted Rand Score	6.6
Adjusted Mutual Info Score	18.7
Homogeneity Score	25.3
Completeness Score	18.8
V Measure Score	21.6

Από τις παραπάνω μετρικές, είναι κατανοητό ότι το συγκεκριμένο dataset δεν προσφέρεται για συσταδοποίηση λόγω του μεγάλου πλήθους διαστάσεων, δηλαδή πολλές διαφορετικές λέξεις και οι ερωτήσεις που ανήκουν στην ίδια κατηγορία δεν συσχετίζονται εύκολα μεταξύ τους με αποτέλεσμα να μη γίνεται σωστή ομαδοποίηση.