Τεχνικές Εξόρυξης Δεδομένων Μεγάλης Κλίμακας (Μ118)
Χειμερινό Εξάμηνο 2018-2019

Ημερομηνία παράδοσης: 10/02/2019
Ομαδική Εργασία (2 Ατόμων)

Σκοπος της εργασίας

Σκοπός της εργασίας είναι η εξοικείωσή σας με τα βασικά στάδια της διαδικασίας που ακολουθούνται για την εφαρμογή τεχνικών εξόρυξης δεδομένων, ήτοι: συλλογή, προ-επεξεργασία / καθαρισμός, μετατροπή, εφαρμογή τεχνικών εξόρυξης δεδομένων και αξιολόγηση. Η υλοποίηση θα γίνει στην γλώσσα προγραμματισμού *Python* με την χρήση του εργαλείου *SciKit Learn* και της βιβλιοθήκης *gensim*.

Περιγραφή

Η εργασία σχετίζεται με την κατηγοριοποίηση δεδομένων κειμένου από ειδησεογραφικά άρθρα. Τα Dataset είναι αρχεία CSV του οποίου τα πεδία είναι διαχωρισμένα με τον χαρακτήρα '\t'(TAB). Περιέχονται δυο αρχεία:

1. train_set.csv (12267 στοιχεία): Το αρχείο αυτό θα χρησιμοποιηθεί για να εκπαιδεύσετε τους αλγόριθμους σας και περιέχει τα εξής πεδία:
    a. Id: Ένας unique αριθμός για το άρθρο
    b. Title: Ο τίτλος του άρθρου
    c. Content: Το περιεχόμενο του άρθρου
    d. Category: Η κατηγορία στην οποία ανήκει το άρθρο
2. test_set.csv (3068 στοιχεία): Το αρχείο αυτό θα χρησιμοποιηθεί για να κάνετε προβλέψεις για νέα δεδομένα. Περιέχει όλα τα πεδία του αρχείου εκπαίδευσης εκτός από το πεδίο 'Category'. Το πεδίο αυτό θα κληθείτε να το εκτιμήσετε χρησιμοποιώντας αλγόριθμους κατηγοριοποίησης.

Οι κατηγορίες των άρθρων είναι 5 και παρουσιάζονται στον παρακάτω πίνακα.

Business
Film
Football
Politics
Technology

Δημιουργία WordCloud

Στο σημείο αυτό καλείστε να δημιουργήσετε ένα Wordcloud για τις πέντε κατηγορίες άρθρων με τα περισσότερα άρθρα. Για την δημιουργία ενός WordCloud θα χρησιμοποιείστε όλα τα άρθρα κάθε κατηγορίας. Παράδειγμα ενός WordCloud παρουσιάζεται στην ακόλουθη εικόνα. Για την δημιουργία του WordCloud μπορείτε να χρησιμοποιήσετε όποια βιβλιοθήκης της Python επιθυμείτε.

Εντοπισμός duplicates

Στο ερώτημα αυτό θα πρέπει να εντοπίσετε παρόμοια κείμενα. Πιο συγκεκριμένα η ομοιότητα ανάμεσα σε δύο κείμενα θα μετρηθεί με cosine similarity ανάμεσα στα term vevctors κάθε κειμένου. Όποιος επιθυμεί μπορεί να χρησιμοποιήσει την τεχνική LSH για πιο γρήγορο εντοπισμό των duplicates. Ο κώδικας θα πρέπει να δέχεται ένα όριο ομοιότητας θ. Τέλος θα πρέπει να επιστρέφουν όλα τα ζευγάρια κειμένων με ομοιότητα μεγαλύτερη από 0.7. Τα αποτελέσματα θα αποθηκευτούν στο αρχείο ‘duplicatePairs.csv’ και θα έχουν την παρακάτω μορφή:

Document_ID1	Document_ID1	Similarity

Υλοποίηση Κατηγοριοποίησης (Classification)

Σε αυτό το ερώτημα θα πρέπει να δοκιμάσετε τις παρακάτω 2 μεθόδους Classification:

- Support Vector Machines (SVM).
- Random Forests.

Επίσης θα ελέγξετε την απόδοση των παραπάνω αλγορίθμων χρησιμοποιώντας τα παρακάτω features:

- Bag of Words (BoW).
- SVD διατηρώντας το 90% του variance (SVD).
- Average word vector για κάθε κείμενο (W2V).

Επίσης θα πρέπει να αξιολογήσετε και να καταγράψετε την απόδοση κάθε μεθόδου χρησιμοποιώντας 10-fold Cross Validation χρησιμοποιώντας τις παρακάτω μετρικές:

- Accuracy
- Precision
- Recall

Beat the Benchmark

Τέλος θα πρέπει να πειραματιστείτε με όποια μέθοδο Classification θέλετε, κάνοντας οποιαδήποτε προ-επεξεργασία στα δεδομένα επιθυμείτε με στόχο να ξεπεράσετε όσο περισσότερο μπορείτε την απόδοση σας στο προηγούμενο ερώτημα.
Θα πρέπει αναλυτικά να τεκμηριώσετε τα βήματα που ακολουθήσατε

Αρχεία Εξόδου

Ο κώδικας θα πρέπει για τα ερωτήματα που αφορούν το Classification θα πρέπει να δημιουργεί τα παρακάτω αρχεία

- EvaluationMetric_10fold.csv
- testSet_categories.csv
- roc_10fold.png

Το format των αρχείων EvaluationMetric_10fold.csv φαίνεται παρακάτω:

Statistic Measure	SVM (BoW)	Random Forest (BoW)	SVM (SVD)	Random Forest (SVD)	SVM (W2V)	Random Forest (W2V)	My Method
Accuracy							
Precision							
Recall							
F-Measure							
AUC							

Το format του αρχείου testSet_categories.csv, το οποίο θα περιέχει τις κατηγορίες των άρθρων που δίνονται στο Test set φαίνεται παρακάτω:

Test_Document_ID	Predicted_Category
1	World News
2	Technology
...	

Για το αρχείο “testSet_categories.csv” θα πρέπει να χρησιμοποιηθεί αυστηρά η παραπάνω μορφοποίηση διαχωρίζοντας τα δυο πεδία με τον χαρακτήρα TAB (‘\t’) και επίσης θα πρέπει στην πρώτη γραμμή να υπάρχουν οι δυο επικεφαλίδες (Test_Document_ID και Predicted_Category) και ακολούθως οι προβλέψεις του μοντέλου σας στις επόμενες γραμμές διευκρινίζοντας το ID του document από το test set και το αντίστοιχο category.

Σχετικά με το παραδοτέο

Ο φάκελος που θα παραδώσετε θα έχει το όνομα:
Ass1_όνοματεπώνυμο1_ΑΜ1_ονοματεπώνυμο2_ΑΜ2.

Ο φάκελος θα περιέχει:

1. ένα κείμενο με τον σχολιασμό στα πειράματα που κάνατε και στις μεθόδους που δοκιμάσατε σε μορφή PDF. Η αναφορά σας θα πρέπει να περιέχει και τους πίνακες με τα αποτελέσματα των αρχείων εξόδου και δε θα πρέπει να ξεπερνάει τις 30 σελίδες.
2. τα ζητούμενα αρχεία εξόδου.
3. τα αρχεία κώδικα που γράψατε.

Το εκτενές κείμενο που θα παραδώσετε, θα περιέχει την περιγραφή των δοκιμών σας και οτιδήποτε σκεφτείτε για να δείξετε τι δοκιμές κάνατε, για ποιο λόγο έχουν τα συγκεκριμένα αποτελέσματα οι μέθοδοι που επιλέξατε, πως λειτουργούν αυτές οι μέθοδοι και σχολιασμό των αποτελεσμάτων σας. Όλες οι εργασίες θα αξιολογηθούν στη βάση της σωστής τεκμηρίωσης και στο βαθμό που υλοποιούν τα ζητούμενα της εργασία.

Χρήσιμα Links:

- https://radimrehurek.com/gensim/models/word2vec.html
- https://code.google.com/archive/p/word2vec/
- https://radimrehurek.com/gensim/models/ldamodel.html