Τεχνικές Εξόρυξης Δεδομένων
Εαρινό Εξάμηνο 2015-2016

2η Άσκηση, Ημερομηνία παράδοσης: 5/6/2016
Ομαδική Εργασία (2 Ατόμων)

Σκοπός της εργασίας

Σκοπός της εργασίας είναι η εξοικείωση σας με την τεχνική Topic-Modeling LDA (Latent Dirichlet Allocation) και η χρήση της για βελτίωση του classification. Η υλοποίηση θα γίνει στην γλώσσα προγραμματισμού Python με την χρήση του εργαλείου SciKit-Learn.

Περιγραφή (Reminder)

Τα Datasets είναι αρχεία CSV των οποίων 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 κατηγόρια 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 κατηγόριες των άρθρων είναι 5 και είναι οι εξής:

- Politics
- Film
- Football
- Business
- Technology

Latent Dirichlet Allocation (LDA)

Για την υλοποίηση του clustering των δεδομενων θα χρησιμοποιησετε την βιβλιοθηκη GenSim. Η μέθοδος LDA έχει μια παραμετρο K η οποία καθορίζει τον αριθμό των latent topics. Το αποτέλεσμα του LDA είναι μια κατανομή σε latent topic για κάθε document (***LDA Features***).

Ζητούμενα

1. Αρχικά θα πρέπει να δοκιμάσετε αν το LDA μπορεί να χρησιμοποιηθεί για το σκοπό του classification. Δηλαδή θα πρέπει να εκπαιδεύσετε τον καλύτερο classifier που βρήκατε στην προηγουμένη εργασία καθώς και τους υπόλοιπους classifiers χρησιμοποιώντας τα

LDA features. *Θα πρέπει να πειραματιστείτε με διαφορετικό αριθμό από latent topics.* Πιο συγκεριμένα, θα δοκιμάσετε τις παρακάτω τιμές K=10, K=100 και K = 100.

2. Θα πρέπει να χρησιμοποιήσετε τα LDA Features μαζί με τα features τα οποία χρησιμοποιήσατε στην προηγουμένη εργασία για να δείτε αν μπορείτε να βελτιώσετε την απόδοση του classification. *Θα πρέπει να πειραματιστείτε με διαφορετικό αριθμό από latent topics.*
3. Τέλος θα πρέπει να χρησιμοποιήσετε το καλύτερο μοντέλο από αυτά των προηγούμενων δοκιμών σας και να υπολογίσετε τα labels που επιστρέφει για το test set.

Αρχεία Εξόδου

Ο κώδικας θα πρέπει να δημιουργεί τα παρακάτω αρχεία:

- EvaluationMetric_10fold_lda_only.csv
- EvaluationMetric_10fold_ex1_features.csv
- testSet_categories.csv

Το format των αρχείων EvaluationMetric_10fold φαίνεται παρακάτω:

Statistic Measure	Naive Bayes	KNN	SVM	Random Forest	My Method
Accuracy K = 10					
Accuracy K = 100					
Accuracy K =1000					

Χρήσιμα Links

1. https://radimrehurek.com/gensim/models/ldamodel.html
2. https://radimrehurek.com/gensim/wiki.html (LDA Tutorial)
3. http://videolectures.net/mlss09uk_blei_tm/?q=david%20blei