Τμήμα Μηχανικών Η/Υ και Πληροφορικής Πανεπιστημίου Ιωαννίνων

ΜΥΕ047: Αλγόριθμοι για Δεδομένα Ευρείας Κλίμακας

Ακαδημαϊκό Έτος 2021-22

1ο Σετ Ασκήσεων: Ανίχνευση Κοντινότερων Γειτόνων σε Συλλογή Εγγράφων

Ευάγγελος Τζώρτζης ΑΜ: 3088

Αναφορά

Στην άσκηση υλοποιήθηκαν διαφορετικοί τρόποι υπολογισμού της ομοιότητας εγγράφων. Συγκεκριμένα υλοποιήθηκαν οι μέθοδοι *brute force* και *locality sensitive hashing (LSH)* χρησιμοποιώντας ως μετρική είτε την *ομοιότητα Jaccard*, είτε την *ομοιότητα υπογραφών* μεταξύ των εγγράφων.

Τα αρχεία που δημιουργήθηκαν σε έκδοση *python 3.9.7* είναι τα:

functions.py

interactionMain.py

Το πρώτο περιέχει τις συναρτήσεις που ζητήθηκαν και το δεύτερο τη διεπαφή με τον χρήστη και καλεί συναρτήσεις από το πρώτο.

Η εκτέλεση γίνεται με την εντολή:

Windows: *python interactionMain.py*

Linux: *python3 interactionMain.py*

Στη συνέχεια ακολουθεί η επιλογή των παραμέτρων από ένα μενού:

```
1. Choose a file.
2. Choose the number of documents to consider.
3. Choose the number of neighbors to locate.
4. Choose the number of permutations of the signature matrix.
5. Choose the similarity metric.
6. Choose the method to use for the average similarity calculation.
7. Calculate/Load the signature matrix.
8. Run the calculation of the average similarity.
9. Compare two documents with each other.
10. Quit.
```

Στον χρήστη εμφανίζονται οι τρέχουσες επιλεγμένες παράμετροι, οι οποίες αρχικοποιούνται σε μη έγκυρες τιμές και ο υπολογισμός της μέσης ομοιότητας δεν επιτρέπεται αν δεν έχουν ορισθεί όλες σε έγκυρες τιμές, πχ:

Selections:

File selected: DATA_1-docword.enron.txt

Number of documents: -1

Neighbors:-1

Permutations: -1

Similarity Metric: -1

Calculation Method: Undefined

Sig Matrix created/loaded: False

Σημείωση: Για τη μέθοδο LSH η μεταβλητή rows per band ζητείται εφόσον επιλεγεί το 8.

Ο υπολογισμός της μέσης τιμής ομοιότητας γίνεται με την επιλογή:

```
8. Run the calculation of the average similarity.
```

Για τον υπολογισμό ομοιότητας μόνο μεταξύ δύο εγγράφων υπάρχει η επιλογή:

```
9. Compare two documents with each other.
```

Όπου εάν έχει επιλεχθεί η Brute Force μέθοδος υπολογίζεται η ομοιότητα μεταξύ των εγγράφων τόσο με τη Jaccard, όσο και με την Signature ομοιότητα (*προϋποθέτει την ύπαρξη του πίνακα υπογραφών, ακόμα και αν έχει επιλεχθεί ως μέτρο ομοιότητας η Jaccard*).Εάν έχει επιλεχθεί η LSH, τότε υπολογίζεται το επιλεγμένο LSH μέτρο (Jaccard ή Signature), και με Brute Force οι ομοιότητες Jaccard και Signature.

Πέρα από την υλοποίηση των μεθόδων που ζητήθηκαν στην εκφώνηση αυτολεξεί, δημιουργήθηκαν οι μέθοδοι:

Για το PROGRAMMING TASK PT5, δημιουργήθηκαν μέθοδοι που υπολογίζουν την μέση Jaccard/Signature ομοιότητα για *ένα έγγραφο με τους κοντινότερους γείτονές του* με την brute force και η συνάρτηση που υπολογίζει τη μέση ομοιότητα:

```
bruteForceNearestNeighborsWithJacSim
bruteForceNearestNeighborsWithSigSim
avgSim
```

Για το PROGRAMMING TASK PT6, πέρα από την lsh, φτιάχτηκαν οι:

```
calcJaccardSimListForLSH
calcSignatureSimListForLSH
avgSimLSH
```

για τον υπολογισμό των ομοιοτήτων για κάθε έγγραφο και στη συνέχεια την τελικής μέσης ομοιότητας.

Επιπλέον, δημιουργήθηκε η μέθοδος `checkNeighbors` για τον έλεγχο ότι όλα έγγραφα έχουν τουλάχιστον τον επιλεγμένο αριθμό γειτόνων και οι μέθοδοι

`combineCandidateFilesWithSimilarities` για τον συνδυασμό των γειτόνων με τις ομοιότητες τους σε λεξικό από λεξικά (πχ {1:{2:0.343, 6:0.045,…}…})

και `sortCandidatesSimilarities` για τη ταξινόμηση των γειτόνων ανάλογα με την ομοιότητα τους.

Ένα στιγμιότυπο εκτέλεσης:

```
================================================================================

Selections:
 File selected: DATA_1-docword.enron.txt
 Number of documents: 3000
 Neighbors:1
 Permutations: 512
 Similarity Metric: Signature
 Calculation Method: Brute Force
 Sig Matrix created/loaded: True

============================

    1. Choose a file.
    2. Choose the number of documents to consider.
    3. Choose the number of neighbors to locate.
    4. Choose the number of permutations of the signature matrix.
    5. Choose the similarity metric.
    6. Choose the method to use for the average similarity calculation.
    7. Calculate/Load the signature matrix.
    8. Run the calculation of the average similarity.
    9. Compare two documents with each other.
    10. Quit.

Please select an option from the menu:8


Calculating Average Signature Similarity...
Document: 0 out of: 3000 in time: 0.0
Document: 500 out of: 3000 in time: 153.7400996685028
Document: 1000 out of: 3000 in time: 307.0912022590637
Document: 1500 out of: 3000 in time: 460.7917585372925
Document: 2000 out of: 3000 in time: 606.7180118560791
Document: 2500 out of: 3000 in time: 753.9450128078461
Time of average Signature similarity for 3000 documents: 903.052
Average Signature similarity: 0.48061393229166666

================================================================================
```

Σημείωση: έχουν δημιουργηθεί αρχεία στα οποία έχουν αποθηκευτεί οι πίνακες υπογραφών και LSH για τα πειράματα. Αυτά με τις κατάλληλες παραμέτρους μπορούν να φορτωθούν αντί να υπολογιστούν.

Σχολιασμός:

Κατά τα πειράματα επιλέχθηκαν κατάλληλοι αριθμοί εγγράφων, γειτόνων, μεταθέσεων υπογραφών, ώστε να περιοριστούν οι χρόνοι εκτέλεσης, ενώ ταυτόχρονα να υπάρχουν ικανοποιητικά αποτελέσματα.

Για το dataset enron επιλέχθηκαν για τα πειράματα 3000 αρχεία , 1 ή 3 γείτονες, 512 μεταθέσεις και 1 γραμμή ανά μπάντα στο LSH.

Για το dataset nips επιλέχθηκαν για τα πειράματα 1500 αρχεία , 1 ή 5 γείτονες, 512 μεταθέσεις και 2 γραμμές ανά μπάντα στο LSH.

Οι γραμμές ανά μπάντα δεν είναι ιδανικές και δίνουν πολύ χαμηλό κατώφλι ομοιότητας, αλλά είναι απαραίτητες για την εύρεση τουλάχιστον του επιλεγμένου πλήθους γειτόνων σε κάθε έγγραφο. Όμως ακόμα και αν έγγραφο δεν έχει τους απαραίτητους γείτονες, ο αλγόριθμος θα θεωρήσει τους υπολειπόμενους με απόσταση 0.

Τα αποτελέσματα με τις διαφορετικές ομοιότητες και διαφορετικούς τρόπους υπολογισμού τους ήταν κοντά μεταξύ τους όπως αναμένεται με τις διαφορές να είναι κυρίως στους χρόνους υπολογισμού.

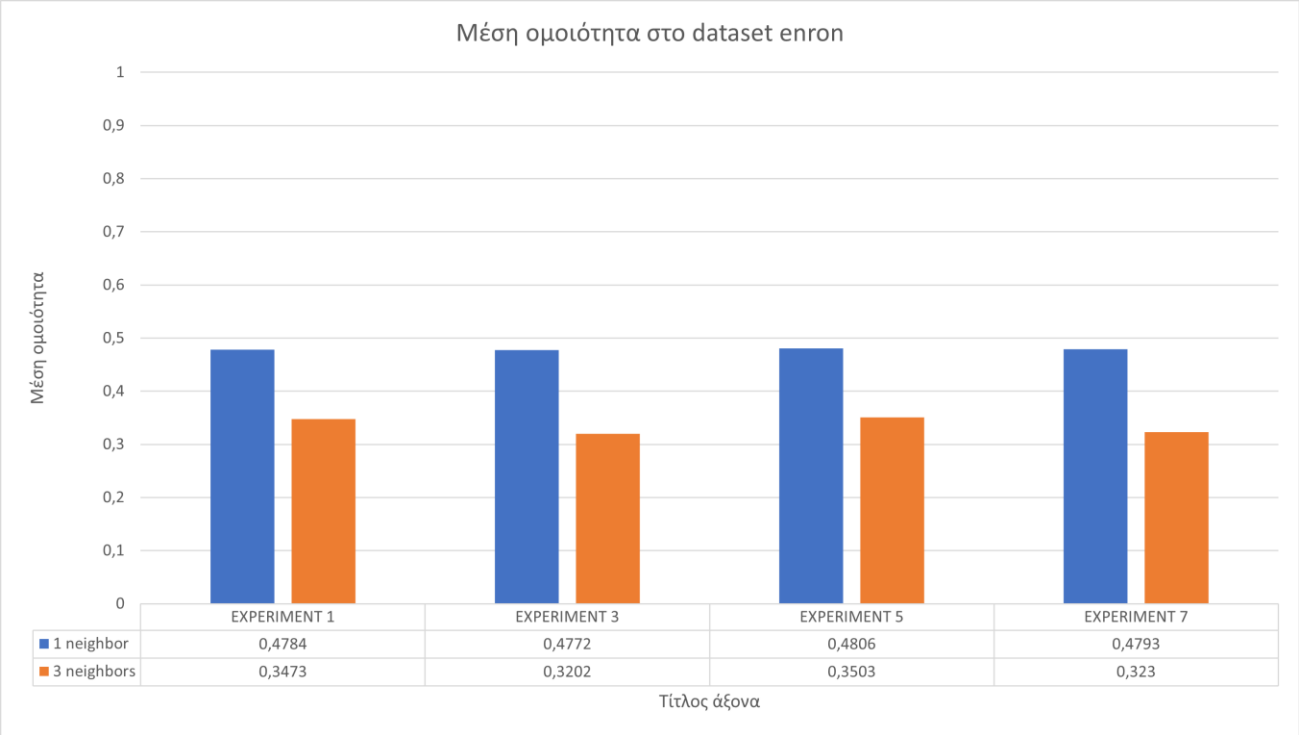
Στα πειράματα παρατηρήθηκε ότι για το dataset enron τα αποτελέσματα ποικίλουν ανάλογα με το πλήθος των γειτόνων εξίσου σε όλες τις περιπτώσεις με πτωτική μέση ομοιότητα όσο αυξάνονται οι γείτονες, κάτι που μας λέει ότι τα έγγραφα επί το πλείστων έχουν λίγους γείτονες (1-2) σχετικά μεγάλη ομοιότητα και οι υπόλοιποι έχουν μικρή και ρίχνουν επομένως τον μέσο όρο.

Στο dataset nips η μέση ομοιότητα είχε διαφορά 2% με 5 ή 1 γείτονες, δηλαδή οι πρώτοι 5 γείτονες έχουν παρόμοιες ομοιότητες.

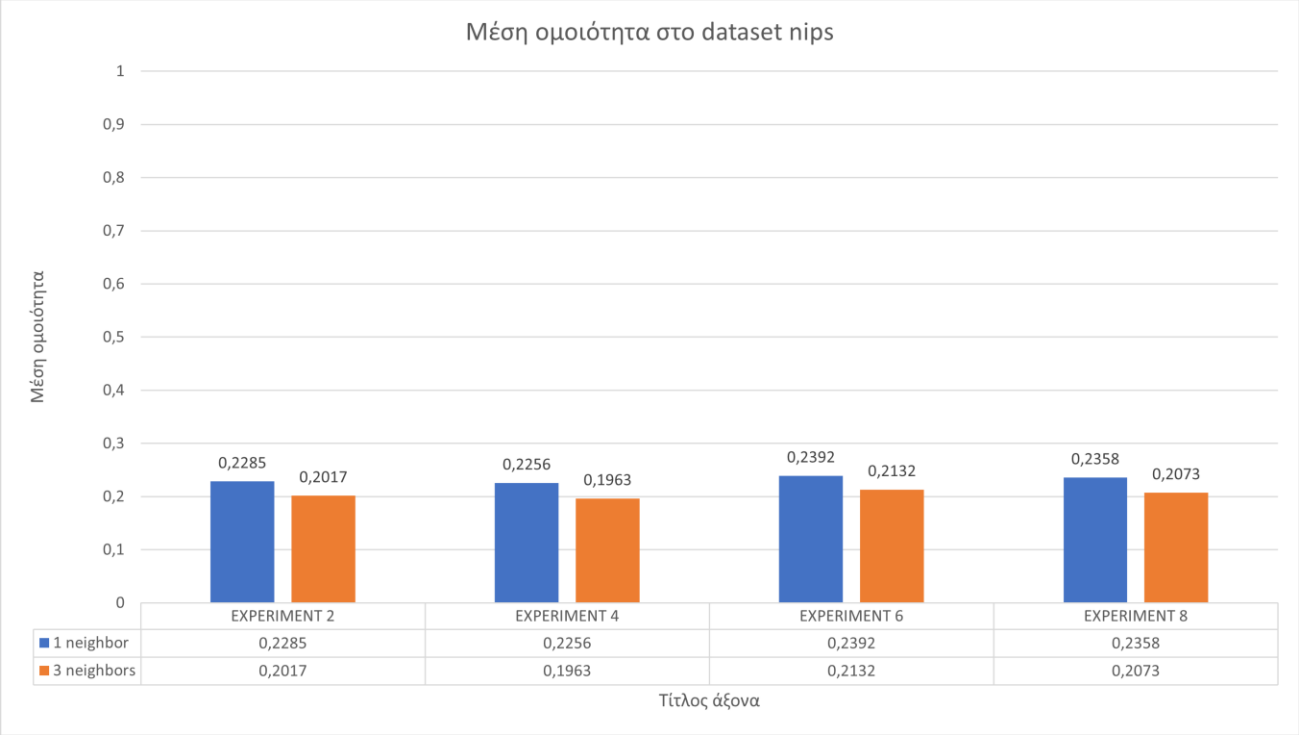
Αν εξαιρέσουμε τον υπολογισμό του πίνακα υπογραφών και του πίνακα LSH , η ταχύτητα μέτρησης της μέσης ομοιότητας βελτιώθηκε αρκετά όταν χρησιμοποιήθηκε LSH. Ωστόσο, στο enron λόγω των περισσότερων αρχείων η signature μέθοδος είναι πιο αργή γιατί δεν μπορεί να γίνει εύρεση των κοινών υπογραφών δύο εγγράφων με ordered lists όπως στη μέθοδο Jaccard, καθώς και για να βρεθούν οι απαιτούμενοι γείτονες χρειάζεται μεγάλο μητρώο υπογραφών. Στο nips με μικρότερο αριθμό εγγράφων ο χρόνος της εύρεσης της μέσης ομοιότητας υπογραφών είναι μικρότερος του αντίστοιχου για Jaccard.

Οι παραπάνω παρατηρήσεις βασίζονται στα παρακάτω διαγράμματα:

Διαγράμματα μέσης ομοιότητας:

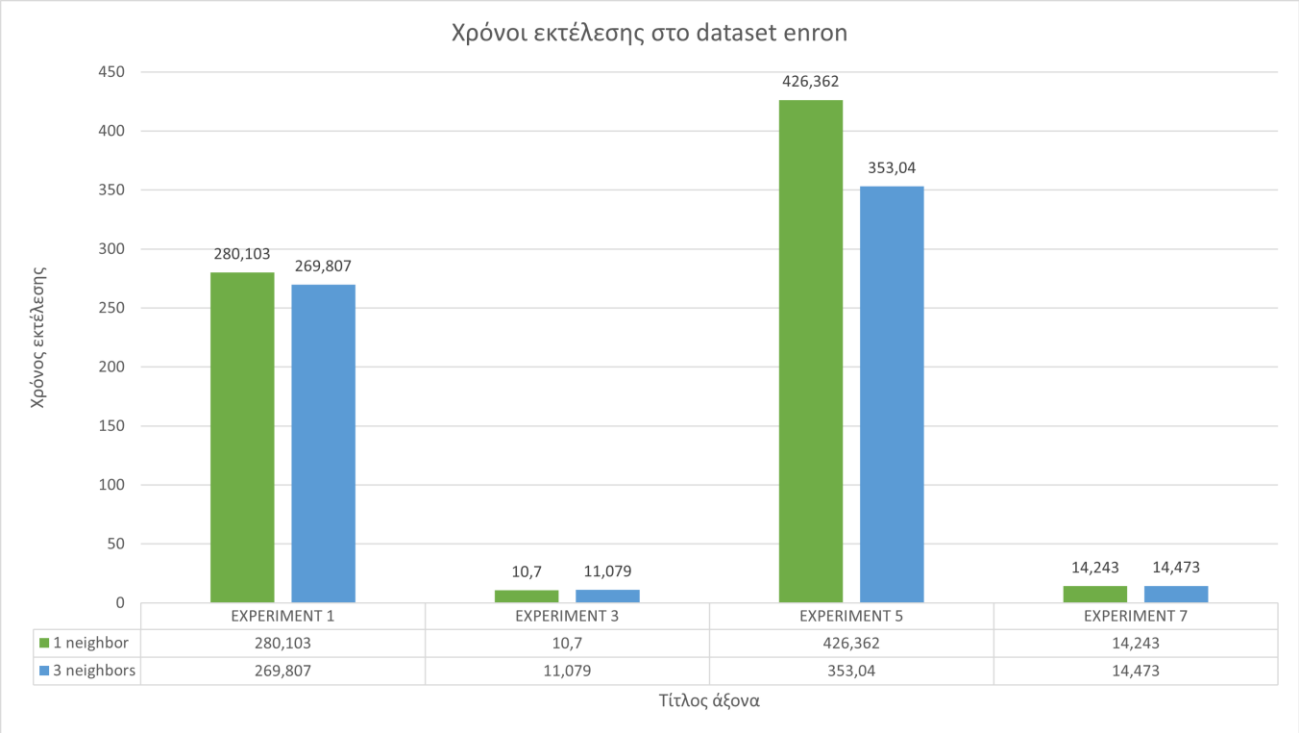


Τα αποτελέσματα είναι κοντά μεταξύ τους, τόσο με 1, όσο και με 3 γείτονες. Τα πειράματα 1,3 όπου χρησιμοποιείται η brute force είναι λίγο πιο κοντά μεταξύ τους, όπως και τα 3,7 με τη LSH.

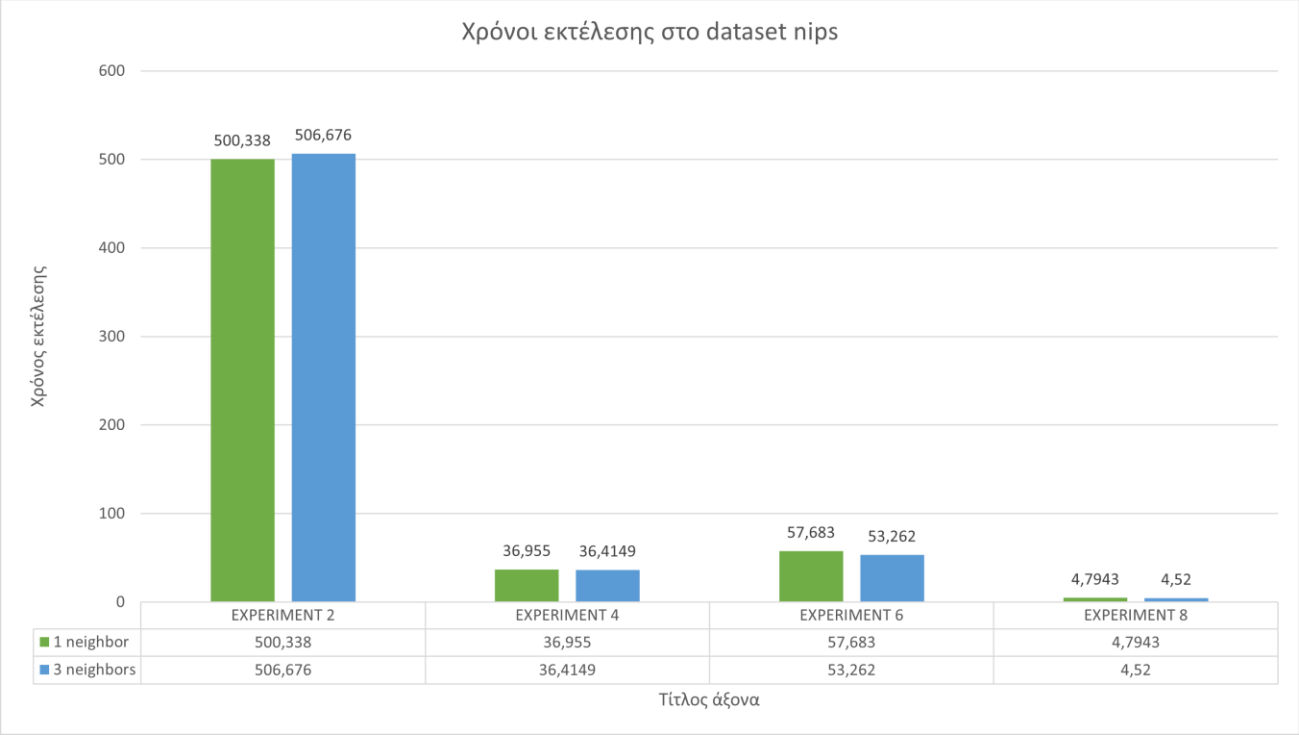


Και εδώ ανάλογα με τους γείτονες τα αποτελέσματα είναι πολύ κοντά, αλλά η διαφορά με 1 και 5 γείτονες είναι πολύ μικρή.

Διαγράμματα χρόνου εκτέλεσης:



Φαίνεται ότι ο χρόνος με τη μέθοδο LSH μειώνεται αρκετά. Στα πειράματα 1,5 και στα 3,7 αντίστοιχα φαίνεται για τη signature ομοιότητα ο χρόνος είναι μεγαλύτερος, ιδιαίτερα στη σύγκριση των 1 και 3.



Εδώ λόγω των λιγότερων εγγράφων, οι χρόνοι Jaccard-Signature έχουν σημαντική διαφορά, τόσο ανάμεσα στα πειράματα 2 και 6 όπου γίνεται υπολογισμός με Brute Force, όσο και στα 4 και 8. Το πείραμα 8 με LSH και ομοιότητα υπογραφών είναι το ταχύτερο.