Ανάλυση Βιβλιοθηκών Java με τεχνικές Ανάκτησης Πληροφοριών

Στο έγγραφο αναλύεται αρχικά ο τρόπος κατασκευής του συνόλου δεδομένων, ενώ στη συνέχεια παρουσιάζεται η μορφή του. Οι τεχνολογίες που αναφέρονται στην κατασκευή του συνόλου δεδομένων είναι χρήσιμες για την κατανόησή του. Παρόλο που η γνώση τους δεν αποτελεί προϋπόθεση για την εκπόνηση της εργασίας, μπορούν να βοηθήσουν στην κατανόηση του dataset και να προσφέρουν ιδέες για την προ-επεξεργασία των δεδομένων.

Τρόπος Κατασκευής του Συνόλου Δεδομένων

Το σύνολο δεδομένων αποτελείται από 80 γνωστές βιβλιοθήκες της γλώσσας προγραμματισμού Java. Οι βιβλιοθήκες αυτές ανήκουν σε 8 κατηγορίες:

- android
- command-line-parsers
- csv-libraries
- http-clients
- json-libraries
- swing-libraries
- testing-frameworks
- xml-processing

Η κατηγοριοποίηση αυτή έχει γίνει από τους προγραμματιστές των βιβλιοθηκών και είναι διαθέσιμη στην αποθήκη βιβλιοθηκών του Maven[1].

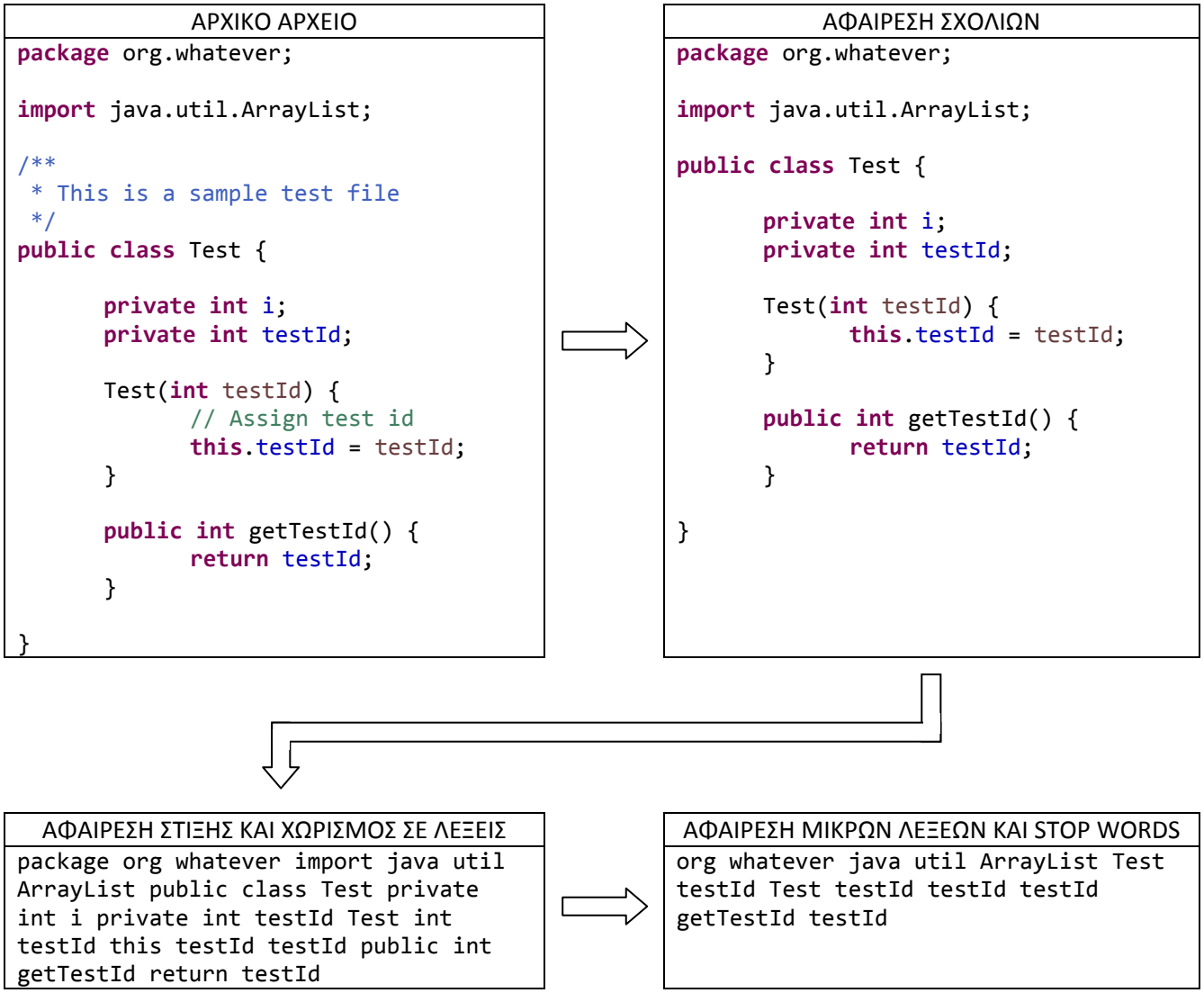
Για την κατασκευή του dataset αρχικά κατεβάσαμε τον κώδικα των βιβλιοθηκών από το GitHub[2]. Στη συνέχεια για κάθε βιβλιοθήκη εντοπίσαμε όλα τα αρχεία .java. Από κάθε αρχείο εξάγαμε τους όρους-λέξεις του, με τα εξής βήματα:

1) Αφαίρεση όλων των σχολίων (και μίας και πολλών γραμμών)
2) Αφαίρεση όλων των σημείων στίξης (τελείες, κόμματα, παρενθέσεις, αγκύλες κτλ.) και χωρισμός σε όρους-λέξεις (tokens)
3) Αφαίρεση όλων των λέξεων που είναι μικρότερες από 3 χαρακτήρες και αφαίρεση stop words της Java (π.χ. private, int, import, κτλ.)

Ένα παράδειγμα εξαγωγής tokens από ένα αρχείο .java φαίνεται στην Εικόνα 1.

Στη συνέχεια, για κάθε βιβλιοθήκη συνενώνουμε όλες τις λίστες από tokens που προκύπτουν για τα αρχεία .java της. Έτσι, θεωρούμε ότι κάθε βιβλιοθήκη είναι ένα μεγάλο document.

[1] Οι κατηγορίες προέκυψαν από τη σελίδα http://mvnrepository.com/. Περισσότερα για το Maven build automation tool στη σελίδα: https://en.wikipedia.org/wiki/Apache_Maven
[2] https://github.com/



Εικόνα 1 *Παράδειγμα εξαγωγής όρων-λέξεων (tokens) από ένα αρχείο java*

Στη συνέχεια, βρίσκουμε τη συχνότητα όλων των όρων σε κάθε document. Π.χ. αν θεωρήσουμε ότι μια βιβλιοθήκη έχει μόνο το αρχείο της Εικόνας 1, οι όροι θα έχουν τις απόλυτες συχνότητες που φαίνονται στον Πίνακα 1.

Πίνακας 1 *Απόλυτες Συχνότητες Όρων σε ένα Document*

Όρος	**Απόλυτη Συχνότητα**
org	1
whatever	1
java	1
util	1
ArrayList	1
Test	2
testId	5
getTestId	1

Μορφή του Συνόλου Δεδομένων

Το σύνολο δεδομένων αποτελείται από 80 γνωστές βιβλιοθήκες της γλώσσας προγραμματισμού Java. Οι βιβλιοθήκες αυτές ανήκουν σε 8 κατηγορίες. Στο πλαίσιο της παρούσας εργασίας, ζητείται να κάνετε ομαδοποίηση των βιβλιοθηκών[3].

Από τις βιβλιοθήκες έχουν εξαχθεί οι 10000 πιο συχνές λέξεις όροι των αρχείων .java καθώς και οι απόλυτες συχνότητες αυτών των όρων. Ένα παράδειγμα συνόλου δεδομένων φαίνεται στον Πίνακα 3.

Πίνακας 2 *Παράδειγμα Συνόλου Δεδομένων*

project	category	args	CSV	Elem	Frame	JSONObj	Lock	StringWriter	android	java
mydroid	android	0	0	0	1	0	0	0	3	1
drlib	android	0	0	0	0	0	1	0	1	0
csvlib	csv-libs	1	1	0	0	0	0	1	0	1
myjson	json-libs	2	0	0	0	0	0	1	0	1
jsonlib	json-libs	0	0	1	0	1	0	0	0	1
xmllib	xml-libs	0	0	1	0	0	0	0	0	1

Η στήλη project περιέχει τις βιβλιοθήκες, ενώ για κάθε βιβλιοθήκη δίνεται επίσης η κατηγορία που ανήκει. Οι όροι που εμφανίζονται σε όλες τις βιβλιοθήκες είναι επίσης στήλες του πίνακα (π.χ. args, android, κτλ.). Για κάθε όρο δίνεται ένας αριθμός που αντιστοιχεί στη συχνότητα εμφάνισης του στη βιβλιοθήκη. Έτσι, για παράδειγμα, παρατηρούμε ότι ο όρος android εμφανίζεται στις βιβλιοθήκες mydroid και drlib, που είναι λογικό καθώς αυτές είναι βιβλιοθήκες που αφορούν το λειτουργικό σύστημα android. Καθώς η τιμή της συχνότητας του όρου είναι μεγαλύτερη για τη βιβλιοθήκη mydroid, είναι σαφές ότι ο όρος είναι αρκετά σημαντικός για αυτή τη βιβλιοθήκη.

Αυτό που ζητείται από εσάς είναι χρησιμοποιώντας τους όρους, **και όχι τη στήλη category**, να ομαδοποιήσετε τις βιβλιοθήκες σε κατηγορίες που θα βοηθήσουν στην κατανόηση των δεδομένων, στην εύρεση παρόμοιων βιβλιοθηκών, κ.α. Η στήλη category δίνεται για τον υπολογισμό μετρικών ποιότητας της ομαδοποίησης.

[3] Σχετική πηγή για την ομαδοποίηση κειμένων https://en.wikipedia.org/wiki/Document_clustering