ΟΜΑΔΑ:
ΒΛΑΧΟΣ ΜΙΧΑΗΛ - 1640  ΣΑΛΑΒΑΣΙΛΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ - 1669

## Επιλογή Δεδομένων

Επιλέξαμε να χρησιμοποιήσουμε ως υποσύνολο των δεδομένων του Yelp το σύνολο των επιχειρήσεων που έχουν αριθμό κριτικών περισσότερες από 2500. Αυτή η επιλογή έγινε για να μειώσει τον όγκο των διθέσιμων δεδομένων καθώς και να παραχθούν καλύτερα και ποιοτικότερα αποτελέσματα για τους χρήστες.

## Προεπεξεργάσια δεδομένων

Αρχικά συλλέγουμε τα έγγραφα που θέλουμε να συμπεριλάβουμε στο ευρετήριο. Τα έγγραφα αυτά τα λαμβάνουμε απο το Yelp dataset που κατεβάσαμε απο το αντίστοιχο site το οποίο είναι το:  https://www.yelp.com/dataset_challenge.

Διαιρούμε το κείμενο σε γλωσσικά σύμβολα (token).  Ένα token είναι μια ακολουθεία απο χαρακτήρες , καθε τέτοιο token είναι υποψήφιο για να εισαχθεί στο ευρετήριο μετά από περαιτέρω επεξεργασία. Επιλέγουμε κατάλληλα τα tokens ανάλογα με την γλώσσα. Κύρια γλώσσα μας θα είναι τα Αγγλικά. Παραδείγματος χάρη ένωση λέξεων με την χρήση του χαραστήρα “-” ή “ ' “ ή του κενού κλπ. Θα μπορούσε να γίνει χρήση μιας Stop List ώστε να αποκλειστούν οι πιο κοινές λέξεις με βάση τη συχνότητά τους οι οποίες έχουν μικρό σημασιολογικό περιεχόμενο, ωστόσο η τάση είναι να μην χρησημοποιούνται.
Μονάδα εγγράφου για το ευρετήριο μας θα είναι ένα αρχείο με τις επιχειρήσεις κάθε πόλης.

Για την κατασκευή του ευρετηρίου επεξεργαζόμαστε τα έγγραφα για να βρούμε τις λέξεις που αποθηκεύονται μαζί με το αναγνωριστικό του αρχείου σε ζεύγη (term, doc-id). Αφού έχουμε επεξεργαστεί όλα τα έγγραφα, το ανεστραμμένο ευρετήριο διατάσσεται με βάση τους επιλεγμένους όρους. Αλλά επειδή θα έπρεπε να διατάξουμε πολύ μεγάλο αριθμό από όρους θα χρησημοποιήσουμε termid αντί του term για καλύτερη απόδοση. Λόγω του μεγάλου όγκου εγγραφών η διάταξη στον δίσκο γίνετε πολύ αργή και απαιτούνται πολλές τυχαίες ανακτήσεις. Γιαυτό χρειαζόμαστε έναν αλγόριθμο εξωτερικής διάταξης. Θα κανουμε χρήση του αλγόριθμου κατασκευής κατά Block (BSBI). Η ιδέα του αλγορίθμου είναι η εξής: χωρίζουμε την συλλογή σε κομμάτια ίσου μεγέθους, στην συνέχεια ταξινομούμε τα ζεύγη termid-docid για καθε κομμάτι στην μνήμη, αποθηκεύντας τα ενδιάμεσα αποτελέσματα στο δίσκο και στο τέλος συγχωνεύουμαι τα ενδιάμεσα αποτελέσματα

## Διάταξη Αποτελεσμάτων

Για την διάταξη των αποτελεσμάτων ο χρήστης θα μπορεί να επιλέξει από διάφορους τρόπους διάταξης όπως την περιοχή της επιχείρησης, τον βαθμό των κριτικών, τον αριθμό των κριτικών και τα check-ins. Τα αποτελέσματα θα εμφανίζονται με βάση την ποιότητα του εγγράφου δηλαδή ανεξάρτητα(στατικά) του ερωτήματος ανάλογα με τον συγκεντρωτικό βαθμό του εγγράφου.

Εμφάνιση Αποτελεσμάτων

Ο αρχικός μας σχεδιασμός μας για την απεικόνιση των αποτελεσμάτων είναι με δυναμικές περιλήψεις δηλαδή μέσα στην περίληψη του αποτελέσματος περιέχονται αρκετοί από τους όρους της αναζήτησης.