Α

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Εαρινό Εξάμηνο 2017-2018

**Εξόρυξη και Ανάκτηση**

**Πληροφορίας στον**

**Παγκόσμιο Ιστό**

**** *Διδάσκουσα: Βακάλη Αθηνά*

*Υπευθ. Εργασίας: Δημητριάδης Ηλίας*

Πρακτικό Μέρος

**Μπόζας Αριστείδης - 740**

**Πέσιος Δημήτριος - 746**

**Ψάλτογλου Άρτεμις - 30**

04 Ιουνίου 2018

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

AΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

**Περιεχόμενα**

[1. Εισαγωγή 3](#_Toc515574831)

[2. Προεπεξεργασία δεδομένων 4](#_Toc515574832)

[3. Ανάλυση και εξερεύνηση του συνόλου των δεδομένων 5](#_Toc515574833)

[4. Εξαγωγή συναισθηματικής πληροφορίας 5](#_Toc515574834)

[5. Δημιουργία web εφαρμογής 5](#_Toc515574835)

1. **Εισαγωγή**

Η παρούσα εργασία ασχολείται με την εξόρυξη και την ανάκτηση πληροφορίας στον Παγκόσμιο Ιστό. Πιο συγκεκριμένα, εστιάζει στην εξαγωγή πληροφορίας από κοινωνικά δίκτυα που έχουν διαμορφωθεί στα πλαίσια μιας πληθοποριστικής πλατφόρμας (crowdsourcing), η οποία συγκεντρώνει πληροφορίες σχετικά με επιχειρήσεις διαφόρων τύπων και τις οποίες μπορούν να αξιολογήσουν οι χρήστες. Η εργασία αυτή χρησιμοποιεί την πλατφόρμα Yelp, και ειδικότερα, ένα σύνολο δεδομένων που περιέχει δεδομένα σχετικά με εστιατόρια που βρίσκονται στην ευρύτερη περιοχή του Las Vegas. Το σύνολο δεδομένων περιέχει:

* Restaurants: 5899 εστιατόρια
* Reviews: 906.000 αξιολογήσεις
* Users: 337.000 χρήστες

Σκοπός της εργασίας είναι η εξαγωγή του γενικού συναισθήματος των αξιολογήσεων των χρηστών μέσω τεχνικών που βασίζονται στην μηχανική μάθηση. Σε αυτήν την τεχνική αναφορά παρουσιάζονται τα βήματα που ακολουθήθηκαν για την εξαγωγή συναισθηματικής πληροφορίας καθώς επίσης και τα αποτελέσματα που προέκυψαν για κάθε διαφορετικό αλγόριθμο ταξινόμησης. Μια συνολική επισκόπηση της εργασίας παρουσιάζεται στο τέλος με τη μορφή μιας διαδικτυακής εφαρμογής (web).

Τα εργαλεία που χρησιμοποιήθηκαν για την εκπόνηση της εργασίας είναι η NoSQL βάση δεδομένων MongoDB σε συνδυασμό με το γραφικό περιβάλλον χρήστη Studio 3T, ενώ η υλοποίηση της εργασίας έγινε στην Python, μια υψηλού επιπέδου γλώσσα προγραμματισμού.

1. **Προεπεξεργασία δεδομένων**

Αυτή η ενότητα περιλαμβάνει μια σειρά από βήματα που αφορούν την προεπεξεργασία του κειμένου των αξιολογήσεων με σκοπό την προετοιμασία τους για περαιτέρω ανάλυση. Πιο συγκεκριμένα, τα βήματα που ακολουθήθηκαν είναι τα εξής:

* Αφαίρεση των σημείων στίξης και αποθήκευση του κειμένου
* Διαχωρισμός του κειμένου σε λεξιλογικές μονάδες (tokenization)
* Αναγωγή των λέξεων στη ρίζα τους (stemming)
* Αφαίρεση των κοινών λέξεων (stop words)
* Μετατροπή όλων των χαρακτήρων σε πεζούς χαρακτήρες
* Υπολογισμός διανύσματος κάθε όρου, με βάση το TF-IDF βάρος που έχει.

1. **Ανάλυση και εξερεύνηση του συνόλου των δεδομένων**

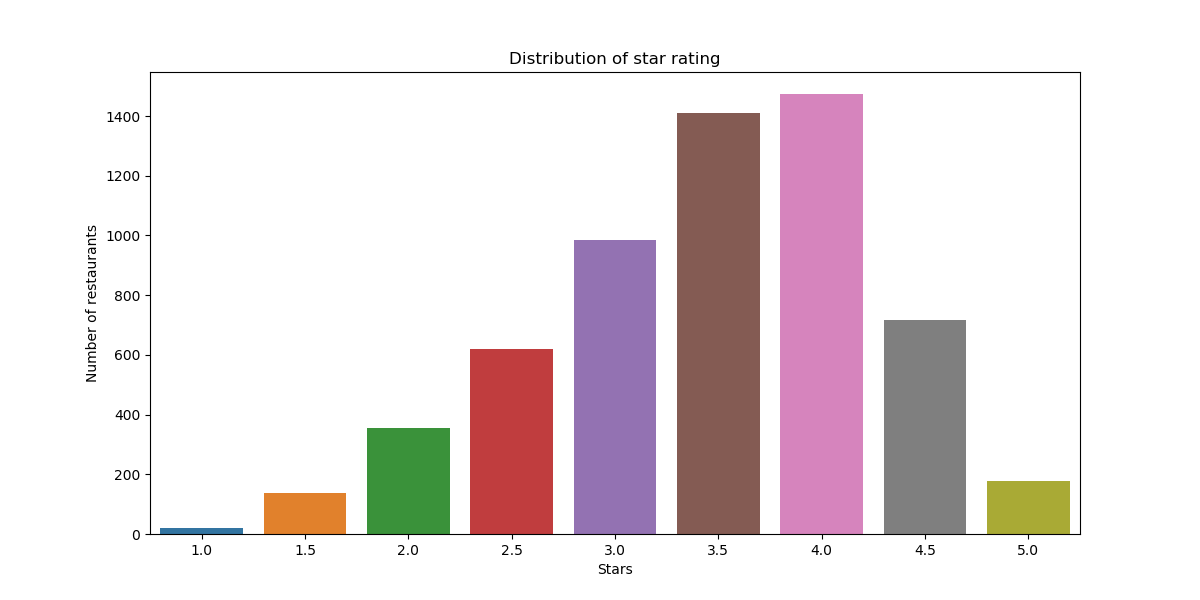
Αυτή η ενότητα διερευνά το σύνολο των δεδομένων που χρησιμοποιείται και παρουσιάζει μια σειρά από διαγράμματα που βοηθούν στην καλύτερη κατανόηση των δεδομένων αυτών. Για τη δημιουργία αυτών των διαγραμμάτων χρησιμοποιήθηκαν 3 βασικές βιβλιοθήκες της python:

* **Numpy**: προσφέρει υψηλού επιπέδου μαθηματικές λειτουργίες και υποστηρίζει την διαχείριση μεγάλων πολυδιάστατων πινάκων και γραφημάτων
* **Pandas**: επιτρέπει τη διαχείριση των δεδομένων σε ευέλικτες δομές (DataFrames) και υποστηρίζει μεγάλο εύρος μεθόδων και τεχνικών επεξεργασίας δεδομένων
* **Matplotlib**: υποστηρίζει την οπτικοποίηση και επεξεργασία γραφημάτων
* **Seaborn**: βασίζεται στην Matplotlib και επεκτείνει τις δυνατότητες οπτικοποίησης και επεξεργασίας γραφημάτων

Στη συνέχεια παρουσιάζονται τα γραφήματα που δημιουργήθηκαν με βάση τις παραπάνω βιβλιοθήκες για την κατανόηση των δεδομένων.

1. **Κατανομή αστεριών αξιολόγησης**

Αρχικά, διερευνούμε το μέσο αστεριών που έχει το κάθε εστιατόριο, ώστε να δούμε ποια είναι η κατανομή των αστεριών αξιολόγησης με βάση τα δεδομένα αυτά. Από το σύνολο δεδομένων που αφορά τα εστιατόρια (restaurants), βρίσκουμε τo πλήθος των εστιατορίων που έχουν μέσο όρο αστεριών 1.0, 1.5, 2.0 κ.ο.κ. Όπως φαίνεται στην παρακάτω εικόνα, παρατηρούμε ότι περίπου το 50% των εστιατορίων έχουν μέσο όρο 3.5 ή 4 αστέρια, και ο μέσος όρος αστεριών είναι 3.45.

****

count 5899.000000

mean 3.458044

std 0.804138

min 1.000000

25% 3.000000

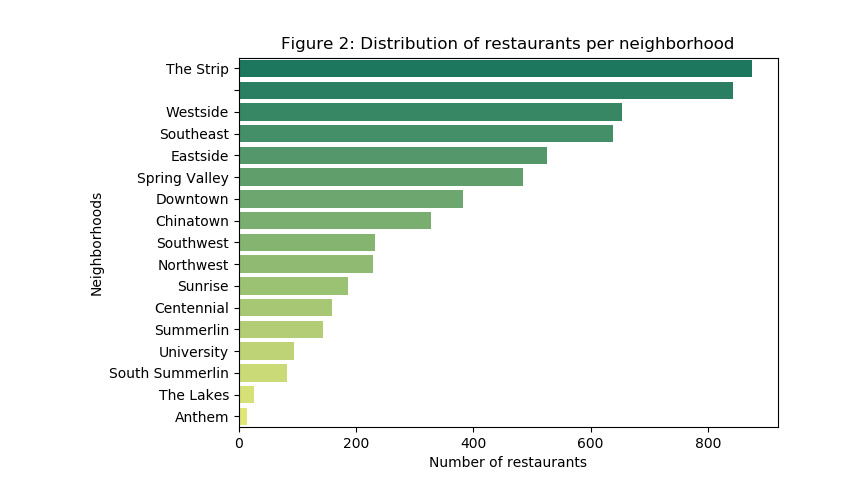
50% 3.500000

75% 4.000000

max 5.000000

1. **Κατανομή εστιατορίων ανά γειτονιά**

Στη συνέχεια, διερευνούμε την κατανομή των εστιατορίων ανά γειτονιά. Όπως φαίνεται παρακάτω, υπάρχουν 16 μοναδικές τοποθεσίες συν όσα εστιατόρια δεν έχουν τιμή για τη γειτονιά στην οποία ανήκουν. Η πιο συχνή περιοχή είναι η The Strip με 875 εστιατόρια.



count 5899

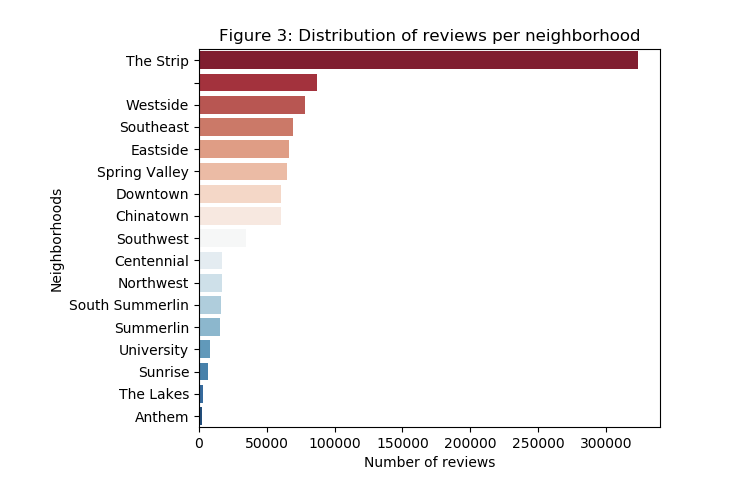
unique 17

top The Strip

freq 875

1. **Κατανομή αξιολογήσεων ανά γειτονιά**

Στη συνέχεια, διερευνούμε την κατανομή των αξιολογήσεων ανά γειτονιά. Όπως φαίνεται παρακάτω,

****

count 5899

unique 17

top The Strip

freq 875

1. **Εξαγωγή συναισθηματικής πληροφορίας**
2. **Δημιουργία web εφαρμογής**