Μετρικές επιτυχημένων ιστότοπων και εταιριών

*Βήματα που έχουν γίνει μέχρι τώρα:*

Για την διεξαγωγή της παρούσας διπλωματικής εργασίας θα πρέπει να μελετηθεί η συσχέτιση ανάμεσα σε μετρικές του περιεχομένου ενός ιστότοπου και της επιτυχίας του ιστότοπου ή της αντίστοιχης εταιρίας. Οι σελίδες που θα ερευνηθούν θα είναι αυτές των ιστότοπων που εμφανίζονται στην κορυφή των Alexa rankings καθώς και των εταιριών Fortune 1000.

Αρχικά θα πρέπει να εντοπίσουμε τους ιστότοπους τους οποίους θα μελετήσουμε. Για τις ανάγκες της παρούσας εργασίας θα χρησιμοποιήσουμε τους ιστότοπους των 500 μεγαλύτερων εταιρειών σύμφωνα με το fortune 500. Για να μπορέσουμε να κατεβάσουμε τις αρχικές σελίδες αυτών των ιστότοπων γράψαμε έναν κώδικα μέσω της γλώσσας προγραμματισμού python (***download\_all\_the\_sites\_500.py***) ο οποίος κατεβάζει την σελίδα της εταιρίας που ανήκει στις 500 μεγαλύτερες σύμφωνα με το fortune500. Για να μπορέσουν να κατεβουν αυτές τις σελίδες φτιάχθηκε ένα αρχείο csv που περιέχει τα ονόμα και τα url κάθε μίας από αυτές τις σελίδες (*fortune500.csv*). Οι σελίδες αποθηκεύτηκαν σε μορφή .html αλλά και σε μορφή .txt για ευκολία. Στο GitHub ανέβηκαν ενδείκτικα οι 50 πρώτες σελίδες που κατέβηκαν. (Σε συνέχεια θα ανέβουν και οι εναπομίναντες.)

Επόμενο βήμα ήταν να ελέγχθουν οι βασικές μετρικές για να μπορέσει να φανεί πόσο διαφέρουν μεταξύ τους οι σελίδες. Αρχικά γράφτηκε ένας νέος κώδικα ο οποίος δημιουργεί ένα csv αρχείο (*info\_images.1csv*) με τον αριθμό και το είδος των εικόνων που περιέχει η κάθε σελίδα (***images.py***).

Αξίζει να σημειωθεί ότι στην προσπάθεια να φτιάχτει ο παραπάνω κώδικας εντοπίστηκε ότι 20 από τις σελίδες του fortune500 δεν μπόρεσαν να ανοίξουν ούτε να αποθηκευτούν στο προηγούμενο βήμα. Για να ξέρουμε λοιπόν ποιες είναι αυτές οι σελίδες μέσα στον παραπάνω κώδικα παράγουμε ένα ακόμα csv αρχείο όπου καταγράφουμε τις 20 αυτές σελίδες για να τις ελένξουμε ξανά. (*not\_downloadables1.csv*)

Στη συνέχεια θα υπολογίσουμε και θα καταγράψουμε τον χρόνο φόρτισης της σελίδας μέσω ενός νέου κώδικα με την χρήση της python (***loading\_time.py***) και τα αποτελέσματα θα εξαχθούν σε ένα csv αρχείο(*loading\_times.csv*). Επειδή ωστόσω το σημείο διαχωρισμού δεκαδικών που μπήκε είναι η τελεία χρησιμοποιούμε ένα μικρό script στην r (infos\_cleansing.R) και το αντικαθιστούμε με το κόμμα για να μην υπάρχουν προβλήματα στην ανάγνωση του αργότερα και ονομάζουμε το νέο αρχείο (*loading\_time.csv*).

Το επόμενο βήμα είναι να καταγράψουμε πόσα διαφορετικά social media χρησιμοποιεί η κάθε σελίδα ώστε να έχουμε μια αίσθητη για την online παρουσία της. Το αρχείο που δημιουργήθηκε ονομάστηκε (*social\_media.csv*) ενώ ο κώδικας ονομάστηκε (***social\_media.py***).

Το επόμενο βήμα είναι ο έλεγχος ορθότητας της html. Ως κριτήρια για να πραγματοποιήσουμε τον έλεγχο αυτό θα πάρουμε το κατά πόσο κάθε <> που ανοίγει κλείνει κιόλας μέσα στον κώδικα </>. Αυτό θα επιτευχθεί μέσω ενός νέου script της python (***correct\_html.py***) το οποίο θα εξάγει ένα νέο csv με την ζητούμενη πληροφορία (*correct\_html.csv*).

Στη συνέχεια δημιουργήθηκε ένα νέο script (***ex\_in\_tot\_links.py***) που υπολογίζει τον αριθμό των εσωτερικών και εξωτερικών δεσμών αλλά και των συνολικών δεσμών που περιέχει η εκάστοτε σελίδα (*links.csv*). Ως εσωτερικούς δεσμούς θεωρούμε αυτούς που δείχνουν σε μια άλλη σελίδα μέσα στο ίδιο site ενώ ως εξωτερικούς τους δεσμούς που δείχνουν σε άλλες σελίδες.

Τέλος δημιουργήθηκε ένα νέο script (***words.py***) που υπολογίζει τον αριθμό των λέξεων και των μοναδικών λέξεων που περιέχει η εκάστοτε σελίδα. Αρχικά για να επιτευχθεί σωστά αυτό το στάδιο πρέπει να δημιουργηθεί ένα λεξικό με το οποίο θα συγκρύνονται οι λέξεις που θα εξαχθούν από τον html κώδικα για να σιγουρευτεί η ύπαρξη τους. Το λεξικό αυτό το βρήκαμε σε μορφή txt στο διαδίκτυο και περιέχει όλες τις αγγλικές λέξεις (dictionary.txt). Αφού λοιπόν εξάγουμε τις λέξεις μέσω του κώδικα και την συχνότητα εμφάνισης τους τις συγκρίνουμε με τις λέξεις του λεξικού και εάν υπάρχουν τις κρατάμε στο τελικό αρχείο (*words.csv*). Ωστόσω το αρχείο αυτό είναι δύσκολα συγκρίσιμο με τις υπόλοιπες πληροφορίες οπότε δημιουργήθηκε ένας δεύτερος κώδικας που κρατάει μόνο το σύνολο των λέξεων για κάθε site όπως και το σύνολο των μοναδικών λέξεων - δηλαδή των λέξεων που εμφανίζονται μόνο μία φορά στο κείμενο - (***unique\_words.py***) και τα αποθηκεύει σε ένα νέο αρχείο (*unique\_words.csv*).

Όλα τα csv αρχεία που δημιουργήθηκαν θα πρέπει να ενωθούν σε ένα ενιαίο αρχείο για να μπορέσει να μελετηθεί με την βοήθεια του RStudio. Για να επιτευχθεί αυτή η ένωση χρησιμοποιήθεικε η φόρμουλα του excel VLOOKUP. Ως αποτέλεσμα δημιουργήθηκαν 2 νέα αρχεία csv για ευκολία. Το ένα περιέχει τα 500 sites (*total.csv*) ενώ το 2ο περιλαμβάνει όλα τα site εκτός των 20 για τα οποία δεν έχουμε ακόμα διαθέσιμα κατεβασμένα sites (*total\_not\_na.csv*).

*Βήματα που έχουν γίνει μέχρι τώρα:*

* Στην συνέχεια για να μπορέσουμε να δούμε και την ανάλυση των εικόνων θα δημιουργηθεί ένα ακόμα κομμάτι κώδικα που θα βρίσκει τις διαστάσεις που χρησιμοποιεί η κάθε σελίδα για να δείξει τις εικόνες της και πόσες φορές χρησιμοποιεί την κάθε διάσταση και θα προστεθούν και στο τελικό csv που θα χρησιμοποιηθεί.
* Έπειτα θα προσπαθήσουμε να εκμεύσουμε από τις σελίδες του fortune 500 στοιχεία για την κατάσταση της κάθε εταιρεία όπως π.χ. market share, revenues, profit change, market value, assets,total ROI
* Στη συνέχεια θα δημιουργηθεί ένα ενιαίο αρχείο με όλα τα στοιχεία που βρέθηκαν και θα πραγματοποιηθεί ανάλυση με την χρήση της R στον συσχετισμό τον τιμών αυτών.