Κατασκευή θεματικού δικτύου από την wikipedia

Εισαγωγή

Η wikipedia είναι μια διαδικτυακή εγκυκλοπαίδεια στην οποία ο καθένας μπορεί να εισάγει ή να τροποποιήσει την πληροφορία που εμφανίζεται υπό κάποιους όρους, όπως ότι το περιεχόμενο πρέπει να έχει αναφορές σε κάποιες πηγές (άρθρα, βιβλία, εφημερίδες κτλ.). Αυτή την στιγμή έχει ~55 εκατομμύρια άρθρα σε 300 γλώσσες και θεωρείτε μία καλή πηγή σε ότι αφορά την γενική γνώση.

Ο σκοπός της εργασίας είναι να φτιάξετε ένα κώδικα σε python που δεδομένου κάποιου αρχικού θέματος δοσμένο με λέξεις κλειδιά (πχ graph theory, social graph, pine tree, Maradona, apple pie) θα κατασκευάζει ένα δίκτυο άρθρων της Wikipedia με συναφή θέματα βάσει των

1. **συνδέσμων στα άρθρα**
2. **σημασιολογικής συσχέτισης δυο άρθρων**

και σε αυτό το δίκτυο θα υπολογίσετε θεματικές περιοχές.

Θα πρέπει να αναλύσετε το παραγόμενο δίκτυο στο Gephi και με την βιβλιοθήκη networkx για να βγάλετε συμπεράσματα τα οποία θα μπορεί να παραχθούν αυτόματα.

Απαιτούμενα

Θα πρέπει να:

- τα άρθρα να προέρχονται μόνο από την wikipedia (english language)
- να κατασκευάζεται ένα αρχείο .gephi που θα περιέχει το θεματικό γράφημα όπως και όλες τις αναλύσεις που έγιναν για μία δική σας επιλογή αρχικού θέματος
- να παραδοθεί ο κώδικας, το αρχείο .gephi όπως επίσης ένα κείμενο που να περιγράφει την ανάλυση του θεματικού δικτύου.
- τα όποια συμπεράσματα για την θεματική περιοχή που ανήκει το θέμα που δόθηκε με τις λέξεις κλειδιά πρέπει να είναι δυνατόν να βγαίνουν αυτόματα

Ποιές έννοιες πρέπει να εξετάσετε

Οι παρακάτω έννοιες από την θεωρία πολύπλοκων δικτύων θα φανούν χρήσιμες όσον αφορά την ερμηνεία του δικτύου σε σχέση με τα θέματα που καλύπτουν

- Διάφορες μετρικές σημαντικότητας κορυφής
- Ομαδοποίηση κορυφών (clustering)
- Χαρακτηριστικά δικτύου όπως διάμετρος, ακτίνα
- Απόγονοι, πρόγονοι κάποιας κορυφής
- κατανομή βαθμών
- συνεκτικότητα

Τι θα χρειαστείτε

Τα παρακάτω εργαλεία θα τα χρειαστείτε για την εργασία

- python
- pandas
- Gephi
- networkx (python module)
- wikipedia (python module)

Τι θα πρέπει να προσέχετε

- το πλήθος των άρθρων που προκύπτουν από τους συνδέσμους που υπάρχουν σε ένα άρθρο μεγαλώνει εκθετικά γρήγορα σε σχέση με το βάθος αναζήτησης.
- δεν είναι όλα τα άρθρα που αναφέρονται σε ένα άρθρο σημαντικά.
- πολλά άρθρα έχουνε συνώνυμους τίτλους
- προσπαθήστε να βρείτε εφαρμογές των εννοιών από την θεωρία δικτύων στο θεματικό δίκτυο
- για να βρείτε το κατα πόσο δυο άρθρα i και j είναι σημασιολογικά παρόμοια, υπολογίστε μια τιμή s(i,j) στο [0,1] αναλόγως πόσο *σχετικά* είναι δύο άρθρα i και j μεταξύ τους, όπου με 1 συμβολίζουμε την μεγαλύτερη συσχέτιση. Για τον υπολογισμό της συσχέτισης s(i,j) μπορείτε να χρησιμοποιήσετε το παρακάτω μοντέλο γλώσσας που θα σας δώσει ενβυθίσεις προτάσεων https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2