Εργασία στο μάθημα Θεωρία Δικτύων 2023-2024

Κατασκευή Θεματικού Δικτύου με βάση την Wikipedia

Νικόλας Ευαγγελίδης, ΑΕΜ:10571

Εισαγωγή:

Στην εργασία αυτή γίνεται προσπάθεια και λόγος τόσο για την κατασκευή ενός δικτύου με βάση την Wikipedia και το αντίστοιχο API. Πριν περάσουμε στην ανάλυση του παραγώμενου δικτύου, είναι αναγκαίο να εξετάσουμε τις σημαντικές έννοιες κατά την κατασκευή του δικτύου αυτού που επηρεάζουν εν τέλει την δομή το και συνεπώς την εννοιολογική του αξία.

Αρχικά μιλάμε για την κατασκευή ενός κατευθυνόμενου γραφήματος με αρχικό κόμβο την σελίδα «Graph Theory» και ακμές τον κάθε σύνδεσμο προς άλλα άρθρα του Wikipedia που εμφανίζονται. Τα βάρη των ακμών, υπολογίζονται με την χρήση του γλωσσικού μοντέλου: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, από το οποίο παίρνουμε τις εμβυθίσεις προτάσεων έχοντας ως είσοδο την πρώτη παράγραφο του κάθε άρθρου. Έπειτα αυτή κανονικοποιείται στο [0,1].

Αν και δεν γράφεται ή μπορεί και να μην είναι ζητούμενο στα πλαίσια της εργασίας, θεωρώ σωστό τόσο για την πληρότητα, όσο και για την ορθή επεξήγηση, την παρουσίαση των βασικών δικτύων που εξετάστηκαν κατά τις δοκιμές, στην αναζήτηση ενός κατάλληλου αλγορίθμου, με βάση το ζητούμενο.

Αν δεν γίνει η επιλογή κάποιου αλγορίθμου αναζήτησης για την ίδια την κατασκευή του γραφήματος, είναι σαφές πως οι ακμές θα δημιουργούνται σειριακά ακριβώς με την σειρά που εμφανίζονται στο άρθρο. Τα παραπάνω κάνουν αναγκαίο τον ορισμό 3 παραμέτρων:

1. max_related_articles_per_article
2. max_nodes
3. similarity_threshold

Με τις δύο πρώτες διασφαλίζουμε το μέγεθος του γραφήματος, βάζουμε φράγμα σε μία απειροστική διαδικασία. Πιο συγκεκριμένα, η παράμετρος 1, χρησιμοποιήθηκε κατά κύριο λόγο για τις πρώτες δοκιμές, όπου τα άρθρα προπελάσονται σειριακά, δίχως κάποιον αλγόριθμο. Η 2, θέτει, όπως είναι προφανές, το μέγιστο για το μέγεθος του γραφήματος. Τέλος η 3, παίρνει τιμές από 0 μέχρι 1 και ελέγχει την σημαντικότητα των άρθρων, με βάση την εννοιολογική ομοιότητα όπως παράγεται από το γλωσσικό μοντέλο που αναφέρθηκε παραπάνω, που επιχειρείται να δημιουργηθούν ακμές.

Μοντέλα/ Δίκτυα που εξετάζονται:

1. Όπως ήδη αναφέρθηκε για αρχή παρουσιάζεται ένα γράφημα κατασκευασμένο δίχως κάποιον αλγόριθμο, έχει περιορισμένο ενδιαφέρον που θα αναλυθεί παρακάτω.
2. Γίνεται χρήση του αλγορίθμου Depth First Search και όπως είναι λογικό, δημιουργείται ένα γράφημα χωρίς μεγάλη εννοιολογική συγγένεια μεταξύ των άρθρων, παρόλα αυτά φαίνονται οι δυνατότητες για την σύνδεση εννοιών μέσω το δικτύου αυτού
3. Γίνεται χρήση του αλγορίθμου Breath First Search και όπως είναι λογικό εμφανίζεται δίκτυο με έννοιες πιο «κοντά» η μία στην άλλη. Σε αυτό βέβαια συμβάλλει και η αύξηση του similarity_threshold από 0.2 σε 0.5, η οποία επιλέχθηκε, μιας και το δίκτυο αυτό εμφανίζει το μεγαλύτερο ενδιαφέρον και αναπτύχθηκε σε 1000 κόμβους αντί για 500 των άλλων δύο.

Σημείωση: Οι αλγόριθμοι αυτοί μπορούν να εφαρμοστούν κατά την κατασκευή του δικτύου, ακριβώς επειδή τα δίκτυα που παρουσιάζονται μπορούν να θεωρηθούν ως στιγμιότυπα ενός νοητού, απειροστικού δικτύου που υπάρχει στις σελίδες της Wikipedia.

Ανάλυση των δικτύων

1. **Directed_Graph_With_Weights_From_Links**

Ένα πρώτο ζήτημα αναφορικά με το δίκτυο αυτό, είναι πως σε κάθε περίπτωση ο αριθμός των ακμών θα είναι ίσος με τον αριθμό των κόμβων, πλην ένα, λόγω του σειριακού τρόπου κατασκευής του. Επιπλέον, βλέπουμε ότι δημιουργούνται 33 κλάσεις, οι οποίες κατά κύριο λόγο αποτελούνται από έναν «κεντρικό» κόμβο ο οποίος κατά κανόνα συνδέεται με τον αρχικό κόμβο Graph Theory, ή με κάποιον ενδιάμεσο κόμβο με μεγάλη κεντρικότητα. Ταυτόχρονα καταλαβαίνουμε πως οι σχέσεις προγόνων-απογόνων είναι οι αναμενόμενες, με τον αρχικό κόμβο να είναι πρόγονος όλων και το ίδιο να ισχύει σε κάθε υποσύνολο του γραφήματος που πολλές φορές μπορεί να αποτελεί και μία από τις 33 κοινότητες.

Αντίστοιχα, εκτελώντας τους αντίστοιχους αλγορίθμους στο περιβάλλον του Gephi, βλέπουμε διάφορες χρήσιμες μετρικές όπως:

		Average Degree	0.998		
Modularity	0.903	Avg. Weighted Degree	0.421	Avg. Path Length	1.863
Statistical Inference	3266.563	Network Diameter	3		
		Graph Density	0.002		

Συμπερασματικά, βλέπουμε ένα πολύ αραιό δίκτυο, με μικρή διάμετρο, ενώ όλες οι παραπάνω μετρικές, όπως και πολύ περισσότερο το αντίστοιχο αρχείο .gexf, κάνουν ξεκάθαρο αυτό που υποθέσαμε αναφορικά με την σημαντικότητα συγκεκριμένων κορυφών

2. **<u>Directed Graph With Weights And Exploration</u>**

Στο δεύτερο δίκτυο χρησιμοποιείται ο DFS. Συνεπώς έχουμε ένα γράφημα αρκετά διαφορετικό από το προηγούμενο. Ο αλγόριθμος ουσιαστικά επιλέγει σχεδόν τον τελευταίο σύνδεσμο από όπου θα ξεκινήσει νέες συνδέσεις. Χαρακτηριστικό είναι το πρώτο παράδειγμα από το άρθρο Graph Theory στο άρθρο Carsten Thomassen (mathematician) και έπειτα στο Technical University of Denmark κ.ο.κ. Ταυτόχρονα είναι χαρακτηριστικό, τόσο από την ανάλογη οπτικοποίηση όσο και από τις μετρικές η μεταβολή του «ρόλου» του αρχικού κόμβου. Δηλαδή, πέρα από τις μετρικές του Βαθμού και του Eccentricity, η αρχική κορυφή δεν είναι σημαντική σύμφωνα με τις άλλες μετρικές. Βλέπουμε επίσης 19 κοινότητες.

Όπως ήταν επίσης αναμενόμενο το γράφημα αυτό επιχειρεί να δημιουργήσει τις πιο «μακρινές» εννοιολογικά συνδέσεις με άρθρα, έχοντας βέβαια τους γνωστούς περιορισμούς. Ουσιαστικά κάθε κόμβος συνδέεται με τα 49 πρώτα άρθρα, τα οποία δεν πραγματοποιούν καμία επιπλέον σύνδεση και έπειτα συνδέεται με το 50κοστό, για το οποίο επαναλαμβάνεται η διαδικασία. Επομένως, η διάμετρος είναι πολύ μεγαλύτερη από το προηγούμενο (23), αλλά και απόγονοι είναι όλοι οι κόμβοι πλην του αρχικού, ενώ πρόγονοι για υπόλοιπους κόμβους είναι οι «κύριοι» κόμβοι για τους οποίους διενεργείται η διαδικασία που μόλις περιγράψαμε.

Τέλος να σημειώσουμε πως είναι εξαιρετικά ενδιαφέρον το ότι το δίκτυο αγγίζει έννοιες και ετικέτες όπως: Dirt Bike Racing, Carbon Dioxide, Destructive Fishing Practices, κ.ο.κ, το οποίο μπορεί να μην αποτελεί αρκετά χρήσιμο εργαλείο αλλά σίγουρα δείχνει τις δυνατότητες των εργαλείων που χρησιμοποιήθηκαν για την μοντελοποίηση προβλημάτων του πραγματικού κόσμο, με προϋπόθεση, την εύρεση του κατάλληλου αλγορίθμου.

3. **<u>Directed Graph With Weights And Exploration BFS</u>**

Στο τρίτο δίκτυο, αφού τρέξουμε τους πρώτους συνηθισμένους αλγορίθμους και πολύ περισσότερο την ομαδοποίηση, βλέπουμε εντυπωσιακά αποτελέσματα τα οποία σίγουρα εμφανίζουν εννοιολογική συγγένεια στα πλαίσια των κοινοτήτων που σχηματίζονται. Σχηματίζονται 19 κοινότητες εκ των οποίων μερικές θα μπορούσαν να έχουν τα εξής ονόματα. Πράσινη: Χρωματισμός Γράφων, Ώχρα: Circle Packing, Κόκκινο: Graph Drawing, Μπλε: Directed Graph κ.ο.κ. Αυτό είναι και το βασικό χαρακτηριστικό για την επιλογή του δικτύου αυτού ως τελικού. Ταυτόχρονα έχουμε διάμετρο 7 και κατανομή βαθμών με μία ίσως υπερβολική συνάρτηση, όπως είναι αναμενόμενο. Βλέπουμε επίσης, σύμφωνα σχεδόν με όλες τις μετρικές την κομβικότητα των άρθρων με τίτλους Graph (discrete mathematics), Simple Graph, Undirected Graph κλπ. Δηλαδή μπορούμε να παρατηρήσουμε πως το

γράφημα, δημιούργηε «κέντρα» εννοιολογικά παρόμοια με την αρχική κορυφή, χωρίς να διατηρηθεί η ίδια η κορυφή ως κεντρική όπως είδαμε στην αρχή.

Σημειώσεις:

Όλα τα scripts έτρεξαν στο Google Colab για λόγους ευκολίας

Για κάθε ένα από τα τρία γραφήματα παρέχεται τόσο ο κώικας, το csv αρχείο, μία απεικόνιση του δικτύου σε μορφή pdf, όσο και το αρχείο .gephi.

Νικόλας Ευαγγελίδης, ΑΕΜ:10571

Απρίλης 2024