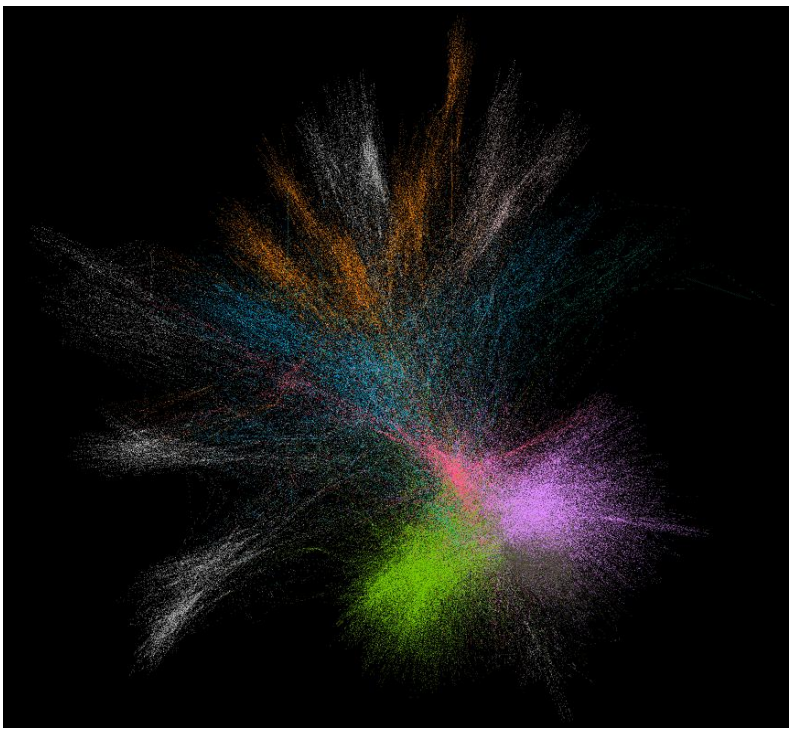
ΘΕΩΡΙΑ ΔΙΚΤΥΩΝ



Εργασία 2020-2021

ΝΙΚΗΦΟΡΙΔΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ 9084

nikifori@ece.auth.gr

Θεσσαλονίκη, Ιανουάριος 2021

Θεματικό Δίκτυο

Σκοπός της εργασίας αυτής, είναι να δημιουργηθεί ένα Θεματικό Δίκτυο με συναφή θέματα, βάσει των συνδέσμων (μόνο εντός Wikipedia) που περιέχονται σε κάθε σελίδα – άρθρο στην Wikipedia.

Η κεντρική ιδέα με την οποία γράφτηκε ο κώδικας στο αρχείο **Thematic_Graph_Maker.py** είναι η εξής:

Να δημιουργηθεί ένα κατευθυνόμενο θεματικό δίκτυο με αρχική ρίζα το θέμα **Tennis.** Στη συνέχεια, δημιουργήθηκε η συνάρτηση **links_nodes_edges(key_word)** η οποία με βάση την τιμή του key_word, μπαίνει στην αντίστοιχη σελίδα στο Wikipedia και βρίσκει τα links τα οποία περιέχονται μέσα στο άρθρο που αναφέρεται στη λέξη key_word. Έπειτα, από αυτά τα links παίρνει τυχαία 350 links, ένα αρκετά μεγάλο και αποτελεσματικό νούμερο, και ελέγχει στο περιεχόμενο τους (content) αν περιέχουν πάνω από 25 φορές τη λέξη tennis. Επιλέχθηκε το 25 για να παίρνουμε όσο γίνεται πιο σχετικά άρθρα με το αρχικό μας θέμα. Επίσης, η τιμή αυτή επιλέχθηκε μετά από αρκετά πειράματα. Τελικά, όσα άρθρα έχουν πάνω από 25 φορές τη λέξη tennis μέσα στο περιεχόμενο τους αποτελούν τα παιδιά της κορυφής key_word. Όπως φαίνεται, δημιουργείται μία δενδροειδής μορφή και όσο κατεβαίνουμε επίπεδα ο αριθμός των κόμβων – άρθρων – links αυξάνεται εκθετικά. Συνεπώς, για εξοικονόμηση χρόνου αρκεστήκαμε στο να τρέξουμε την παραπάνω συνάρτηση για 4 επίπεδα. Το αποτέλεσμα όπως θα δούμε και παρακάτω πιστεύω είναι αρκετά ικανοποιητικό.

Αφού τρέξαμε τον κώδικα για αρκετές ώρες, δημιουργήθηκε ένα δίκτυο με 217 κορυφές (άρθρα) και 2275 ακμές. Επίσης, η παραπάνω συνάρτηση έτρεξε περίπου 5000 φορές, το οποίο σημαίνει ότι έτρεξε και αρκετές φορές για το ίδιο άρθρο, κάτι το οποίο δείχνει και το ότι η συνάρτηση επέλεγε αρκετά σχετικά άρθρα με το αρχικό θέμα μας. Προφανώς, αυτό δε σημαίνει ότι δημιουργούνταν πολλές φορές η ίδια κορυφή, αλλά το ότι προσθέτονταν σε αυτή νέες ακμές με άλλες

κορυφές. Να αναφερθεί ότι αρκεστήκαμε σε έναν χαμηλό αριθμό κορυφών ώστε βλέποντας το θεματικό δίκτυο στο gephi να μπορούμε να κάνουμε έναν ποιοτικό σχολιασμό των άρθρων που επιλέχθηκαν.

Επιπλέον, δημιουργήθηκε το αρχείο **Graph_Statistics.py** ώστε να παίρνουμε διάφορες μετρήσεις για το θεματικό δίκτυο που δημιουργήσαμε. Επιπλέον, παρόμοιες μετρήσεις μας δίνει και το gephi. Τελικά, τα αποτελέσματα και το πόρισμα των παραπάνω μετρήσεων παραθέτονται παρακάτω:

- **Graph_statistics.py**
- Κατευθυνόμενος γράφος: **Ναι**
- Αριθμός κορυφών: **217**
- Αριθμός ακμών: **2275**
- Μέσος βαθμός κορυφής: **10.4839** (προφανώς έσω και έξω βαθμός είναι ίσοι)
- Πυκνότητα ακμών δικτύου σε σχέση με τον μέγιστο αριθμό ακμών που θα μπορούσε να έχει με 217 κορυφές: **0.0485**
- Μεταβατικότητα: **0.22429** (ή αλλιώς τριαδικό κλείσιμο)
- Αριθμός ακμών που γυρνάν στην ίδια την κορυφή (selfloops): **5**
- Μέσος συντελεστής ομαδοποίησης για όλες τις κορυφές: **0.2387**
- Top 10 degree centrality vertices

Top 10 degree centrality vertices	
Tennis	0.60648
Grand Slam (tennis)	0.52314
The Championships, Wimbledon	0.47222
Rafael Nadal	0.40277
Novak Djokovic	0.39351
Björn Borg	0.38425
Andy Roddick	0.37962
Pete Sampras	0.365740
Jimmy Connors	0.36111
Roger Federer	0.351851

- Top 10 eigenvector centrality vertices

Top 10 degree centrality vertices	
Association of Tennis Professionals	0.207590
The Championships, Wimbledon	0.19789
Grand Slam (tennis)	0.18525
US Open (tennis)	0.18481
Tennis	0.176695
International Tennis Federation	0.164920
Fred Perry	0.15501
History of tennis	0.15280
Novak Djokovic	0.15208
John McEnroe	0.148263

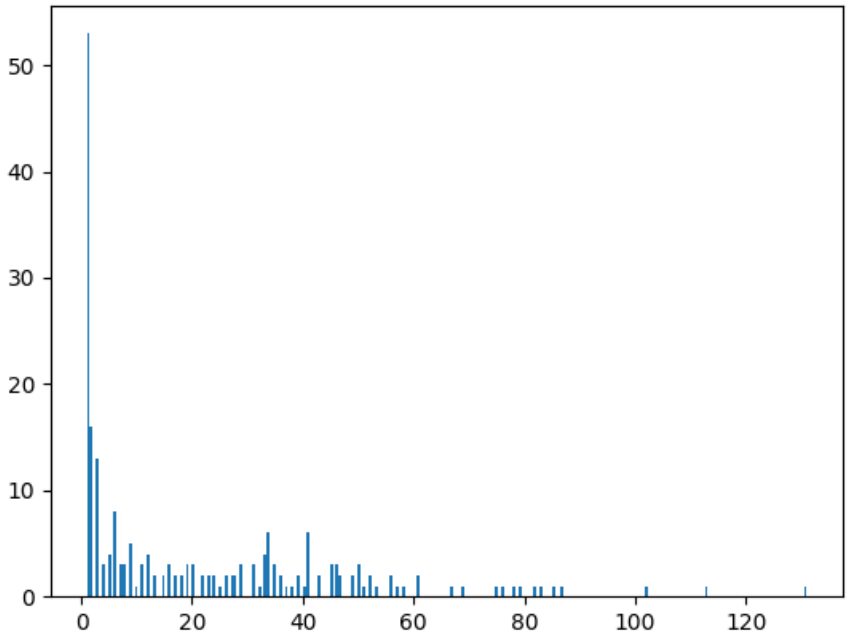
- Top 10 closeness centrality vertices

Top 10 degree centrality vertices	
Association of Tennis Professionals	0.296771
The Championships, Wimbledon	0.29488
US Open (tennis)	0.293014
Tennis	0.29117
Grand Slam (tennis)	0.29117
International Tennis Federation	0.28333
History of tennis	0.27889
Novak Djokovic	0.27889
Rafael Nadal	0.27722
Roger Federer	0.26916

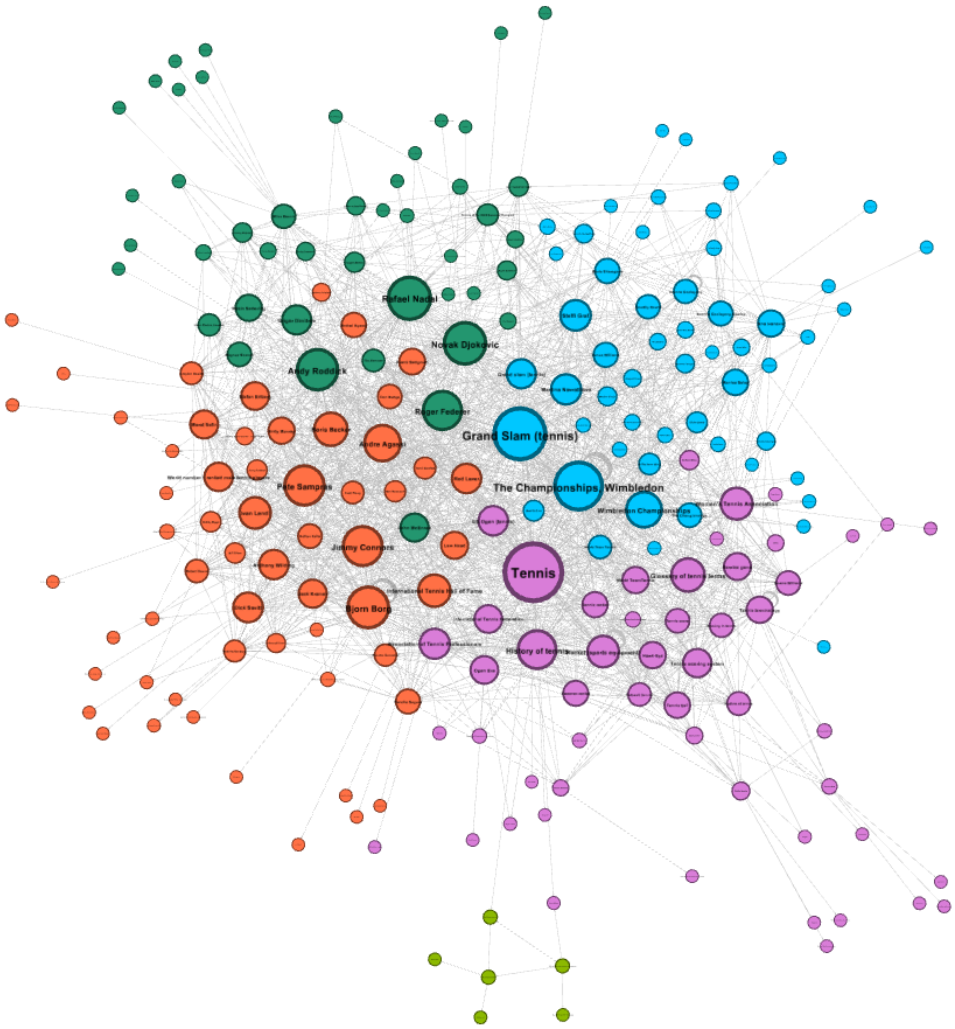
- Top 10 betweenness centrality vertices

Top 10 degree centrality vertices	
Tennis	0.087348
Grand Slam (tennis)	0.03175
The Championships	0.027549
Andy Roddick	0.02566
Rafael Nadal	0.02492
Glossary of tennis terms	0.0234915
Roger Federer	0.022481
Novak Djokovic	0.022237
Pete Sampras	0.018118
Jimmy Connors	0.01764

- Κατανομή βαθμών κορυφών



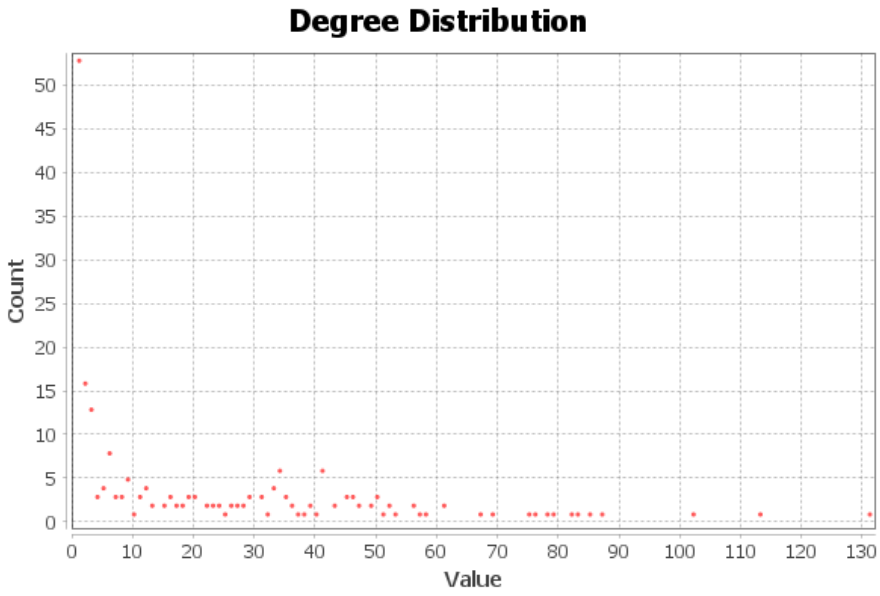
- **Gephi**
- Κατευθυνόμενος γράφος: **Ναι**



- Αριθμός κορυφών: **217**
- Αριθμός ακμών: **2275**
- Μέσος βαθμός κορυφής: **10.4839** (προφανώς έσω και έξω βαθμός είναι ίσοι)

Results:

Average Degree: 10.484



- Πυκνότητα ακμών δικτύου σε σχέση με τον μέγιστο αριθμό ακμών που θα μπορούσε να έχει με 217 κορυφές: **0.049**
- Modularity – Αρθρωτότητα: **0.264**
- Αριθμός ακμών που γυρνάν στην ίδια την κορυφή (selfloops): **5**
- Μέσος συντελεστής ομαδοποίησης για όλες τις κορυφές: **0.227**
- Μέσο μήκος συντομότερων μονοπατιών: **2.4528**
- Διάμετρος δικτύου: 6
- Ακτίνα δικτύου: 0

- **Πόρισμα**

- Από την πυκνότητα ακμών βλέπουμε ότι έχουμε ένα σχετικά αραιό δίκτυο με μέσο βαθμό κορυφής 10.4839. Σχετικά με όλα τα πειράματα που έγιναν το 10.4839 είναι ένα σχετικά ψηλό νούμερο.
- Γίνεται φανερό ότι η μεταβατικότητα του δικτύου είναι διάφορη του μέσου συντελεστή ομαδοποίησης.
- Στον πίνακα με τις 10 σημαντικότερες κορυφές με βάση τον βαθμό σημαντικότητας βλέπουμε ότι προφανώς στο νούμερο 1 είναι το θέμα μας (Tennis) , αφού είναι και η λέξη που ελέγχουμε

αν υπάρχει πάνω από 25 φορές στο content της υποψήφιας κορυφής για να προστεθεί στο γράφημα.

- Στον πίνακα με τις 10 σημαντικότερες κορυφές με βάση τη κομβικότητα ιδιοδιανύσματος βλέπουμε ότι έχουμε κορυφές που έχουν σημαντικούς γείτονες ή έχουν πολλούς όχι και τόσο σημαντικούς γείτονες.
- Στον πίνακα με τις 10 σημαντικότερες κορυφές με βάση την κομβικότητα απόστασης βλέπουμε ότι έχουμε κορυφές οι οποίες έχουν την μικρότερη μέση απόσταση από όλες τις υπόλοιπες κορυφές.
- Στον πίνακα με τις 10 σημαντικότερες κορυφές με βάση την σημαντικότητα θέσεως έχουμε τις κορυφές που βρίσκονται σε πολλά συντομότερα μονοπάτια που ενώνουν 2 άλλες κορυφές.
- Η κατανομή των βαθμών των κορυφών στη python προφανώς είναι ίδια με αυτή του gephi.
- Στην κατανομή των βαθμών βλέπουμε ότι ένας μεγάλος αριθμός κορυφών έχει βαθμό < 5.
- Παρόλο που το modularity είναι μικρό, το gephi έκανε έναν αρκετά καλό διαχωρισμό του δικτύου που μπορούμε άνετα να τον εξηγήσουμε ποιοτικά.
- Με πράσινο χρώμα, είναι κυρίως άντρες τεννίστες της τωρινής εποχής. Με πορτοκαλί χρώμα, φαίνονται κυρίως άντρες τεννίστες παλιάς εποχής, δηλαδή που έχουν σταματήσει να παίζουν. Με μπλε χρώμα, φαίνονται κυρίως γυναίκες τεννίστριες ανεξαρτήτου εποχής . Με μοβ χρώμα, φαίνονται κυρίως άρθρα που έχουν σχέση με το Tennis χωρίς να είναι απαραίτητα άνθρωποι όπως πχ, κανόνες του Tennis και ιστορία του Tennis. Τέλος με λαχανί χρώμα, φαίνονται άρθρα σχετικά με ένα ιαπωνικό cartoon που έχει βασικό θέμα το Tennis.
- Γενικότερα, βλέποντας το γράφημα στο gephi φαίνεται να έχουν μπει στο θεματικό δίκτυο αρκετά σχετικά θέματα με το Tennis, όπως αυτός ήταν και ο σκοπός μας.

Τέλος