TEL AVIV UNIVERSITY

stackoverflow

קורס רשתות מורכבות – פרויקט מסכם
מרצה: דר' אסף אלמוג

מגישים: נדב שטרן (203016100) וישי שפירא (203016217)
תאריך הגשה: 22/06/21

תקציר מנהלים

בפרויקט זה ניתחנו נתונים מהאתר Stack Overflow, הנתונים הם שאלות מתחום התוכנה בשפות Python, R, Java, JavaScript ו-PHP, שאלות אלו נשאלו בשנת 2020.

מטרת הפרויקט והמוטיבציה לבחור נושא זה היא לנסות למצוא דמיון בין השאלות השונות באתר ולנסות למנוע חזרתיות, דבר זה יכול להועיל גם למשתמשים וגם לאתר עצמו.
בנוסף רצינו לחקור האם ניתן לקלסטר את הנתונים לפי השפות השונות, למצוא קשרים מעניינים בתתתי הנושאים בכל שפת תכנות ובין השפות השונות.

בנינו שתי רשתות בעזרת הנתונים שבידינו וביצענו ניתוחים מעולם ה-complex networks, הנושא בו עסקנו בקורס.

הרשת המרכזית היא רשת השאלות בה כל צומת היא שאלה והקשר בין השאלות התאפיין על ידי מדד דמיון בין רשימת השאלות של כל שאלה, מדד זה מבוסס על המדד Jaccard Index עם שינויים שעשינו בשביל להתאים את המדד למקרה שלנו, ביצענו ניתוח threshold בשביל להחליט מעל איזה סף של דמיון אנו מקשרים בין 2 שאלות, החלטנו שהסף הנכון הוא 0.6.
ברשת זו קיבלנו תוצאות מעניינות, חלקן צפויות וחלקן פחות, כמו שציפינו, התפלגות הדרגות לא הייתה power law, הגיוון בנושאים בשאלות הוא גדול מאוד ואין ציפייה שייוצרו hubs – ששאלות מסוימות יהיו קשורות לאלפי שאלות אחרות, מסיבה זו גם קיבלנו רשת לא מחוברת.

בניתוח הויזואלי של הרשת אפשר לראות בקלות שקיבלנו רשת שמורכבת מחמישה חלקים, כל חלק הוא באמצע שפה אחרת, דבר שגרם לנו לחשוב על אפשרות לקסלטר את הנתונים לשפות השונות, ביצענו community detection בעזרת greedy modularity, ושמנו לב שהקלסטרים שקיבלנו דומים מאוד לחמשת השפות השונות בהן עוסקות השאלות, השווינו את הקלסטרים שקיבלנו לחלוקה לחמשת השפות וקיבלנו דיוק של 83%! במילים אחרות הצלחנו לענות על האתגר של Kaggle לסווג את השאלות לשפות השונות.
השלב הבא היה להתמקד בתוך כל קלסטר (שפה), הצלחנו לבצע זיהוי לתתי הנושאים בכל שפה.

בשביל לקבל זוית נוספת על הנתונים וללמוד עוד על האתר בנינו רשת נוספת – רשת התגיות, רשת זו היא הרשת ההפוכה של רשת השאלות כאן הצמתים הן התגיות השונות והקשר בין התגיות חושב על ידי השאלות המשותפות.
רשת זו היא רשת מחוברת, כל תגית יכולה להגיע לכל תגית אחרת במקסימום 4 צעדים, ההסבר שלנו לתכונת העולם הקטן שקיבלנו היא שישנן תגיות שרלוונטיות לתחומים ושפות רבות, תגיות אלה בעצם חיברו בין הנושאים השונים.
ההתפלגות של רשת זו הייתה power law ברוב טווח הדרגות, דבר המעיד על הימצאותם של hubs, בניתוח הויזואלי של רשת זו ראינו שחמשת ה-hubs הם חמשת השפות שלנו (שגם שימשו כתגיות), תגיות אלה גם היו עם ערכי betweenness גדולים, ההסבר שלנו לכך הוא שהמשמעיות של תגיות אלה הן כלליות וקשורות להרבה תגיות אחרות, כך תגיות אלה בעצם מקשרות בין המון תגיות אחרות.

דבר מעניין ששמנו לב אליו שבחנו את הנתונים של שתי הרשתות הוא שבהרבה פרמטרים קיבלנו תוצאות הפוכות כמעט לחלוטין, ההסבר שלנו לתוצאות אלה הוא שרשתות אלה הפוכות לחלוטין, מה שמשמש ברשת הראשונה כצמתים ברשת השניה מאפיין את הקשתות ולהפך, כך שבעצם רשת אחת היא מעין הרשת הדואלית של הרשת השניה.

בשביל לקחת את הרשת הזו צעד אחד קדימה וממצב שבו אנו זהים שפות ותתי נושאים בתוך כל שפה על ידי הרשת למצב בו אנו ממש מחברים שאלות זהות אנו צריכים לשכלל את מדד הקורלציה בין השאלות על ידי הוספה של נתונים חדשים שלא זמינים לנו כרגע בדאטה סט שבידינו.

תוכן עניינים

מבוא

תיאור הנתונים

Stack Overflow הוא אתר אינטרנט פרטי פופולרי שנוצר ב-2008, העוסק בתחום שאלות ותשובות בתחומי המיחשוב ופיתוח תוכנה וחומרה. הוא אתר הדגל של רשת Stack Exchange והוא אחד האתרים הפופולריים בעולם.

הנתונים בהם השתמשנו נלקחו מהאתר Kaggle, כל רשומה זו שאלה מהאתר, הנתונים כוללים את השדות הבאים:

- Stack id - שדה מזהה של שאלה
- Title - תיאור של השאלה
- Tags – רשימה של תגיות שקשורות לשאלה
- Views – כמות הצפיות של השאלה
- Score – הדירוג הממוצע של השאלה על ידי הגולשים
- Done – האם השאלה נענתה (כן/לא)
- Label – השפה/ פלטפורמה בה השאלה עוסקת

מצאנו שלושה דאטה סטים שונים עם מידע זה, איחדנו את שלושתם לדאטה סט אחד והסרנו כפילויות של רשומות עם אותה שאלה.

מוטיבציה לפרויקט

במסגרת הפרויקט רצינו לחקור את הקשרים בין השאלות השונות, השאלה המרכזית שהתמקדנו בה היא האם ניתן למצוא דמיון בין שאלות שונות וכך ללמוד מהתשובות הקיימות בשאלה אחת על שאלות אחרות שטרם נענו, במידה ואכן קיימים קשרים מסוג זה, זה יכול לעזור למשתמשי האתר בקבלת מידע ומענה על שאלותיהם גם אם לא קיבלו מענה מספק בצורה ישירה, דבר זה יכול לשמש גם כבסיס למודל חיזוי תשובה או הצעה של תשובה אפשרית לשאלה ברגע שהשאלה נשאלת, מבלי לחכות לתשובה אנושית, מעין מערכת אחזור מידע שתוכל לעזור בצורה מיידית למשתמשים עד לקבלת מענה אנושי.
בנוסף רצינו לחקור את מבנה השאלות והאתר, האם יש שאלות מרכזיות שקשורות להמון שאלות (hubs)? האם אפשר לקטלג את השאלות לקטגוריות שונות? ואיך המאפיינים השונים שיש לנו בדאטה סט באים לידי ביטוי ברשת.

תיאור הרשת

כמו שהזכרנו לעיל, השאלות השונות מייצגות את הצמתים, קשרים בין שאלות יבואו לידי ביטוי על ידי דמיון של תגיות משותפות, ככל שיש יותר תגיות משותפות ככה הדמיון גדול יותר.
בשביל למדוד את בדמיון בין השאלות בחנו מספר מדדי דמיון מתורת הקבוצות, ביניהם Dice coefficient, Tversky index ו-Jaccard index, בחרנו להשתמש ב-Jaccard index, הנוסחה למדד זה היא:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

לאחר בחינה של הדמיון שקיבלנו ראינו שמדד זה לא מתייחס לכמות התגיות המשותפות אלא לאחוז התגיות המשותפות, דבר זה גרם לכך ששאלות עם תגית אחת בלבד קיבלו דמיון מקסימלי השווה ל-1, דבר זה לא מעיד בהכרח על כך ששתי השאלות דומות ולכן ערכנו מדד זה והוספנו חזקה למונה – מספר התגיות המשותפות והוספנו כלל למכנה שאם האיחוד של התגיות שווה ל-1 (משמע יש רק תגית אחת לכל שאלה) אז המכנה שווה ל-2 ולא ל1, באופן זה התייחסנו לכמות התגיות המשותפות ולא רק לאחוז התגיות המשותפות, קיבלנו את המדד הבא:

$$J_{modified}(A,B) = \frac{|A \cap B|^{2}}{max\{|A \cup B|, 2\}}$$

בשביל לקבל זוית נוספת על האתר ועל הנתונים שבידינו בנינו רשת נוספת, רשת התגיות, רשת זו היא בעצם רשת הפוכה לרשת המרכזית שבנינו, נפרט על רשת זו בפרק הרלוונטי.

ניתוח נתונים

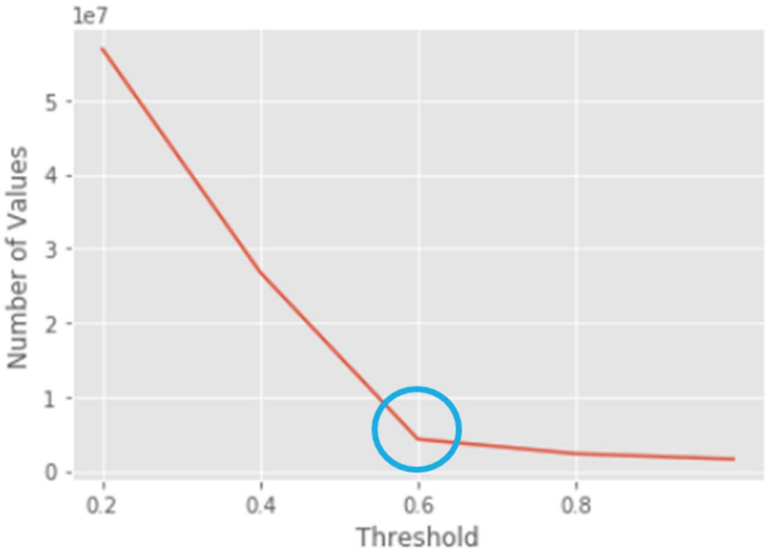
מבנה הרשת

על מנת לקצר זמני ריצה ותוך הבנה כי מספר הצפיות בשאלה מהווה פילטר הגיוני לחשיבות השאלות חתכנו את הנתונים אשר יש להם פחות מ50 צפיות. כעת יש לנו 15,662 נודים.

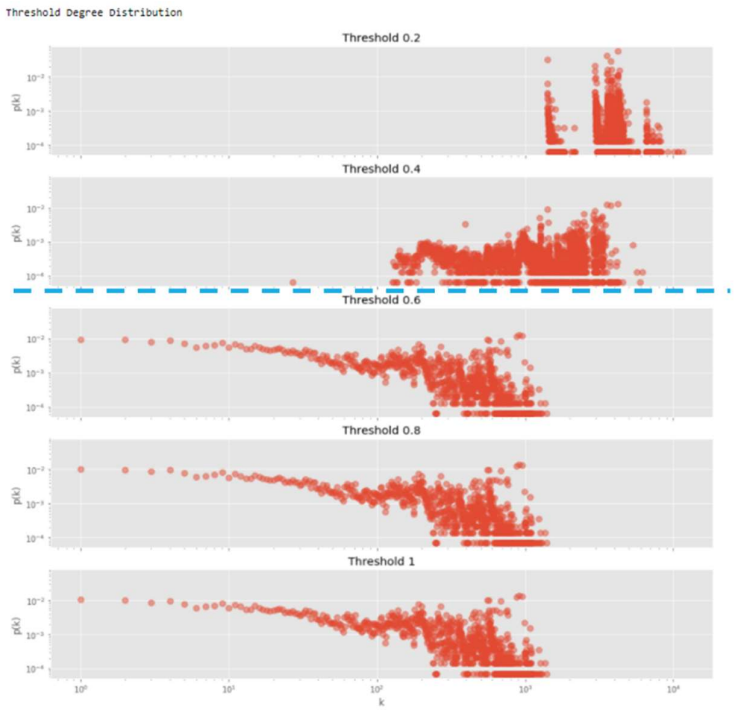
בנינו את הרשת כאשר כל נוד זה שאלה והקשר בין שתי שאלות הוא מאופיין בתגיות משותפות והקורלציה מחושבת באמצעות הפרמטר $J_{modified}(A,B)$, הרשת היא רשת לא מכוונת מכיוון שקשר בין שתי שאלות הוא סימטרי, הרשת היא רשת ממושקלת כאשר המשקל של כל קשת הוא הוא ערך הקורלציה המוזכר לעיל.

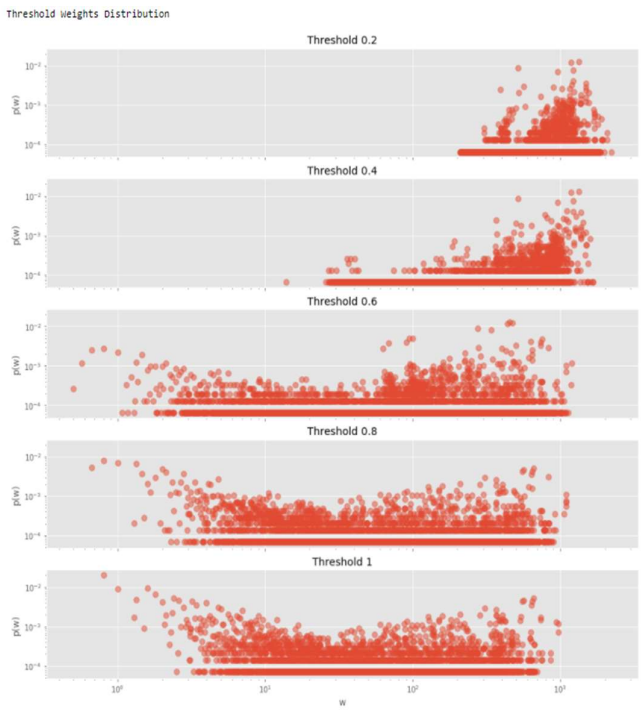
ניתוח threshold

תוך כדי בניית הרשת השמנו בסיפריות בעלות פילטר קורלציה משתנה את האייטמים הרלוונטיים.
כך שיצרנו 5 רשתות עם ספים משתנים שנעו מ0 עד 1 בקפיצות של 0.2.
חשוב לציין שהמדד שלנו עולה מעל 1 אך לרוב מתפלג בין 0 ל1 ולכן התמקדנו בתחום הזה.
בגרף הבא ניתן לראות את מספר התצפיות כתלות בסף הנבחר. הברך נמצא בסף 0.6.



כעת בדקנו את התפלגות K עבור כל סף. ניתן לראות שתי סדרות של חמש גרפים. מימין גרף התפלגות משקולות ומשמאל גרף התפלגות דרגות. ניתן לראות כי בסף נמוך מ- 0.6 אנחנו מקבלים גרף fully connected שלא נותן לנו מידע. מעל 0.6 אנחנו מקבלים התפלגות והיא לא משתנה ככל שאנחנו מגדילים את הסף ולכן נבחר את הסף המינימלי שנותן לנו את ההתפלגות שאיננה fully connected. לכן נבחר בו – סף של 0.6 בהמשך הפרוייקט.





נתוני הרשת עם ה-threshold שנבחר

מספר הצמתים הוא N = 15,516.

מספר הקשרים לפי ה-threshold = 0.6 שנבחר הוא L = 2,137,796.

הדרגה הממוצעת היא <K> = 275.5.

C = 0.7452 - זהו מקדם גבוה שנובע מהחיבוריות הגבוהה ברשת.

הרשת שקיבלנו לא מחוברת, ישנן שאלות שכלל לא מחוברות לשאלות אחרות, דבר הגיוני כי יש שאלות מנושאים שונים ואין סיבה שיהיה ביניהן קשר, אפילו לא עקיף, ולכן קוטר הגרף הוא אינסופי.

הדרגה המקסימלית ברשת היא 1,403 והדרגה המינימלית היא 1.

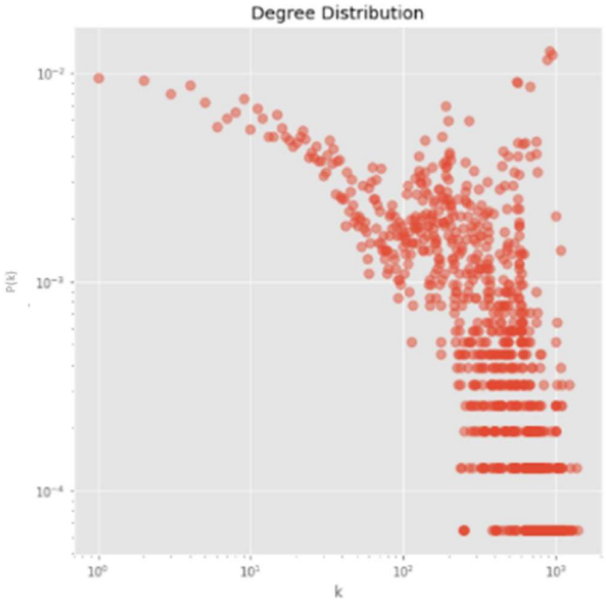
הרכיב הכי גדול ברשת הוא בגודל 15,499, נתון זה הפתיע אותנו מכיוון שיש שונות גבוהה בשאלות – שאלות משפות שונות ותחומים שונים (web, ניתוח נתונים ועוד), ציפינו לקבל מספר רכיבים, שכל אחד יהיה תת רשת של תחום אחר, ההסבר שלנו לנתון הוא שלמרות השוני והגיוון בשאלות ישנן תחומים גנריים אשר מחברים בין התחומים השונים, דוגמא לתחום כזה הוא בסיסי הנתונים ואכן יש לנו שאלות בנושא זה.

התפלגות דרגות P(K)

בגרף מטה ניתן לראות התפלגות בקנה מידה לוג לוג. ניתן לזהות שהרשת איננה מתנהגת בהתפלגות power low. בגרף אפשר לראות פיזור יחסי עם התקבצות של נתונים עם דרגה בין מאה לאלף והסתברויות להתחבר לK בתחום רחב מאוד, בנוסף מהגרף ניתן להסיק כי ככל ההדרגה של הצומת קטנה יותר ההסתברות שלה גדולה יותר.

זה הגיוני ברשת זו משום ששאלות עם חיבוריות גבוהה הן נדירות יחסית, לכל שאלה יש בערך שש תגיות, במצב זה אין אפשרות לקבל hubs -ששאלה תיהיה קשורה לאחוז גבוה של שאלות מכלל השאלות בנתונים, זה גם הגיוני כי שאלה היא על תחום מסוים או בעיה מסוימת ואין סיבה דבר זה יהיה רלוונטי לכל כך הרבה שאלות.

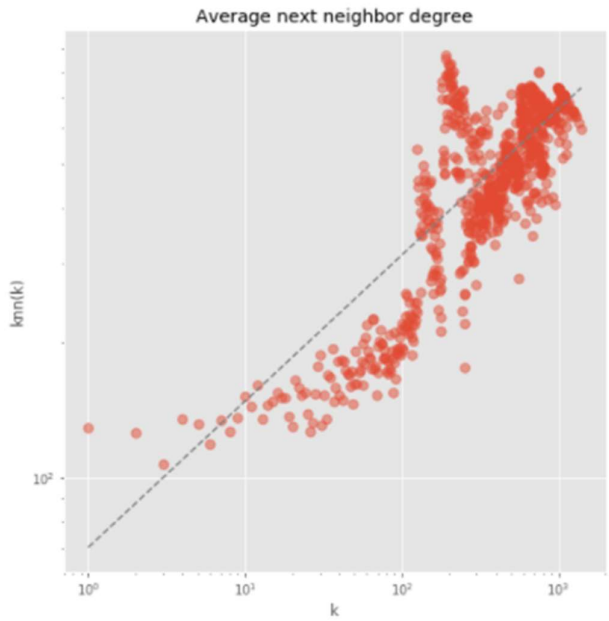
כלומר שאלות עם הרבה תגיות נפוצות יהיו בעלות קשרים מרובים ולכן בעלות דרגה גבוהה אך המצב הנל איננו נפוץ. לרוב שאלות יהיו בעלות מעט תגיות ולכן המדד שלנו $J_{modified}(A,B)$ אשר מנסה להבליט את השאלות בעלות קשר משמעותי עם כמה שיותר תגיות משותפות ולהקשות על הקשר בין שאלות עם מעט תגיות יחזק את ההתפלגות שאנו מקבלים עם נטיה יחסית ריבועית שלילית.

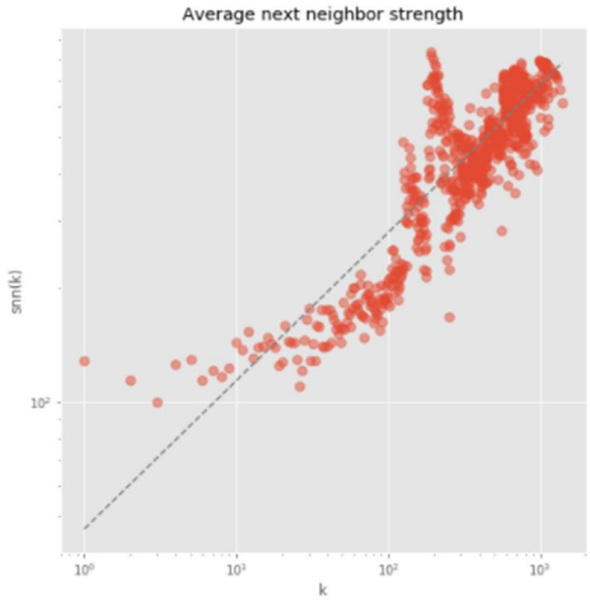


Degree Correlation

נרצה לבחון את הקורלציה בין דרגות הצמתים ברשת שלנו, קיבלנו שמדד ה-assortative של הרשת שלנו הוא r = 0.17302, המשמעות היא שהרשת שלנו היא assortative מכיון ש-r גדול מאפס, אפשר לראות זאת בגרף שאכן ככל שהדרגה של הצומת גבוהה יותר, כך הסיכוי שהשכנים הישירים של אותה צומת יהיו גם בעל דרגה גבוהה גדול יותר, המשמעות היא שככל הדרגה של הצומת גבוהה יותר כך הדרגה הממוצעת של השכנים הישירים גם גבוהה יותר.

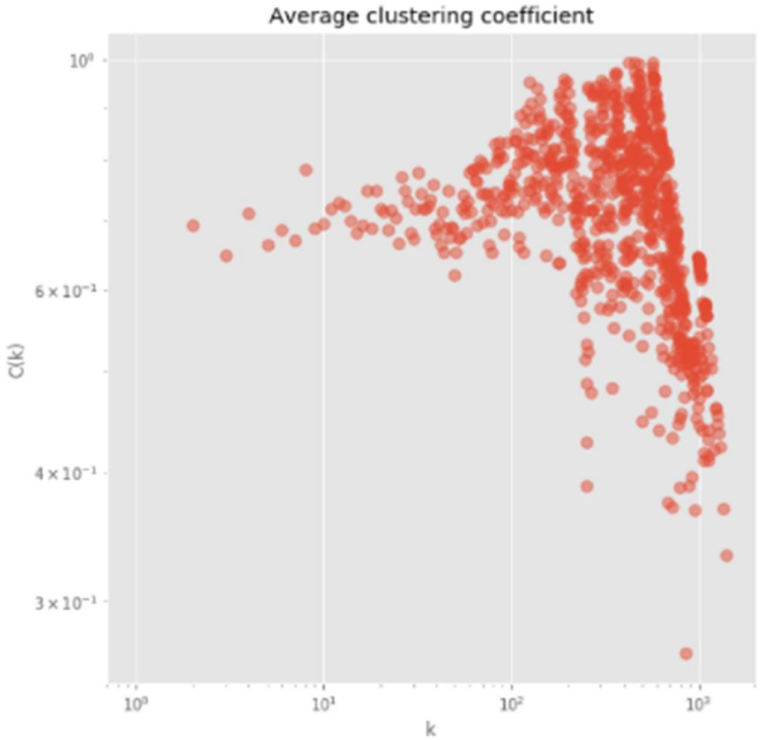
תוצאה זו לא הפתיעה אותנו מכיון שישנן תגיות פופולריות שקיימות ביותר שאלות, שאלות אשר יש להן תגית אחת כזו או יותר יקבלו קורלציה גבוהה עם הרבה שאלות אחרות שגם הן יהיו מחוברות להרבה שאלות, יוצא שנוצרים בעצם מעין "מועדונים" של שאלות מדרגות שונות, ככל שהדרגות יותר גבוהות – תגיות יותר פופלאריות, כך ה"מועדונים" גדולים יותר.





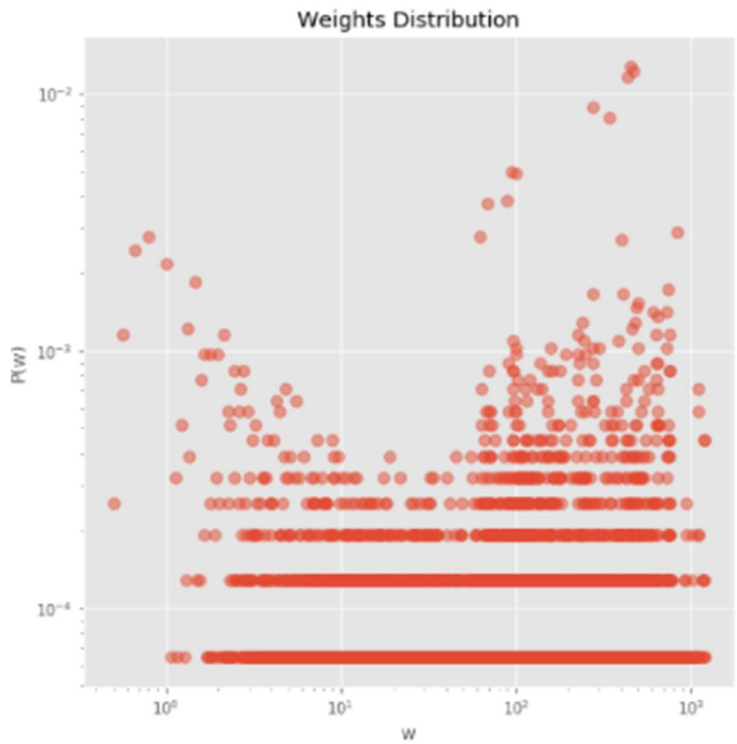
Average clustering coefficient

בגרף הבא הבוחן את ממוצע מקדם ההתקבצות כתלות בדרגה ניתן לראות כי מרבית הנקודות מפוזרות.
כלומר לא קיים קשר מובהק בין הדרגה לבין ממוצע מקדם ההתקבצות של כל דרגה, זאת אומרת שאי אפשר לומר שככל שהדרגה גבוהה יותר כך השכנים שלי מחוברים יותר/פחות.
מכיוון שלרוב השאלות מקדם ההתקבצות גבוה המסקנה היא שהרשת דחוסה יחסית. לא קיימת מגמה ליניארית בגרף כלומר לא ניתן לומר כי קיימת היררכיה בגרף. מקדם ההתקבצות ההמוצע הוא C = 0.7452.



התפלגות המשקולות

כמו שהזכרנו, הרשת שלנו הינה רשת ממושקלת, בדקנו את התפלגות המשקולות לפי הסף שנבחר, המשקולות אצלנו זה הקורלציה לפי $J_{modified}(A,B)$. ההתפלגות היא יחסית אקראית ולא דומה לשום התפלגות מוכרת, דבר זה נכון גם לספים האחרים שבדקנו, כפי שניתן לראות בגרפי התפלגות המשקולות בפרק הthreshold לעיל.



ניתוח Centrality

ביצענו ניתוח סנטרליטי באמצעות Page Rank ו- closeness centrality על מנת לזהות את הנודים המשמעותיים ברשת שלנו.
ראשית כל ניתן לזהות כי השפות(הלייבלים) העיקריות שהיו הכי משמעותיות בניתוח ה-centrality הן PHP ו JavaScript. כלים שמאוד קשורים לפיתוח web. סביבה עתירת כלים ואינטגרציה ביניהם, לכן יחסית צפוי כי הן יהיו הלייבלים המובילים במדדים שלנו.

ניתוח *Page Rank*.

בניתוח, השאלות שזוהו כמרכזיות הן למעשה שאלות עם מספר טיוגים המשתייכים ישירות ללייבלים. טיוגים נפוצים יחסית אך בשאלות אלה יש מספר טיוגיפ נפוצים. PHP, JavaScript. בנוסף השאלות הן מאוד גנריות ורלוונטיות להמון כלים אתרים/אפליקציות שמפתחים ברמה היומיומית.
למשל Mobile number and message text checking. שאלה לגבי ולידציה מאוד נפוצה שקיימת היום בכל ממשק.

title	id	tags	views	score	done	language
How to preview an image before and after upload?	16207575	['php', 'javascript', 'jquery', 'html']	94757	23	TRUE	javascript
Can I load the standard windows calculator using web tecnolgies	19160218	['javascript', 'php', 'jquery', 'html', 'windows']	710	-1	FALSE	php
Mobile number and message text checking	20266219	['javascript', 'php', 'jquery', 'html']	2178	0	TRUE	php
Disable page refresh after button click event	22719118	['javascript', 'php', 'jquery', 'html']	6930	0	FALSE	php
What is the purpose of the html form tag	31066693	['javascript', 'php', 'jquery', 'html', 'forms']	15998	86	TRUE	php

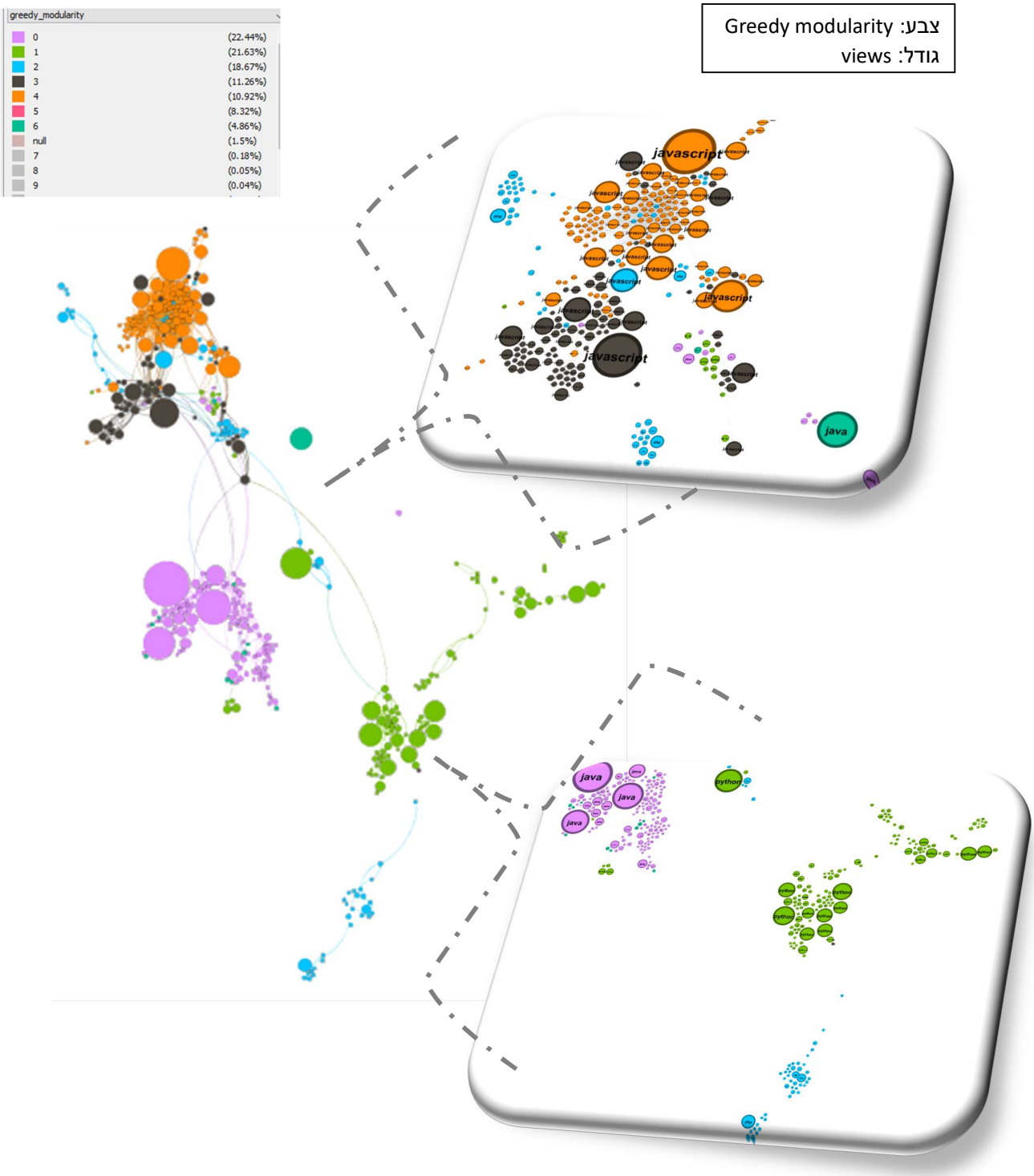
ניתוח *Closeness centrality*.

בניתוח זה ניתן לראות כי שאלות אשר חוצות בין שפות וכלים הן המשמעותיות. למשל בשאלות מבוססות web שבהם יש שליחה של נתונים מכלי כלשהו לPHP. לא מפתיע מאוד, בתחום ה-web יש מספר רב של כלים שהבהם מתעסקים בבת אחת ולכן הגיוני ששאלות שנשאלות הן בעיקר על האינטגרציה ביניהם. מכיוון שאופן ההגדרה של החשיבות של שאלות אצלנו מוגדר באמצעות מדד על מספר טיוגים משותפים נצפה ששאלות כאלו ידורגו גבוה.

title	id	tags	views	score	done	language
Put JSON data into html form input hidden?	8576339	['javascript', 'php', 'html', 'json']	46684	21	TRUE	php
Python 'str' object has no attribute 'read'	19474832	['php', 'python', 'arrays', 'json', 'python-3.x']	22466	0	TRUE	python
Sending a json array with ajax to php	22122486	['javascript', 'php', 'jquery', 'ajax', 'json']	938	-2	FALSE	javascript
jQuery Ajax return multiple data	38380952	['javascript', 'php', 'jquery', 'json', 'ajax']	4104	0	TRUE	javascript
MIT APPInventor error on openning. How to fix it?	41504328	['javascript', 'java', 'php', 'android']	1194	3	FALSE	php

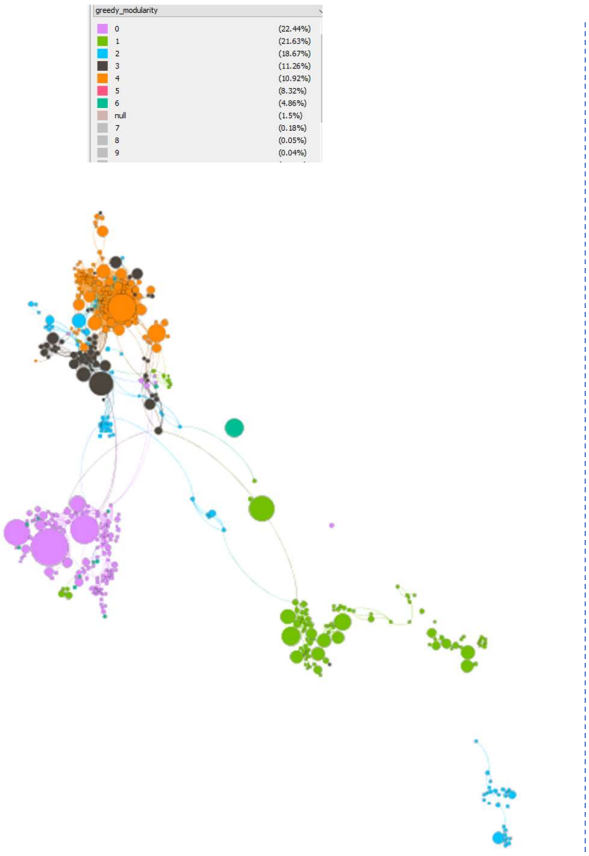
ניתוח ויזואלי ו-Community Detection

ביצענו community detection באמצעות אלגוריתם גרידי מודולריטי אותו יצרנו בזרת networkx.
בגרפים הבאים ניתן לראות בחינה של היכולת לסווג את שפת התכנות עליה נשאלה שאלה באתר stack overflow. תוך כדי למידה של איזה שפות דומות יותר אחת לשניה ואיזה שאלות תופסות חשיבות מרכזית.
את הגרפים יצרנו בתוכנת גפי אשר נלמד עליה בקורס.
לכל גרף צירפנו מקרא צבעים, מה משפיע על גודל הנוד ומה הלייבל.
בגרף הראשון נציג תמונה כללית של הרשת.
ניתן לזהות כי בחלקו העליון של הגרף יש את JavaScript, במרכז את java, החלק הירוק התחתון זה פייטון ולבסוף PHP.

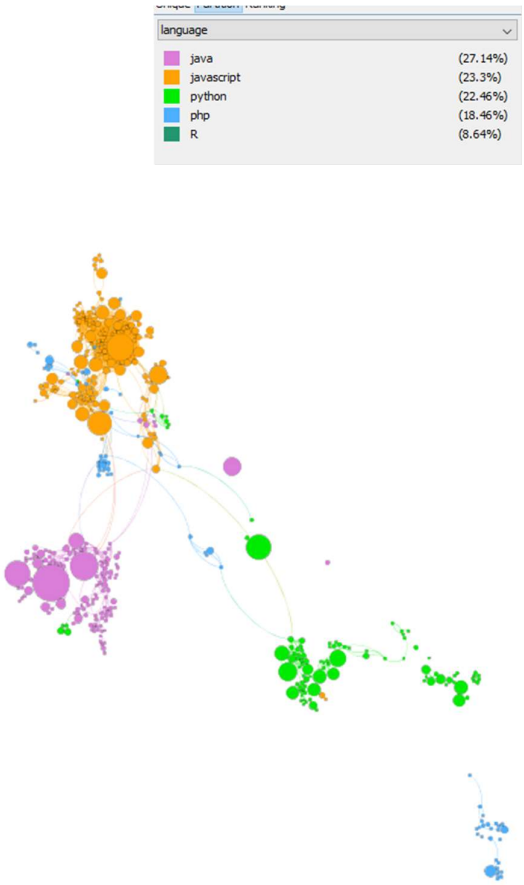


כבר בשלב זה ניתן להתחיל לזהות כי ישנו פוטנציאל להתאמה בין חלוקת הצבע (Greedy modularity) לבין הלייבל (שפה). לכן ביצענו את ההשוואה.

Greedy modularity



Language

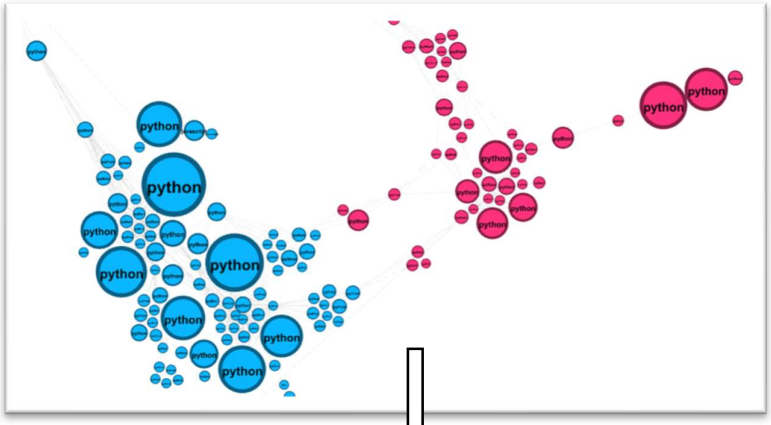


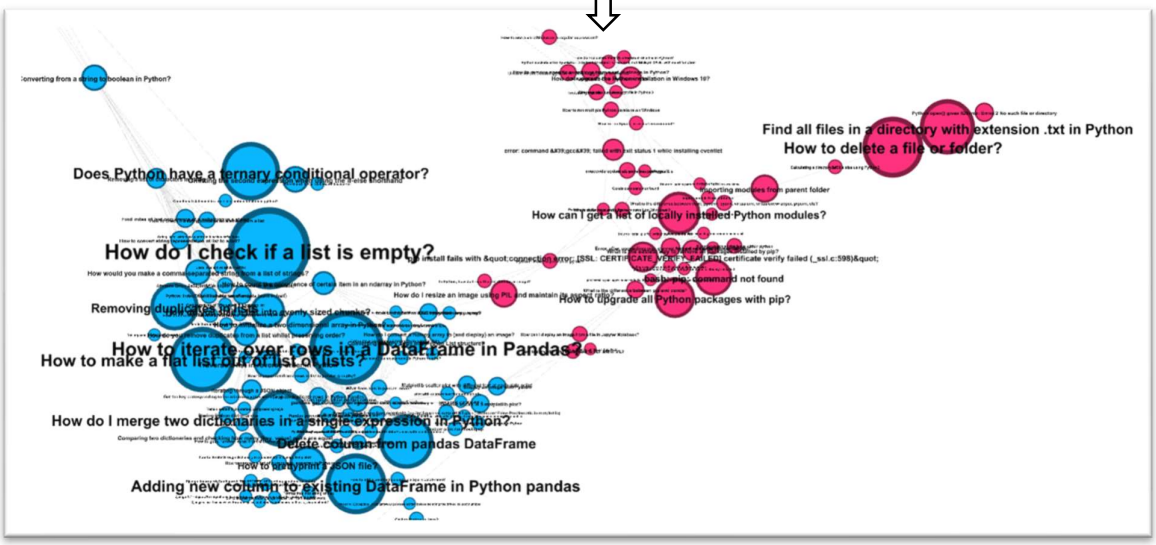
מה שמעניין בניתוח הנ"ל זה עד כמה החלוקה דומה בין הלייבל (השפה) לבין שיטת המודולריטי. בעצם המודולריטי הצליחה לסווג באופן יחסית מרשים את הלייבל של השאלה על סמך התגיות. כלומר ניתן להשתמש בשיטת ה-community detection שלנו על מנת לשייך לייבלים לדטה לא מלא,לסווג נתונים חדשים או להשתמש בקילסטור כפיטשר משמעותי במודל חיזוי. בבדיקת ה-accuracy שביצענו מצאנו כי רמת הדיוק של המודל הוא מעל 83% בסיווג השפה.

בשלב זה נרצה לצלול שלב אחד יותר לעומק. כלומר נחקור את החלוקה על פי המודולריטי של גפי בתוך כל קלסטר (שפה). תוך אבחון המשמעות של החלוקה שנוצרה.

צבע: Modularity
גודל: views
לייבל : שפה ← ואז טייטל

PYTHON:



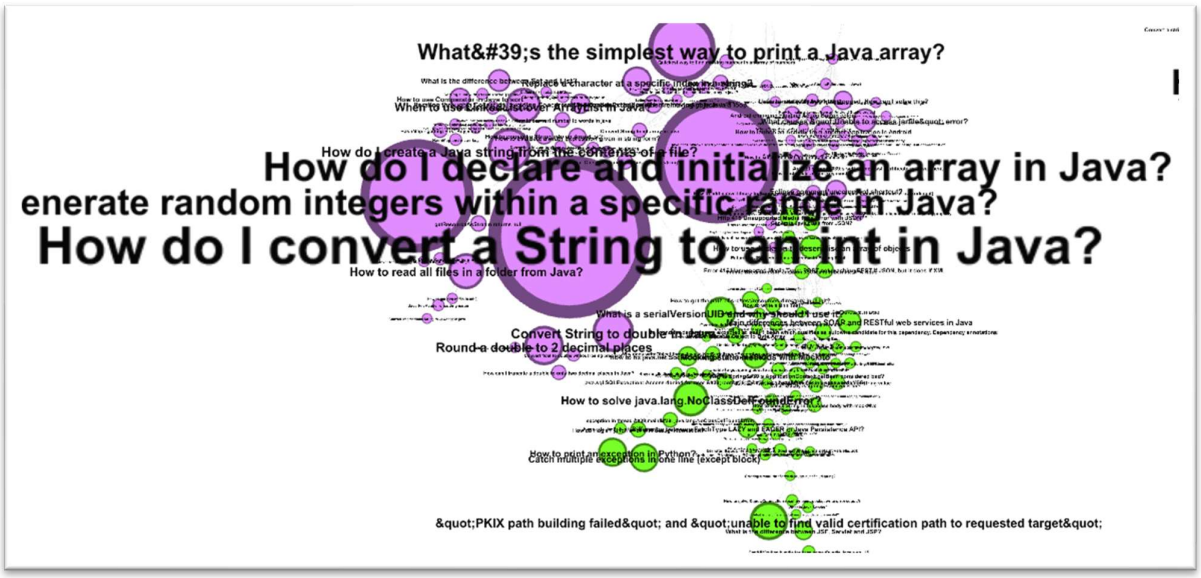


בגרף זה אנו רואים את הנודים אשר נופלים תחת הלייבל של השפה פייטון והצבע מחולק לפי מדד המודולריטי.
נבחן מדוע אנו מקבלים שני צבעים במקום אחד.

בורוד ניתן לראות שאלות מתחום ההתקנות ועבודה עם קבצים בפייטון. בנוסף הקלסטר הזה קרוב יותר לPHP אשר יותר קשור לאופרציות קלאסיות של שפת תכנות מאשר ML.

בכחול אנחנו רואים את עולם הML: Pandas, אופרציות על DF.כלומר החחלוקה הגיונית, אמנם השאלות בשני החלקים קשורים לפייטון, כמו בהרבה שפות, יש לפייטון שימושים רבים וניתן לראות חלוקה מאוד הגיונית.
הצלחנו למצוא סיווג בתוך כל קלסטר לתתי נושאים בעלי משמעות.

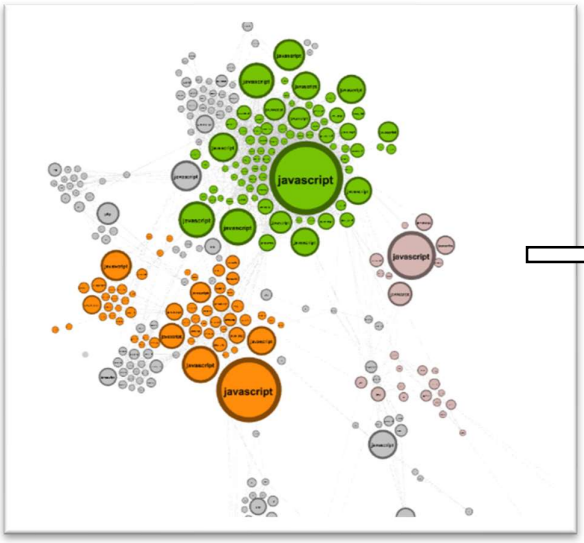
JAVA:

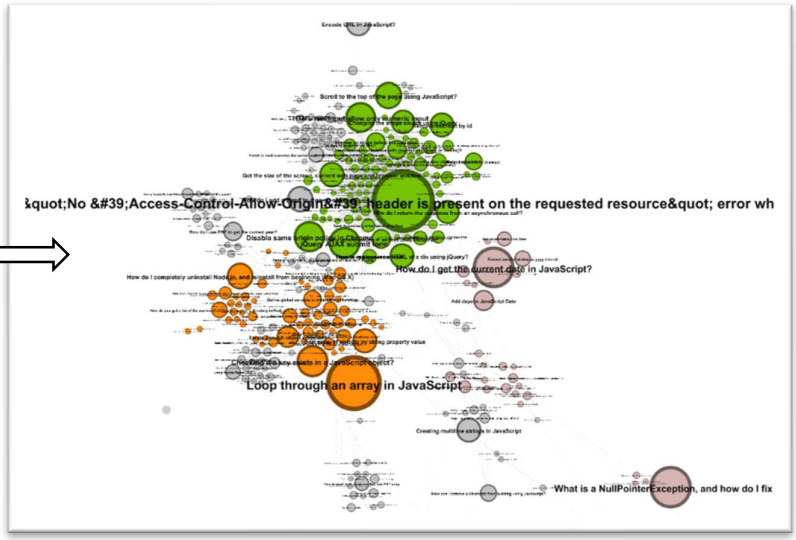


בסגול - אנו רואים שאלות בתחום מניפולציות דטה.

בירוק - יותר תחום הweb והטראבל שוטינג של שגיאות.

JAVASCRIPT:





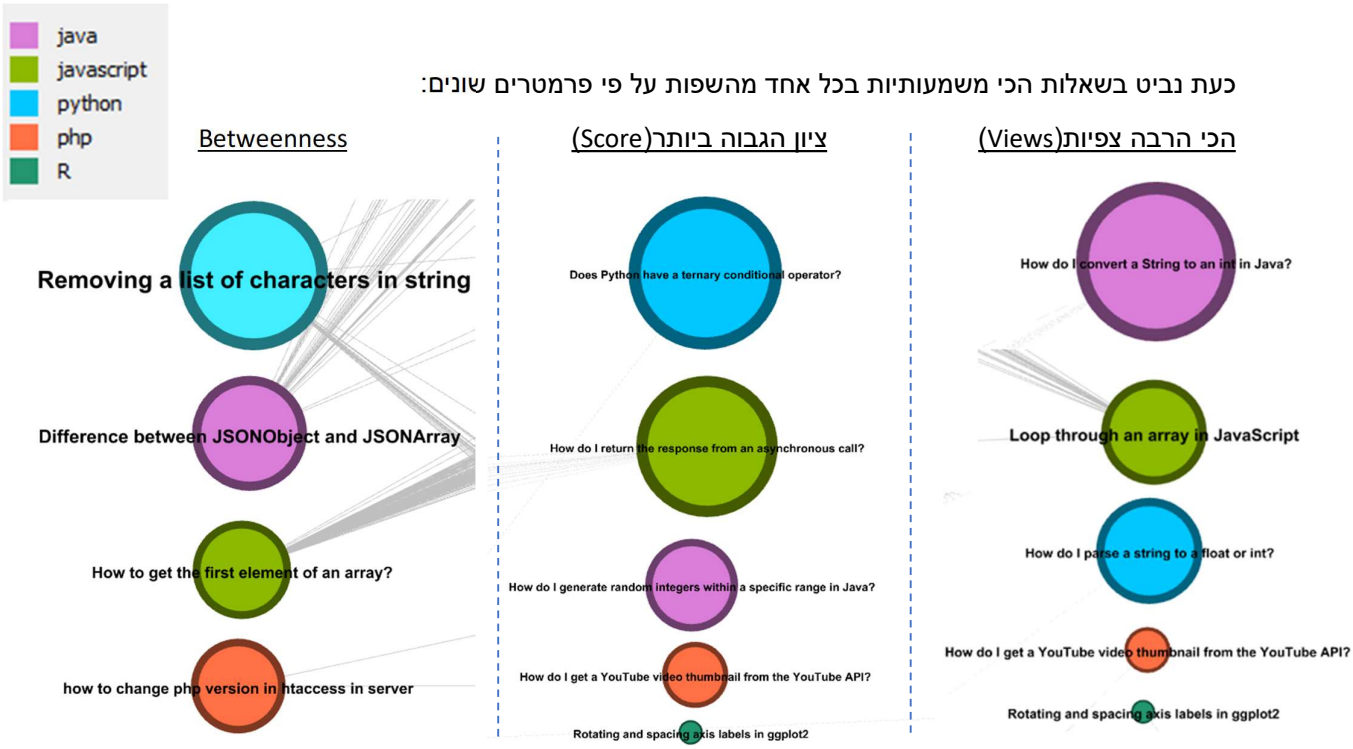
בירוק – ולידציית נתונים

בכתום – פעולות על ערכים

באפור – מידע על משתנים מסוג סטרינג.

בחום - תאריכים

כעת נביט בשאלות הכי משמעותיות בכל אחד מהשפות על פי פרמטרים שונים:



על פי הצפיות ניתן לראות כי השאלות מאוד פופולריות. שאלות שרובנו שאלנו או חיפשנו באתר.
הגיוני ששאלות כאלה הן המשמעותיות ברשת שכן הן שאלות קלאסיות, גנריות ואינן דורשות ידע טכני מעמיק.
לעומת זאת כאשר המדד שלנו הוא ה SCORE דבר שמעיד על חשיבות ואיכות השאלה בעיניי היוזרים אנו רואים שאלות ברמה טכנית יותר גבוהה.
כאשר הBetweenness הוא הפרמט שלנו השאלות הבולטות הן שאלות שקשה להבין באיזה שפה הן נופות או שאלות שהן בעצמן על חיבוריות בין דומיינים.

בהתייחסות לDone, מספר הצפיות וציון של שאלות שלא ענו עליהם נמוך מאוד ביחס לאלה שכן ענו עליהם.
בנוסף אין שאלות בולטות בקבוצה של השאלות שלא ענו עליהם.

הרשת ההפוכה – רשת התגיות

בשביל לקבל זוית נוספת וללמוד עוד דברים על האתר ועל הנתונים שבידינו בנינו רשת נוספת שהיא בעצם רשת הפוכה לרשת הראשונה שהצגנו, אם ברשת הראשונה השאלות היו צמתים והקורלציה בין השאלות חושבה על ידי התגיות המשותפות, כאן הצמתים הן התגיות השונות והקשרים בין שתי תגיות זה שאלות משותפות ששתי התגיות הופיעו בהן.

הציפייה שלנו היא שנקבל רשת עם קלסטרים, כל קלסטר יהיה עם תגיות שרלוונטיות לאחת מחמש השפות שיש לנו בנתונים.

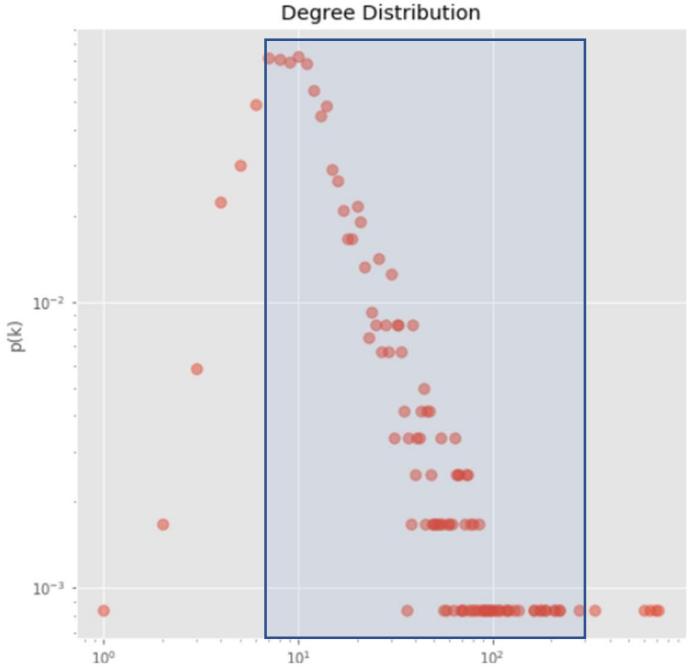
בסה"כ יש לנו 6,145 תגיות, חלק מאוד גדול מהם מופיע בשאלה אחת בלבד כך שאין לתגיות אלה תגיות משותפות, בנוסף רצינו להתמקד בתגיות היותר מרכזיות ולכן החלטנו שנסתכל רק על תגיות שהופיעו בחמש שאלות או יותר, באופן זה פילטרנו את המידע ונשארנו עם 1,200 תגיות בדיוק.

נתוני הרשת

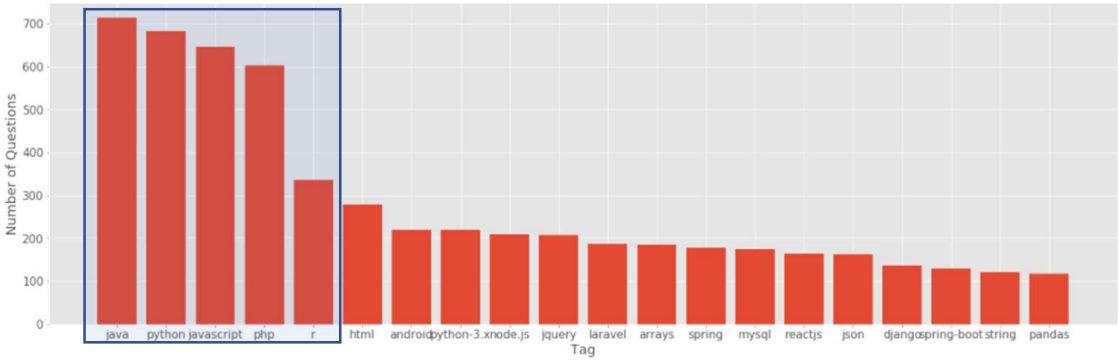
מספר הצמתים הוא N = 1,200, מספר הקשרים הוא L = 13,424.
הדרגה הממוצעת היא <K> = 22.3.
הרשת שקיבלנו היא רשת מחוברת, קוטר הרשת הוא d = 4.
המשמעות היא שאפשר להגיע מכל תגית לכל תגית אחרת, דבר זה בהחלט הגיוני כי ישנן תגיות שלא קשורות לשפה ספציפית אלא רלוונטיות להמון שפות ותחומים, לדוגמא "sql", "knn" או "leetcode". תגיות אלו בעצם מחברות לנו את קלסטרים השונים לרשת אחת גדולה.
d = 4 זה אומר שאנו מקבלים ברשת זו את תכונות העולם הקטן, המשמעות היא שאפשר להגיע מכל תגית לכל תגית אחרת תוך 4 מעברים לכל היותר, זה די הפתיע אותנו כי מגוון התחומים של התגיות רחב מאוד.

התפלגות דרגות

אפשר לראות בגרף למטה שבטווח 10^1-10^2 אנחנו מקבלים סוג של התפלגות power-law.
דבר זה הגיוני בגלל שכאן, בניגוד לרשת השאלות באמת הגיוני שתגיות מסוימות קשורות למספר גדול של תגיות אחרות כי תגיות אלה מופיעות בהמון שאלות.



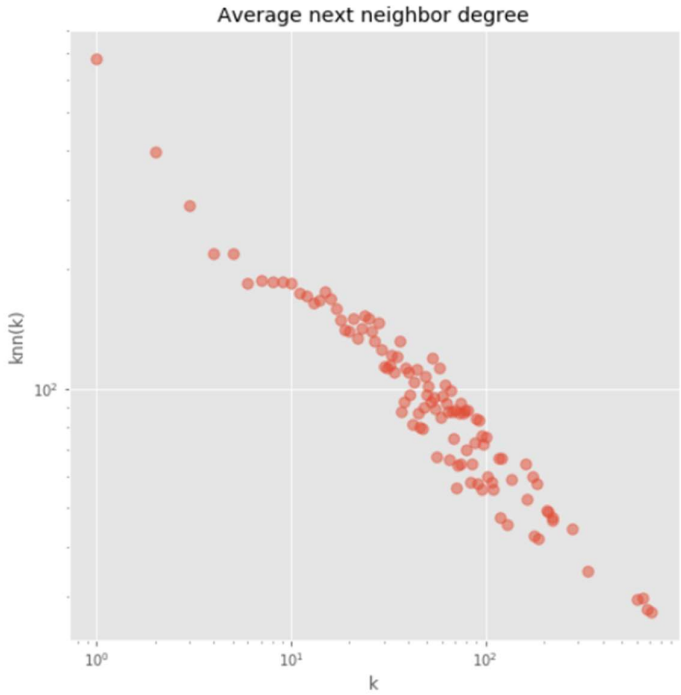
בגרף הבא אפשר לראות את 20 התגיות הכי פופולאריות, ניתן לראות שאכן יש תגיות שמופיעות במספר גדול מאוד של שאלות, חמשת התגיות הכי פופולאריות הן חמשת שפות התכנות שיש לנו בנתונים, דבר זה מסתדר עם ההנחה שלנו שהרשת תכלול חמישה קלסטרים סביב השפות השונות.



Degree Correlation

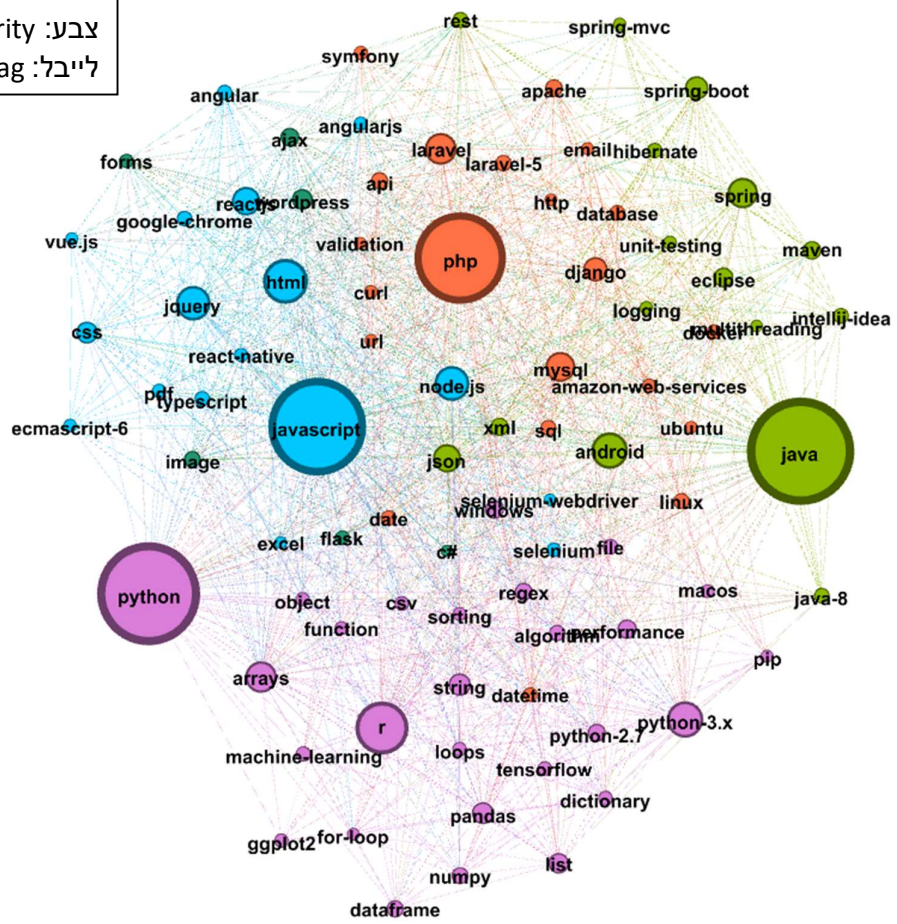
אפשר לראות שכאן, בניגוד לרשת הרגילה, הרשת היא disassortative, מדד ה-assortative של הרשת שלנו הוא r = -0.202104, המשמעות היא שככל שהדרגה של הצומת גדולה, הסיכוי שלשכנים הישירים שלי יש דרגה גבוהה קטן יותר.

פה לא הייתה לנו אינטואיציה לגבי הקורלציה בדרגות, מה שמשפיע על יצירת הקשתות והדרגה של הצומת אלו השאלות המשותפות ואין משהו שמאפיין את השאלות שנראה לנו שיכול להשפיע על הקורלציה.



ניתוח ויזואלי של רשת התגיות

גודל: Degree
צבע: modularity
לייבל: tag



בגרף ניתן לראות את הרשת השניה שלנו.

כאשר כל נוד הוא תגית ומה שמאפיין את הקשר בין התגיות הן השאלות המשותפות.
ראשית כל בולט לעין כי השפות העיקריות הן המשמעויות ביותר.

יש לנו 4 חלקים עיקריים בגרף:

סגול – בעל שני נודים מרכזיים (פייטון וR) עם תגיות נוספות בעלות משמעות מאוד קשורה לשפות הנ"ל. ML, pandas, TensorFlow, pip ועוד. הקשר בין השניים מאוד צפוי שכן לשני הכלים יש יכולות מאוד דומות ואפילו מתחרות אחת בשניה.

ירוק – בעל נוד אחד מרכזי (JAVA). אשר קרוב לתגיות כגון אקליפס (הקומפיילר הפופולרי שמריצים בו ג'אווה, אנדרויד שאת האפליקציות כותבים במקרים רבים בג'אווה).

כתום – בעל נוד אחד מרכזי (PHP). מאפיין את צד השרת ולכן הנודים המקושרים אליו הגיוניים.laravel, validation, database, sql.

כחול – בעל נוד אחד מרכזי (JavaScript). מאפיין את שפת הweb ובעיקר את צד הלקוח. לכן מאוד הגיוני שנודים יחסית גדולים אחרים בקרבתו הם HTML, CSS, jquery.

מסקנות מנתוני רשת התגיות

דבר מאוד מעניין ששמנו לב אליו זה שבהרבה מהמדדים והגרפים קיבלנו תוצאות שונות ואפילו הפוכות מהתוצאות שקיבלנו ברשת הראשונה - רשת השאלות, כמו שאמרנו הרשתות הן הפוכות לחלוטין, ברשת השאלות הצמתים הם השאלות והקשרים בין השאלות מחושבים על ידי התגיות, ברשת התגיות לעומת זאת הצמתים הם התגיות והקשרים בין תגיות זה השאלות המשותפות.

בטבלה הבאה סיכמנו את ההבדלים המרכזיים בין שתי הרשתות:

רשת	**רשת השאלות**	**רשת התגיות**
Connectivity	weekly connected	fully connected
Degree distribution	no scale free	~ scale free
Degree correlation	assortative (r = 0.17302)	disassortative (r = -0.202104)
Average clustering coefficient	0.7452	0.5141

אפשר לראות שבלא מעט מדדים הרשתות עם מבנה הפוך, ניסינו להבין את המשמעות של זה, חשבנו שזה קשור לכך שאחת הרשתות היא בעצם הבעיה הדואלית של הרשת השניה אבל לא הצלחנו למצוא התייחסות לכך במאמרים אקדמיים, ננסה לחקור נושא זה בהמשך לימודינו האקדמיים.

סיכום ומסקנות

במחקר זה ניסינו להבין את הקשרים בין שאלות באתר stack overflow מתוך רצון למצוא דמיון בין שאלות ליעל את האתר למצוא קשרים מעניינים בתוך שפות וביניהם וכמובן לנסות לסווג את השפה שבה השאלה נשאלת.

לצורך כך לקחנו מידע משנת 2020 על 5 שפות עיקריות (PHP, Python, Java, JavaScript, R) עם 75K שאלות ייחודיות כאשר לכל שאלה 5-7 תגיות.
בנינו שתי רשתות, רשת אחת, רשת השאלות, שבה הצמתים הם השאלות וקשר בין שאלה לשאלה התבססה על חישוב סימילריטי של Jaccard Index אשר פעל על מנת להבליט קשר בין שאלות עם כמה שיותר תגיות משותפות. רשת שניה, רשת התגיות, בעצם רשת הפוכה. ברשת התגיות כל צומת היא תגית ייחודית והקשרים בין הצמתים נבנים על בסיס שאלות משותפות (לפחות 5).

את הניתוח ביצענו לאחר בדיקת סף יסודית אשר מסקנתה הייתה לבחור סף של 0.6 קורלציה כסף ליצירת קשר בין השאלות.
מהניתוח שביצענו על רשת השאלות לא קיבלנו התפלגות scale free כפי שציפינו.
בבדיקת degree correlation (KNN) קיבלנו גרף אסורטטיבי עם r=0.17 כלומר צמתים בעלי דרגה גבוהה לרוב מתחברים עם צמתים בעלי דרגה גבוהה.

מהניתוח הויזואלי למדנו תחילה שהשאלות הבולטות (על פי מספר הצפיות) הן שאלות קלאסיות, כללית ולא דורשות ידע טכני מעמיק. למדנו כי ישנה חלוקה יפה ל-communities באמצעות מודל הגרידי מודולריטי שבנינו ב-networkx. החלוקה התגלתה כסיווג עם דיוק של מעל 83% אל מול הלייבל שלנו (השפה שעליה נשאלה השאלה), דבר השווה להדגיש משום שהאתגר ב-Kaggle הוא הסיווג הנ"ל והמודל הצליח לענות על האתגר.
בשלב זה העמקנו וחיפשנו האם ניתן למצוא חלוקה בתוך כל שפה ל communities בעלי משמעות.
החלוקה הוכיחה עצמה כמעניינת מאוד, שכן מצאנו בתוך כל קלסטר תתי נושאים בעלי משמעות. למשל פייטון שישנה חלוקה מאוד ברורה לML ולהתקנות.

ברשת השניה, רשת התגיות, קיבלנו התפלגות power law בטווח ערכים מסוים. הגיוני משום שעבור התגיות כן נצפה למצוא האבים – התגיות שהן בעצם כלליות יותר כגון השפות שלנו.
גילינו כי ברשת שהיא בעצם ההיפוך של הרשת הראשונה המדדים גם כן הפוכים. יתכן כי רשת התגיות היא הדואלית של רשת השאלות.
בניתוח הויזואלי מצאנו כי הצמטים הבולטים הן השפות. אשר מהוות קלסטרים של פרונטאנד, בקאנד, סטטיסטיקה/ML וJAVA.

<u>צעדים להמשך</u>

על מנת לאפשר סיווג בעל רמת דיוק של שאלות זהות נמליץ להשתמש בפיטשרים נוספים כגון ניתוח טקסט, שכן, רוב המידע יושב בתצורה הזו.
היינו מעוניינים להוסיף זמן כאלמנט על מנת לראות אם יש טרנדים, תוך הוספת שפות נוספות על מנת להפיק מסקנות נוספות.
הוספת מידע מעולם המשרות/קורסים טכניים וחיבור לדטה שלנו יכול להניב תוצרים מעניינים, למשל: מה מעניין אנשים והאם הם מגישים לתפקידים שעליהם הם שואלים?מהם התחומים החמים בשוק? האם אנשים שואלים יותר/פחות ככל שההכשרה שלהם מתקדמת ועוד.