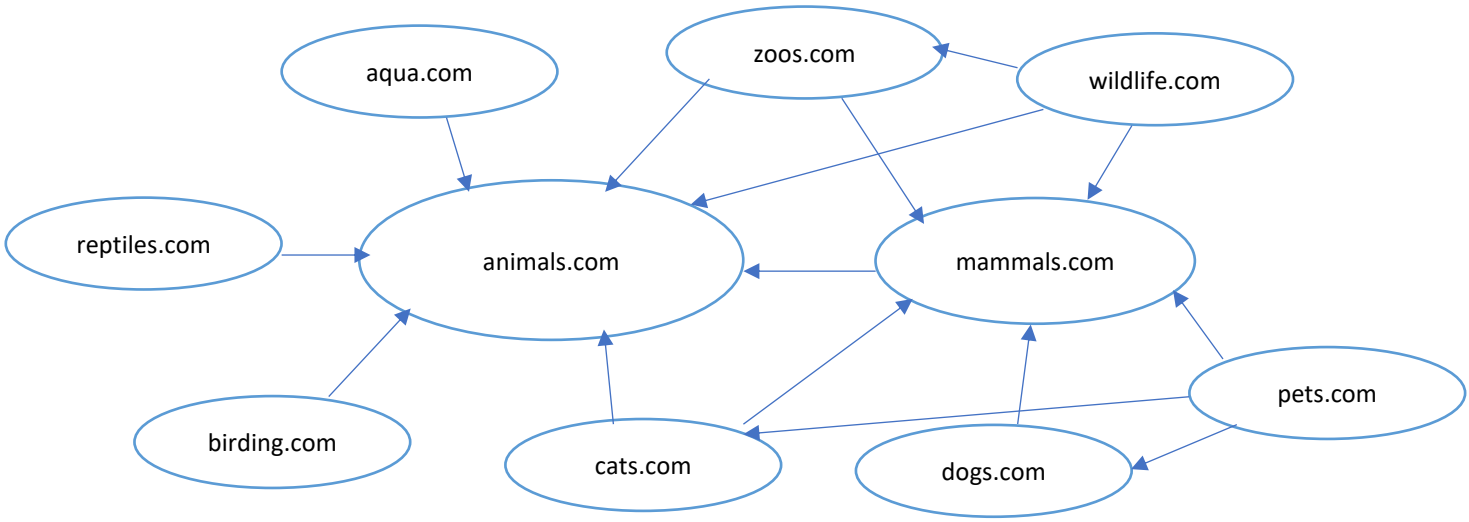
Osher Elhadad – אושר אלחדד - 318969748 Gili Gutfeld – גילי גוטפלד - 209284512

ניהול נתוני עתק- תרגיל 2

שאלה 1

בנינו מיני רשת עם 10 אתרים וזוהי הדיאגרמה שמייצגת את הלינקים ביניהם:



שאלה 3

קיבלנו כתוצאה מהרצת הפונקציה invertedIndex(myData(),mySearchString() את הפלט הבא:

{'cat': [['cats.com', 0.13264666955734586], ['pets.com', 0.07958800173440753], ['mammals.com', 0.0497425010840047], ['animals.com', 0.0497425010840047]], 'mammal': [['mammals.com', 0.125]], 'animals': [['pets.com', 0.13979400086720375], ['zoos.com', 0.13979400086720375]], 'cute': [['dogs.com', 0.10457574905606754], ['pets.com', 0.10457574905606754], ['animals.com', 0.0653598431600422]], 'big': [['cats.com', 0.17429291509344585], ['dogs.com', 0.10457574905606754], ['animals.com', 0.0653598431600422]]}

שאלה 4

קיבלנו כתוצאה מהרצת הפונקציה pageRankSimulation(myData(),100000,0.8) את הפלט הבא:

[['animals.com', 0.369191], ['mammals.com', 0.196821], ['dogs.com', 0.063671], ['zoos.com', 0.06320100000000001], ['cats.com', 0.062171000000000004], ['wildlife.com', 0.049621000000000005], ['birding.com', 0.049541], ['aqua.com', 0.048671000000000006], ['pets.com', 0.048571], ['reptiles.com', 0.048551000000000004]]

האלגוריתם PageRank לוקח בחשבון את מספר ואיכות הקישורים המפנים לאתר וגם את דירוגם של האתרים שמצביעים אליו כדי לקבוע את חשיבותו של האתר.
בדוגמה שלנו, האתר עם הדירוג הגבוה ביותר הוא animals.com שתואם את האינטואיציה שלנו כי כמעט לכל שאר האתרים יש קישורים אליו. לאחר מכן, יש לנו את mammals.com שעשוי להיות האתר השני בחשיבותו כי כמעט כל החיות הן יונקים ולמעשה יש מהן קישורים אליו. בנוסף, ל-mammals.com יש קישור ל-animals.com אך לא להפך, מה שמחדד לנו את חשיבותו של animals.com לעומת "סגנו", mammals.com. לאתרים האחרים יש דירוג נמוך משמעותית כפי שצפינו ולכן הפלט של הפונקציה אכן אינטואיטיבי לנו.

שאלה 5

הפונקציה שבחרנו היא שורש סכום הריבועים (נציין כי $x, y \in \mathbb{R}$ מכיוון שאלו ערכים אפשריים עבור $tfIdf$ ו$PageRank$):

$$f(x,y) = \sqrt{x^2 + y^2}$$

נראה כעת כי הפונקציה מונוטונית עולה עבור שני המשתנים x, y , כלומר יהיו $x, x_1, y, y_1 \in \mathbb{R}$ ומתקיים $x_1 \geq x \geq 0 \wedge y_1 \geq y \geq 0$ נראה כי $f(x_1, y_1) \geq f(x,y)$.

$$x_1 \geq x \geq 0 \wedge y_1 \geq y \geq 0 \Longrightarrow$$

$$x_1^2 \geq x^2 \geq 0 \wedge y_1^2 \geq y^2 \geq 0 \Longrightarrow$$

$$x_1^2 + y_1^2 \geq x^2 + y^2 \Longrightarrow$$

$$\sqrt{x^2 + y^2} \geq \sqrt{x^2 + y^2} \Longrightarrow$$

$$f(x_1, y_1) \geq f(x,y)$$

מונוטונית כדרוש.

בחרנו בפונקציה זו מכיוון שהיא מקבלת תוצאות שהן יותר קטנות מאשר רק סכום של הריבועים למשל, מכיוון שהיא נותנת השפעה שווה לשני המשתנים x, y, ואכן ערכיהם של הPageRank וtfIdf יצאו די קרובים (ערכים קטנים) וכמובן כי הינה מונוטונית.

שאלה 6

קיבלנו כתוצאה מהרצת הפונקציה top1(invertedIndex(myData(),mySearchString()),pageRankSimulation(myData(),100000,0.8)) את הפלט הבא:

```
Sorted access to animals.com at the PageRank index
Random access to animals.com at the invertedIndex cat
Random access to animals.com at the invertedIndex mammal
Random access to animals.com at the invertedIndex animals
Random access to animals.com at the invertedIndex cute
Random access to animals.com at the invertedIndex big
Sorted access to cats.com at the invertedIndex cat
Random access to cats.com at the PageRank index
Random access to cats.com at the invertedIndex mammal
Random access to cats.com at the invertedIndex animals
Random access to cats.com at the invertedIndex cute
Random access to cats.com at the invertedIndex big
Sorted access to mammals.com at the invertedIndex mammal
Random access to mammals.com at the PageRank index
Random access to mammals.com at the invertedIndex cat
Random access to mammals.com at the invertedIndex animals
```

Random access to mammals.com at the invertedIndex cute
Random access to mammals.com at the invertedIndex big
Sorted access to pets.com at the invertedIndex animals
Random access to pets.com at the PageRank index
Random access to pets.com at the invertedIndex cat
Random access to pets.com at the invertedIndex mammal
Random access to pets.com at the invertedIndex cute
Random access to pets.com at the invertedIndex big
Sorted access to dogs.com at the invertedIndex cute
Random access to dogs.com at the PageRank index
Random access to dogs.com at the invertedIndex cat
Random access to dogs.com at the invertedIndex mammal
Random access to dogs.com at the invertedIndex animals
Random access to dogs.com at the invertedIndex big
Sorted access to cats.com at the invertedIndex big
Sorted access to mammals.com at the PageRank index
Sorted access to pets.com at the invertedIndex cat
Sorted access to animals.com at the invertedIndex mammal
Sorted access to zoos.com at the invertedIndex animals
Random access to zoos.com at the PageRank index
Random access to zoos.com at the invertedIndex cat
Random access to zoos.com at the invertedIndex mammal
Random access to zoos.com at the invertedIndex cute
Random access to zoos.com at the invertedIndex big
Sorted access to pets.com at the invertedIndex cute
Sorted access to dogs.com at the invertedIndex big
Sorted access to dogs.com at the PageRank index
Sorted access to mammals.com at the invertedIndex cat
Top 1 page - animals.com with score- 0.4109362426991182

אכן האלגוריתם החזיר את האתר הכי חשוב ורלוונטי (animals.com) כפי שניתן לראות מההגדרות של האתרים- גם רוב האתרים מכילים קישור אל אתר זה ופרט גם האתר mammals.com שאליו גם כמעט כל האתרים מכילים קישור ומכיוון שאתר זה מכיל קישור לאתר animals.com לכן זה עוד יותר מעלה את חשיבותו ולכן מבחינת הPageRank מכיוון שאלגוריתם PageRank לוקח בחשבון את מספר ואיכות הקישורים המפנים לאתר וגם את דירוגם של האתרים שמצביעים אליו כדי לקבוע את חשיבותו של האתר. כמו כן ניתן לראות שמבחינת tfIdf אזי אמנם לanimals.com אין דירוג ראשון באינדקסים השונים אך כן הוא מופיע ברובם (כלומר רלוונטי למחרוזת החיפוש) לעומת רוב האתרים האחרים שברוב האינדקסים אין להם כלל דירוג tfIdf (כלומר לפי האלגוריתם שלנו יש להם דירוג tfIdf 0), ולכן בשכלול 2 הדירוגים כפי שציפינו אכן animals.com הינו במקום הראשון.
בשאלה זו בחרתי לממש את האלגוריתם TA מכיוון שהוכחנו בביתה שהוא מאוד יעיל.
למעשה ביצענו באלגוריתם זה סה"כ 44 גישות- כאשר 30 רנדומיות ו14 ממוינות.
באופן נאיבי היינו צריכים לעבור על כל 10 האתרים עבור כל אחד משש העמודות (עמודה אחת לPageRank ו5 לtfIdf עם invertedIndex) כלומר היינו צריכים לבצע 60 גישות ולכן למעשה חסכנו באופן זה 16 גישות מתוך ה60 שהיינו אמורים לבצע שזה יוצא 26.6667% חסכון בגישות שלנו. זאת מכיוון שאנו משתמשים בthreshold ובגישות ממוינות ורנדומיות באלגוריתם מה שמאפשר לנו ברגע שישנו score מסוים לpage אשר ערכו גדול מהthreshold שמשתנה בכל גישה ממוינת וקטן אזי אנחנו יכולים להחזיר את הtop1 ולכן אנו לא עוברים על כל 60 התאים שהיינו אמורים לעבור בהם.