אחזור מידע – פרויקט להגשה – אביב התשפ"ד

יש להגיש למוודל עד לתאריך 25.8.24 בשעה 23:59

*מבוסס על תרגילי הבית בקורס*



מגישים

אלעד פישר - 318882800

חן חנצין - 208636761

גלעד סגל - 314865841

ניצן ממן – 205798515

קישור לגיט עם כל מטלות הקורס:

<https://github.com/gilseg10/Information-recovery-class>

תרגיל בית 1

תרגיל בית 2

תרגיל בית 3

1. מהו תחום העיסוק המרכזי של האתר? מהו המידע הזמין למשתמשי האתר? ענו בפסקה אחת. צרפו את הקישור לאתר.

**תחום העיסוק המרכזי** של אתר Steam הוא מכירת והפצת משחקי מחשב ותוכן דיגיטלי נוסף, כגון תוכנות, סרטים, ותוכן להורדה (DLC). האתר מאפשר למשתמשים לרכוש, להוריד, ולהתקין משחקים ותוכנות שונות, כמו גם לנהל את הספרייה הדיגיטלית שלהם, לקרוא ולכתוב ביקורות, להשתתף בקהילות משחקים, ולצפות בטריילרים ובשידורי משחקים חיים. Steam מציע גם אפשרויות לרכישות בתוך המשחק ותמיכה במודינג (יצירת תוכן על ידי המשתמשים).

1. רשמו שלוש שאילתות מעניינות שהייתם רוצים לקבל עליהן תשובה באתר, והאתר אינו **עונה עליהן** כעת. עבור כל שאילתא כזו, ציינו מהם פרטי המידע הנדרשים לצורך מענה על השאילתא.

שאילתה 1: מהו ממוצע זמן המשחק ביחס לכלל השחקנים שרשומים אליו.  
הסבר: מתוך ה-API, אנו משתמשים ב EndPoint הבא: <http://api.steampowered.com/IPlayerService/GetOwnedGames/v0001/>  
שמאפשר החזרה של רשימת משחקים שנרכשו (בתשלום/חינם) עבור כל משתמש, וגם זמן משחק כולל של כל משתמש על כל אחד מהמשחקים.   
התגיות הרלוונטיות בהקשר שאילתה זו הן:  
**playtime\_forever** - זמן המשחק הכולל של משחק מסוים (בדקות).  
השאילתה מחזירה: רשימה של משחקים, כאשר עבור כל משחק: מתעדים את **כמות המשתמשים** וגם כן את **הזמן המשחק של כל משתמש**, ובנוסף גם **הזמן הממוצע ששיחקו במשחק**.

שאילתה 2: חילוץ המילים הבולטות ביותר בהישגים של משחק. **(Achievement או Badges).**הסבר: לכל משחק שרשום בSteam, יש שלל הישגיים (Badges) שמשתמש יכול להשיג לאורך זמן המשחק שלו. לכל הישג שכזה יש תיאור ושם שמקושר למשחק, וגם אחוז גלובלי שמציג כמה מכלל המשתמשים בעולם שרכשו את המשחק השיגו את ההישג הזה.  
מידע זה ניתן לחלץ מה-API דרך הEndPoint הבא:   
<http://api.steampowered.com/ISteamUserStats/GetGlobalAchievementPercentagesForApp/v0002>   
התגיות הרלוונטיות בהקשר לשאילתה זו הן:  
**gameid** – אמצעי זיהוי ייחודי עבור כל משחק.  
השאילתה מחזירה: רשימה של כלל ההישגים (שמות ואחוז הישג גלובלי) למשחק מסוים. דרכו, אנו נחלץ את המילים הבולטות ביותר מכלל שמות ההישגים הקיימים למשחק.

שאילתה 3: חילוץ המילים הנפוצות ביותר בתיאור של כל משחק.  
הסבר: לכל משחק שמופץ ב Steam קיים עמוד משלו ובתוך העמוד ישנו תיאור המשחק תחת הכותרת "ABOUT THIS GAME". אנו נחפש את המילים הנפוצות ביותר בכל תיאור משחק אשר לרשותנו דרך שאיבת עמוד ה-HTML של המשחק וחילוץ הפסקאות המתארות את המשחק. ניתן לחלץ את עמוד ה-HTML של כל משחק דרך נקודת הגישה הבאה: https://store.steampowered.com/app/{appid}  
התגיות הרלוונטיות בהקשר לשאילתה הן:  
appid – אמצעי הזיהוי של משחק מסוים לחקירה.  
game\_area\_description – ה-id של ה-div המכיל את הפסקאות של תיאור המשחק.  
השאילתה מחזירה: רשימה של המילים הנפוצות ביותר בתיאור משחקים ממוינות בסדר לא עולה.

1. בנו זחלן המחזיר את התוצאות לשאילתות שהגדרתם. מומלץ (אך לא חובה) לבנות את הזחלן בפייתון. הציגו את קוד הזחלן.
2. רשמו טכנולוגיות שונות מעניינות שהשתמשתם בהם בפרויקט.
   1. נעזרנו ב Steam Web Api – אשר דרכו השתמשנו בשאילתות מוכרות לצורך חילוץ מידע על משחקים רבים: איזה הישגים (Badges) יש למשחק, זמן משחק ממוצע ודומה.
   2. השתמשנו גם BeautifulSoup – ספריית פייתון שנועדה לחלץ מידע מאתרי HTML. נעזרנו בה על מנת לחלץ טקסט של התיאור (Description) לכל משחק מדף האינטרנט שלו וגם לצורך חילוץ רשימת משחקים לפי קטגוריות הקיימות באתר Steam.
3. כמה זמן רצו השאילתות שלכם?במה זה תלוי? האם לדעתכם ניתן לשפר זמן זה?

השאילתות רצות בטווח של 10-30 דקות כתלות בכמות המשחקים (דפים שחיפשנו) וכמות המשתמשים שבהם השתמשנו למאגר. לצורך שיפור אפשר לצמצם את כמות המשתמשים על ידי סינון מקדים של משתמשים בעלי זמן משחק נמוך או כמות משחקים נמוכה באופן כללי.

1. בנו inverted index ל- 15 המילים הנפוצות שחזרו (מספיק להתיחס ל -20 הדפים הראשונים שחזרו)
2. בחרו את אחת השאילתות, וחשבו tfIdf של המושגים בשאילתא.

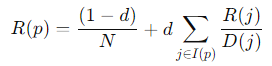
באקסל

1. האם בדפים שהוחזרו קיימים hubs? Authorities? נמקו.

עבור השאילתה למציאת משחקים בהתבסס קטגורית משחק, אנו סיפקנו לזחלן את הHubs (הקטגוריות) – ודרך כך, קיבלנו את הAuthorities (המשחקים) ששייכים לאותו Hub. פירוט נוסף בסעיף הבא.

1. בחרו 10 דפים שהזחלן החזיר, ואשר יש ביניהם קישורים .חשבו pageRank לכל דף, מומלץ להעזר בסקריפט מוכן. הציגו את החישובים ואת הדירוג הסופי.

במסגרת מטלה זו הרצנו שאילתא נוספת שבה הזחלן מקבל קטגוריות של משחקים אליהן התייחסנו כ-hubs ועל ידי שימוש בהם הזחלן החזיר לנו את המשחקים הכי פופולריים לאותם קטגוריות, על ידי חילוץ קישור מתאים (שימוש בתגית <a>). משחקים אלו הגדרנו כ-Authorities והזחלן שמר לנו אותם בטבלה בקובץ csv (מצורף בהגשה). מתוכם, בחרנו 10 דפים: 4 דפי קטגוריות (hubs) ו-6 דפי משחק (Authorities) וביצענו חישוב pageRank בהתבסס הנוסחא שלמדנו בכיתה:



בהקשר שלנו:

הזנחנו את 'd'.

N שווה ל-10 דפים.

I(p) מייצג סט של קטגוריות שמצביעות ל-p (משחק).

D(j) מייצג סט של משחקים שקטגוריה j מצביעה עליהם.

שלב ראשון:

באתחול כל דף מקבל דירוג של 1/10.

שלב שני:

הבנת אופן חישוב הדירוג של כל דף בהתבסס על הדפים שמצביעים אליו:

* מכיוון והדפי קטגוריות מצביעים רק לדפים אחרים (דפי משחק) ואף אחד לא מצביע אליהם אז הדירוג שלהם יישאר **1/10.**
* דפי המשחק מקבלים הפנייה מכמה קטגוריות שונות ולכן נדרש היה לחשב כל אחד בצורה עצמאית.

שלב שלישי:

חישוב כל דירוג של כל משחק:

**Game Name: Counter-Strike 2**

I(Counter−Strike 2)={ Action, Strategy }

D(Action)=4 and D(Strategy)=3.

R(Counter-Strike 2) = 0.1 + 0.1 / 4 + 0.1 / 3 = 0.1583

**Game Name: ELDEN RING**

I(ELDEN RING)={ Action, RPG }

D(Action)=4 and D(RPG)=4.

R(ELDEN RING) = 0.1 + 0.1 / 4 + 0.1 / 4 = 0.15

**Game Name: The Elder Scrolls® Online**

I(The Elder Scrolls® Online)={ Adventure, RPG, Action }

D(Adventure)=3 and D(RPG)=4 D(Action)=4.

R(The Elder Scrolls® Online) = 0.1 + 0.1 / 3 + 0.1 / 4 + 0.1 / 4 = 0.1833

**Game Name: Baldur's Gate 3**

I(Baldur's Gate 3)={ Adventure, RPG, Strategy }

D(Adventure)=3 and D(RPG)=4 D(Strategy)=3.

R(Baldur's Gate 3) = 0.1 + 0.1 / 3 + 0.1 / 4 + 0.1 / 3 = 0.1916

**Game Name: The First Descendant**

I(Baldur's Gate 3)={ Adventure, Action}

D(Adventure)=3 and D(Action)=4.

R(Baldur's Gate 3) = 0.1 + 0.1 / 3 + 0.1 / 4 = 0.1583

**Game Name: Once Human**

I(Once Human)={ Adventure, Action, RPG, Strategy}

D(Adventure)=3 and D(Action)=4 D(RPG)=4 D(Strategy)=3.

R(Once Human) = 0.1 + 0.1 / 3 + 0.1 / 4 + 0.1 / 4 + 0.1 / 3 = 0.2166

1. הראו לשני משתמשים שונים את הדירוג מהסעיף הקודם, בקשו מהם לסמן relevance feedback .

חשבו precision ו - recall.האם ניתן להציע שאילתא מותאמת על מנת לשפר תוצאות?

הצגנו את השאילתה לשני חברים לתואר אשר משחקים משחקי מחשב ובעלי חשבון Steam. מתוך השאילתה הצגנו 10 משחקים (מסמכים) שעלו מהשאילתה, ושאלנו אותם כמה הם רלוונטיים עבורם. כלומר, בהתבסס המשחקים שהם אוהבים/רוכשים באתר של Steam. שני החברים בחרו איזה משחקים רלוונטיים עבורם. להלן רשימת המסמכים אשר הצגנו להם:

|  |  |
| --- | --- |
| AppId | Game name |
| 730 | Counter-Strike 2 |
| 1245620 | ELDEN RING |
| 306130 | The Elder Scrolls® Online |
| 1086940 | Baldur's Gate 3 |
| 2074920 | The First Descendant |
| 2139460 | Once Human |
| 230410 | Warframe |
| 1222670 | Sims 4 |
| 271590 | Grand Theft Auto V |
| 1085660 | Destiny 2 |

משתמש ראשון: המשתמש הראשון הביע העדפה חזקה למשחקי הרפתקאות ושיתף שהוא נהנה במיוחד ממשחקים כמו "The Elder Scrolls® Online" ו-"Baldur's Gate 3". הוא הרגיש שמשחקים אלה מדורגים גבוה בקטגוריית הרפתקאות, מכיוון שהם מספקים את החוויות האופייניות לז'אנר. עם זאת, המשתמש הצביע על חוסר משיכה ספציפית למשחק "Once Human". הוא ציין שמשחק זה לא נחשב בעיניו כמשחק הרפתקאות מסורתי. הוא המליץ שמשחק זה צריך להיות בעל דירוג נמוך יותר או אפילו יוסר מקטגוריית ההרפתקאות. המשתמש הדגיש העדפה לראות משחקי הרפתקאות קלאסיים יותר, כגון "Grand Theft Auto V" ו-"The First Descendant". **סך הכל, המשתמש הראשון סימן 7 משחקים רלוונטיים מתוך הרשימה.**

משתמש שני: המשתמש השני החל לשחק לאחרונה ב-"Baldur's Gate 3" והוא מאוד התרשם מהמיקום שלו בקטגוריות RPG ואסטרטגיה. הוא הדגיש את העומק האסטרטגי של המשחק, שמתיישר היטב עם העדפות המשחק שלו. בנוסף, הוא הביע עניין רב במשחקי Online, מה שהופך משחקים כמו "The Elder Scrolls® Online" ו-"Destiny 2" לרלוונטיים במיוחד לטעמו. המשתמש היה מרוצה גם מהדירוג של "The Elder Scrolls® Online". המשתמש גם ציין שהמשחק "Once Human" אינו מועדף עליו במיוחד ולכן היה מדרג נמוך יותר לעומת משחקים אחרים. **סך הכל, המשתמש השני סימן 6 משחקים רלוונטיים מתוך הרשימה.**

חישוב Precision ו-Recall: עבור Precision החישוב מתבצע על פי כמות המשחקים הרלוונטיים חלקי ה-10 מסמכים שהחזרנו. לעומת זאת, ה-Recall אנו נחשב על פי כמות המסמכים הרלוונטיים חלקי הערכה עבור כמות המסמכים שנמצאים מהאתר של Steam. בהקשר השאילתא שלנו, אנחנו החזרנו משחקים פופולריים מתוך קטגוריות מסוימות. באתר קיימים בסביבות ה-100,000 משחקים שונים וכ-60 קטגוריות (ז'אנרים) שונים של משחקים. לכן עבור כל קטגוריה אפשר להניח כי קיימים באזור 1,700 משחקים.

User 1:

Precision = 7/10 = 0.7

Recall = 7/1700 = 0.00411

User 2:

Precision = 6/10 = 0.6

Recall = 6/1700 = 0.0035

1. הציגו את כל התוצרים של סעיפים א-ו בדף HTML יחיד. הדף יכול להיות דינמי או סטטי (כלומר מציג צילומי מסך של התשובות לשאלות). העלו את הדף לgit pages. צרפו קישור לדף שהעליתם.
2. בתאריך 12.8 תציגו את הדף מסעיף 11, הכולל את סעיפים 1-10. תקבלו משוב מחבריכם. ענו בטבלה:

|  |  |
| --- | --- |
| איזה שינוי הוצע | האם לדעתכם יש מקום לשיפור זה? אם כן, הסבירו כיצד אפשר לשפר. אם לא, נמקו. |
| "אולי לחלץ את המילים הכי נפוצות ב-comments בעזרת מודל של Word Embedding במקום ספירת מילים הכי נפוצות." | אפשרי לחלוטין להשתמש במודלים או אלגוריתמים אחרים למציאת שכיחות מילים בטקסט. |
| "חישוב כמות משחקים פיר כל קטגוריה אולי.  להוסיף כמות זמן, AI ימליץ לפי טווח זמנים." | יש מקום לשיפור זה. כפי שנאמר בהגנות של הפרויקט, בהינתן והיה לנו עוד זמן, היינו מפתחים AI שמקבל קלט של טווח זמנים למשחק וקטגוריה ואז מוציא המלצה עבור המשתמש בהתאם להיסטוריית החיפוש שלו. |
| " אני חושבת שההחזרה צריכה להיות ספציפית יותר" | לא ברור לנו על איזו החזרה מדובר וכיצד היא אמורה להיות ספציפית...  ניתן להבין מהערה זו שהסטודנטית המדוברת לא מבינה במשחקיות, ולא מבינה כי השאילתות יוצרות יחדיו בסיס ל-UI שמספק כלי נחוץ לגיימרים ביום יום. |
| "שימוש בscraping טהור ללא שימוש בapi" | שימוש ב-scraping טהור הוא אפשרי בהחלט באתר של Steam אך מציע מידע הרבה יותר גולמי ומסובך לעיבוד מאשר השימוש ב-API שמצד אחד מספק יותר נתוני משתמשים בצורה מאורגנת אך מצד שני באמת לא חשוף למשתמשים בדפים השונים באתר. |

1. מהו היו האתגרים בקורס / פרויקט כצוות?

אחד האתגרים המרכזיים בפרויקט כצוות היה התיאום בין חברי הצוות, במיוחד במקרים שבהם היינו צריכים לחלק משימות טכניות שונות. בנוסף, נדרשנו להתמודד עם בעיות טכניות שנבעו ממורכבות המידע והפורמטים השונים שהאתר Steam מספק. לבסוף, ניהול הזמן היה קריטי כדי לעמוד בלוח הזמנים של הקורס ולהבטיח שהפרויקט יעמוד בדרישות הקורס.

כמו כן, חלוקת תרגילי הבית בקורס הייתה מאתגרת מכיוון שנדרשנו להתמודד עם חומר תיאורטי, שאמנם נלמד בהרצאה אך היה לא פשוט להטמיע בתרגילים השונים. היינו צריכים לאזן בין פתירת המטלות התיאורטיות לבין התקדמות בפיתוח הזחלן כקבוצה. זה דרש תכנון קפדני של לוח הזמנים שלנו, וכמובן גם שיתוף פעולה ותמיכה הדדית כדי שכל חברי הצוות יעמדו בדרישות הן של תרגילי הבית והן של הפרויקט.

יש להגיש קישור לריפו שלכם ב -GIT, הכולל את כל קבצי האתר, ותיקייה בשם project שבה קובץ וורד ובו מענה על השאלות.

ניתן לפנות אליי בכל שאלה.

בהצלחה!