**שם הקורס: ניהול מידע מבוזר (096224)**

**מגישים: נימרוד סולומון (ת.ז 206574733) ומתן שילוני (ת.ז 208634469)**

**מספר תרגיל הבית: פרויקט חלק א', חלק 3**

**תאריך הגשה: 09.06.2022**

**חלק ג' – Design**

במסמך זה נציע תיכון למסד הנתונים בהתבסס על התובנות אליהן הגענו בחלק הקודם של הפרויקט. כפי שציינו בחלק הקודם, התובנות שלנו היו מבוססות ומונחות בעיקר על הערים, מתוך תקווה למצוא תובנות מעניינות שיהיו רלוונטיות לתכנון החלוקה האופקית והאנכית של הטבלאות במסד. בחלק הקודם מצאנו לא מעט תובנות מעניינות על המסד, שברובן נשתמש לתיכון.

בשורות הבאות נציג את התיכון למסד, טבלה-טבלה. קצרה היריעה מלהרחיב על כל עמודה ושורה בכל טבלה, לכן נתייחס לחלוקה בקווים מעט יותר כלליים, בדגשים, וככל שלא נציין פרט מסוים, ניתן יהיה להניח שהוא אינו קריטי לתיכון בהתבסס על התובנות שמצאנו בחלק הקודם.

**הטבלה movies**:

חלוקה אופקית - דגשים

1. **ז'אנרים**
2. **תל אביב**: נשים לב כי מהתובנות שמצאנו עולה כי הז'אנרים מהווים חלק הארי מבין כלל השאילתות בהן מופיעה התייחסות לתל אביב. כלומר, נוכל להשמיט ז'אנרים שאינם נמנים על הקבוצה לעיל לטובת הפרגמנטציה באתר .
3. **ירושלים**: ניתן לראות מהתובנות כי בעיר זו הז'אנרים מאוזנים יחסית זה לזה ואין רוב מובהק לאף קבוצה מצומצמת (בגודל שכדאי לעשות עבורו פרגמנטציה).
4. **חיפה**: מהתובנות שמצאנו בחלק הקודם עולה כי הז'אנר מחזיק ברוב מובהק בהופעות בשאילתות מעיר זו. על כן, באתר נוכל להשמיט סרטים שאינם מז'אנר זה כלל (כלומר מספיק שאחד הז'אנרים של סרט הוא על מנת שהוא כן יישמר באתר זה).
5. **אילת**: ניתן לראות מהתובנות כי בעיר זו הז'אנרים מאוזנים יחסית זה לזה ואין רוב מובהק לאף קבוצה מצומצמת (בגודל שכדאי לעשות עבורו פרגמנטציה).
6. **טבריה**: נשים לב כי מהתובנות שמצאנו עולה כי הז'אנרים מהווים חלק הארי מבין כלל השאילתות בהן מופיעה התייחסות לטבריה. כלומר, נוכל להשמיט ז'אנרים שאינם נמנים על הקבוצה לעיל לטובת הפרגמנטציה באתר .
7. את הרשומות המכילות סרטים מז'אנרים שלא צוינו עד כה נשמור באתר אקראי, לצורך העניין - (בהמשך נדאג לשמור גם באתרים אחרים לצורך איזון), ובלבד שהן גם לא עונות על אף תנאי שיצוין בהמשך (אחרת נשמור אותו במקום אחר בהתאם לתנאי).
8. **שפה**
9. ניתן להסיק מהתובנות של ניתוח הנתונים כי בערים תל אביב וירושלים יש עניין רב בסרטים באנגלית ובעברית, בעוד שאר השפות אינן זוכות להתעניינות משמעותית בערים אלו/בערים האחרות.
10. מהתבוננות קצרה בדאטה קל להתרשם שרוב הסרטים הם, בין היתר, גם באנגלית.
11. לכן, את כל הסרטים שאינם בעברית לא נפריד ע"פ מאפיין השפה.
12. הסרטים בעברית ימוקמו בתל אביב וירושלים בצורה משוכפלת, לאור ההתעניינות שצוינה לעיל.
13. **מדינה**
14. מכיוון שהסרטים הישראליים פופולריים מאוד, ואך ורק בירושלים ובתל אביב, נמקם את הרשומות בהן מופיעים סרטים ישראליים בצורה משוכפלת בתל אביב ובירושלים.
15. בערים אחרות אין התעניינות במאפיין זה.
16. לכן, רשומות שאינן עוסקות בסרטים ישראליים, נשמור באתר אקראי, לצורך העניין – , ובלבד שהן גם לא עונות על אף תנאי שצוין לעיל או יצוין בהמשך (אחרת נשמור אותו במקום אחר בהתאם לתנאי).
17. **חברות הפקה**
18. נשים את כל הסרטים של חברות ההפקה: Pixar Animation Studios, Walt Disney Pictures, Warner Bros באילת, שכן חברות ההפקה מאוד פופולריות באילת ופופולריות כמו כל חברת הפקה אחרת בשאילתות הנוגעות לערים אחרות.
19. את הרשומות הנותרות נשמור באתר אקראי, לצורך העניין – ובלבד שהן גם לא עונות על אף תנאי שצוין לעיל או יצוין בהמשך (אחרת נשמור אותו במקום אחר בהתאם לתנאי).
20. **שנה**
21. מהתבוננות קצרה על הדאטה ניתן לראות כי ישנם סרטים שהופקו לפני שנת 1990, ואולם כפי שראינו בחלק הקודם, השאילתות מתייחסות רק לסרטים משנת 1990 ואילך. לכן, נמקם רשומות של סרטים שיצאו לפני 1990 באחד האתרים באקראי שכן אף שאילתה לא ניגשת אליהם. לצורך העניין – , ובלבד שהן גם לא עונות על אף תנאי שצוין לעיל או יצוין בהמשך (אחרת נשמור אותו במקום אחר בהתאם לתנאי).

חלוקה אנכית - דגשים:

1. את העמודות: overview, revenue, tagline, title, יחד עם המפתח movie\_id נשים באחד האתרים באקראי, לצורך העניין שכן אף שאילתה לא ניגשת אליהן.
2. מימשנו את האלגוריתם[[1]](#footnote-1) לביצוע פרגמנטציה אנכית שנלמד בכיתה והרצנו אותו על רלציה זו, בתקווה למצוא קורלציות חזקות בין עמודותיה השונות.
3. Text

   Description automatically generatedלהלן התוצאות:

המספר 3 מייצג את העמודה , והיא מהווה לצורך העניין פרגמנט נפרד משאר העמודות - .

1. לאחר הרצת האלגוריתם, הסקנו כי אין קורלציה משמעותית בין העמודות, וכי אין בידינו נתונים שעל בסיסם ניתן לחלק אנכית.
2. לכן, לא נפריד את הרלציה לפי העמודות אליהן יש התייחסות בשאילתות המשתמשים.
3. כלומר, בכל אתר תהיינה העמודות ששאילתות ניגשות אליהן, ואילו את העמודות הנותרות נשמור באחד האתרים באקראי, לצורך העניין .

**הטבלה credits**:

חלוקה אופקית - דגשים:

1. **שחקנים**
2. ניתן להסיק מהתובנות של ניתוח הנתונים כי בערים תל אביב וירושלים יש עניין בזהות השחקן המשחק בסרט, וכמעט אך ורק בשחקנים .
3. בערים אחרות יש שאילתות מעטות מאוד בהן השדה אינו ריק. על כן, הסקנו כי למשתמשים בערים אלו זהות השחקן פחות חשובה.
4. לכן, בערים תל אביב וירושלים נשמור רשומות בהן מופיעים אחד מהשחקנים לעיל.
5. את הרשומות הנותרות נשמור באתר אקראי, לצורך העניין – .

חלוקה אנכית - דגשים:

1. ברלציה זו קיימות מעט מדי עמודות כדי לבצע חלוקה אנכית שתייעל את מסד הנתונים (שתי עמודות למעט המפתח).

**הטבלאות tickets ו-users**:

הטבלאות tickets ו-users שונות מהטבלאות movies ו-credits: הן קטנות יותר מבחינת כמות המידע (ניתן אף לחזות בכך אם נתבונן בגודל הקבצים של הטבלאות שסופקו) ולמעט הערים השונות שמופיעות בכל רשומה (ב-tickets תחת השדה city וב-users תחת השדה user\_location) לא ניתן למצוא פרדיקטים ששאילתות מפרידות רשומות בטבלאות הללו על פיהם. לכן נחלק את הטבלה חלוקה אופקית בלבד, ונשים בכל אתר את הרשומות שבהן מופיעה העיר המתאימה לאתר.

**הטבלה queries**:

מההבחנה הפשוטה כי אף שאילתה לא ניגשת לרלציה זו, נוכל למקם את כולה או חלקים ממנה בכל אתר שנרצה.

**נספח – אלגוריתם , מימוש**

Text

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Scatter chart

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

1. ראה נספח מצורף בסוף המסמך [↑](#footnote-ref-1)