אוניברסיטת בן גוריון

הפקולטה למדעי הטבע

המחלקה למדעי המחשב

נושאים באנליזה בלתי מונחית של מידע

דו"ח מסכם

מאת:

אביאל ברזין - 318174711

יובל עמרמי - 316417591

בהנחיית:

פרופ' סיון סבתו

11/07/2021

תוכן עניינים

מטרות:

בפרויקט זה ברצוננו ללמוד על שיטות למידה בלתי מונחית של מידע על ידי ניתוח מאגרים. שיטת העבודה בה נתמקד בעבודה זו הינה **Clustering** - ניתוח אשכולות, קיבוץ אובייקטים לתתי קבוצות של מאגר המידע כך שהאובייקטים השייכים לקבוצה מסוימת דומים האחד לשני יותר מאשר הם דומים לאובייקטים בקבוצות האחרות.

בנוסף, במהלך העבודה נרצה לבצע היכרות עם שיטות שונות ללמידה בלתי מונחית וכן עם טכנולוגיות תכנותיות שונות.

רקע ומאגר מידע:

מאגר המידע עליו נעבוד הינו מאגר Movielens 1M המייצג דירוג של סרטים על פי הצבעת צופים. בעבודה זו ננסה לבצע ניתוח אשכולות למאגר, במטרה למצוא קבוצות סרטים שצופיהם זהים ובכך ליצור מערכת המלצות לסרטים. ההערכה היא שצופים יעדיפו לצפות בסרטים דומים לסרטים בהם צפו עד כה. לדוגמה צופה שצפה בסרט "להרוג את ביל" יעדיף לצפות ב"ספרות זולה" יותר מאשר בסרט "בת הים הקטנה". לכן חלוקה לאשכולות של הסרטים על סמך נתוני הצפייה של הצופים הקודמים תאפשר הצעה של סרט מאותה הקבוצה לצופה חדש בעל נתוני צפייה דומים.

בשל המרכיב האנושי בדירוג הסרטים נתייחס לרשומות הדירוג כאינדיקציה שצופה צפה בסרט אך לא ניתן משמעות לערך הדירוג עצמו.

לטובת ניתוח זה נגדיר פונקציית עלות שמעידה על טיב ההתאמה בין סרטים שונים. ככל שפונקציית העלות תהיה קטנה יותר כך נגדיר את הקלסיפיקציה כטובה יותר.

יהיו $k \geq 2$ מספר הסרטים, $M = \{m_1, m_2, \ldots\ldots m_k\}$ קבוצת הסרטים,

ו$C := \{C_1, C_2, \ldots\ldots C_n\}$ חלוקה כלשהי של הסרטים לn אשכולות כך ש $\cup_{i=1}^{n} C_i = M$ וגם $i \neq j, C_i \cap C_j = \Phi$.

עבור קבוצת צופים כלשהי וסרט m כלשהו מקבוצת הסרטים, נגדיר את $p(m)$ להיות ההסתברות שצופה אקראי צפה בסרט. כמו כן עבור הסרטים $m_1, m_2 \in M$ נגדיר את $P(m_1, m_2)$ להיות ההסתברות שצופה אקראי צפה בשני הסרטים יחד.

באמצעות ערכים אלו נגדיר:

$$cost(C) = \Sigma_{i=1}^{n} cost(C_i)$$

$$cost(C_i) := \begin{cases} \Sigma_{m_1 \in C_i} \Sigma_{m2 \in C_i \; and \; m_2 \neq m_1} \frac{1}{|C_i| - 1} \log\left(\frac{1}{p(m_1, m_2)}\right) & |C_i| \geq 2 \\ \log\left(\frac{1}{p(m)}\right) & |C_i| = 1, C = \{m\} \end{cases}$$

להיות פונקציית העלות.
נציין שהדוגמה בדף ההנחיות מתארת מצב בו בסכמה מתקיים $m_1 < m_2$.
מכיוון שהתוצאות מעניינות יותר עבור ההשוואה בין פונקציות העלות עם הנחה זו נחשב כך את התוצאות.
מכיוון שחישוב הערך המינימלי לפונקציה זו הוא בעיה קשה חישובית נחפש אלגוריתמים שונים שיספקו קירוב לבעיה.

האלגוריתם הנאיבי:

אלגוריתם זה ניתן לנו בהגדרות העבודה ותפקידו לעשות קירוב לפונקציית העלות על ידי פישוטה. ראשית נעבור על רשימת הסרטים ונבדוק עבור כל זוג $m_1, m_2 \in M$ האם הם מקיימים קורלציה חיובית כלומר $p(m_1, m_2) \geq p(m_1) \cdot p(m_2)$ או שמא הקורלציה שלהם שלילית. נסמן זוגות בעלי קורלציה חיובית ב"+" ושלילית ב"-".

כעת נבצע על הזוגות ניתוח Clustering בשם *Correlation Clustering*. גם בעיה זו נחשבת בעיה *NP* קשה אולם קיים קירוב פולינומי, בעל עלות שהיא לכל היותר פי 3 מהעלות של הפתרון האופטימלי, בו נשתמש בניתוח זה. האלגוריתם בוחר סרט *m* כלשהו מהרשימה ויוצר אשכול המכיל את כל הסרטים המסומנים ב"+" ביחס לסרט *m*. לאחר מכן חוזר האלגוריתם על הפעולה עבור תת קבוצת הסרטים שלא הוכנסו באשכול. האלגוריתם מסתיים כאשר כל הסרטים הוכנסו לקבוצה.

אלגוריתם משופר לחישוב חלוקה:

שאלה 1 בהגדרות העבודה

על מנת למצוא אלגוריתם שיספק תוצאות טובות יותר לחלוקה דנו במספר כיוונים.
ראשית, שקלנו לבצע חישוב המבוסס על תורת הגרפים בניסיון למצוא מרכזי כובד של גרף הסרטים. כמו כן שקלנו לבצע חישוב של עצים פורשים בעלי משקל מקסימלי כאשר משקולות הגרף יהיו תלויות בכמות הצופים שצפו יחד באותו הסרט.
אולם לאחר מחשבה נוספת, ובשל הרצון ללמוד על תחומים נוספים בלמידה בלתי מונחית, בחרנו להתמקד באלגוריתם למציאת k-מרכזים.
באלגוריתם זה שני שלבים עיקריים:
האחד, <u>שלב ההקצאה</u>, בו משויך כל אלמנט, במקרה שלנו סרטים, למרכז כלשהו מבין k המרכזים במרחב. השיוך נעשה לפי קרבתו האוקלידית במרחב הרב ממדי של האלמנט למרכז. כלומר האלמנט משויך למרכז הקרוב ביותר אליו.
השני, <u>שלב העדכון</u>, בו לאחר ההקצאה של כל אלמנט למרכז כלשהו מחושב מיקום המרכז במרחב על ידי מיצוע של כל מיקומי האלמנטים המשויכים אליו.
התכנסות האלגוריתם מוגדרת על ידי השלב בו לא קיים שינוי בין מיקום המרכזים לפני ואחרי שלב העדכון, כלומר לא קימת תזוזה משמעותית של אלמנטים בין

מרכזים שונים.
בכל שלב בתהליך כל אלמנט משויך בדיוק למרכז אחד ולכן החלוקה למרכזים כשלעצמה מהווה חלוקה תקנית לk אשכולות על פי הגדרת החלוקה.
בתחילה חשבנו לפרוש את דגימות הסרטים שלנו ביחס למרחב הצופים כולו כך שסרט שנצפה על ידי מספר צופים יקבל ערך בכל אחד מהממדים הללו אולם בשל גודל הנתונים הבנו כי כמות ממדים הנקבעת על פי כמות הצופים מאטה את ריצת האלגוריתם במידה ניכרת.
על מנת לפתור בעיה זו הרחבנו את האלגוריתם לאלגוריתם k מרכזים כפול כאשר מרחב הסרטים נפרש על ידי מרחב המרכזים של הצופים ובמקביל מרחב הצופים נפרש על ידי מרחב המרכזים של הסרטים. ככל שמרכז מסוים מכיל יותר אלמנטים כך עוצמתו על חישוב המרחק תשפיע יותר.
הזזת צופים שונים ממרכז למרכז תתבצע על פי רמת הקרבה בין הסרטים בהם הם צפו לבין סרטים ש"נראו" על ידי המרכז כך שככל שהמרכז ממוקם על פני יותר סרטים בהם צפה הסיכוי שישתייך הצופה למרכז גדלה. בצורה דומה הסרטים ישויכו לקבוצות צופים דומות.
נציין שבתחילת ריצה האלגוריתם נבחרים המרכזים על ידי חלוקה אקראית של האלמנטים השונים לקבוצות שונות.
שיפור שהוסף לאלגוריתם בשלבים מאוחרים הינו קבצי קאש ששומרים את טעינת הרשומות של הצופים והסרטים. קבצים אלו מתעדכנים במהלך הריצה ומייעלים את ריצת האלגוריתם על אותם הנתונים בפעם הבאה.

אתגרים במהלך העבודה:

שאלה 3 בהגדרות העבודה

אחד הנושאים שהעסיק אותנו במהלך העבודה היה שפת התכנות בה העבודה תכתב. מכיוון שרצינו אלגוריתם שיעמוד בדרישות לזמן ריצה סביר החלטנו להשתמש בשפה המבצעת ייעול והפשטה של פעולות ולכן בחרנו בשפה פונקציונלית – Scheme.

פרט מידע ששמנו לב אליו הינו שהרנדומליות של השפה scheme דטרמיניסטית.
לפיכך, החלוקה האקראית שמתרחשת בתחילת האלגוריתם חוזרת על עצמה עבור כל הרצה של אותם הנתונים ולכן תבטיח את מסלול השיפור של האלגוריתם מאיטרציה לאיטרציה ותוביל לאותן נקודות קיצון בכל ריצה, כך שבהרצה חוזרת של האלגוריתם יתקבל אותו הפתרון.
אם כך, אם נרצה לבצע שימוש סטטיסטי באלגוריתם לטובת מציאת k מרכזים אופטימליים נאלץ לקודד את החזרה על הקוד בתוכנית המקורית בניגוד להרצה בשפות אחרות שלא תדרוש זאת. אולם, מכיוון שאנו מעוניינים בפתרון יחיד עבור כל דאטאסט, נושא זה לא השפיע על העבודה הנוכחית.

בנוסף לשאלת השפה נדרשנו לבחור מתודולוגיית עבודה על בסיסה יעבוד האלגוריתם. אלגוריתם k המרכזים מתבסס על ערך k מסוים ממנו הוא מתחיל. אף כי קיימת האפשרות שמרכזים מסוימים 'ייעלמו' אם כל האלמנטים שבהם יעברו לקבוצה אחרת באותה האיטרציה בהכרח לא ייווצרו מרכזים חדשים לאחר תחילת ריצת האלגוריתם. לפיכך שאלת בחירת הk תשפיע על התוצאה הסופית של פונקציית העלות ותהליך ה*clustering* . ראינו כי קיים טווח בו הגדלת *k* עבור הצופים גרמה לירידה משמעותית של ערך העלות, אולם במקביל זמן הריצה של האלגוריתם והזמן שלקח לו להגיע להתייצבות עלה.
פרט זה תואם את התיאוריה שנאבד מידע בעת איחוד של קבוצות גדולות של צופים לידי מרכז אחד בעל ממד אחד אם ערכי הצפייה שלהם שונים.
ערכי ברירת המחדל של האלגוריתם הינם *k=35* לסרטים ו- *k=50* לצופים.
ניתן לשנות ערכים אלו באלגוריתם על ידי הוספת הדגלים ACoordc -- עבור הסרטים ו- BCoordc -- עבור הצופים ואחריהם מספר.
שיפור אפשרי לאלגוריתם שיצרנו הינו הוספה של מנגנון חיפוש שתפקידו להריץ את המערכת עם מספר ערכים שונים ולראות מי מבניהם מספק את התוצאות הטובות ביותר עבור קבוצת סרטים נבחרת וזמן ריצה נתון מראש.
השערת בסיס טובה הינה *k* השווה לשורש הערכים. זאת בשל האיזון בין כמות הקבוצות וגודלן הממוצע.

שאלה נוספת בנושא האלגוריתם הייתה שאלת התזמון.
מכיוון שאנו מבצעים את האלגוריתם על שתי מערכות תלויות, האחת של הסרטים והשנייה של הצופים, נשאלת השאלה כמה איטרציות לבצע עבור כל אחת מהמערכות לפני המעבר לעדכון המערכת השנייה. בתחילה ניסינו לבצע תזמון התלוי בנקודת השבת של כל מערכת. כלומר, כאשר אין שינוי בין שלבי עדכון שונים, האלגוריתם יעבור לעדכון המערכת השנייה. אף כי תוצאות מערכת תזמון זו היו הטובות ביותר עבור מאגרי נתונים קטנים, כאשר עברנו למאגרים גדולים יותר, ובמיוחד כאשר הגדלנו את ערכי k בהתאמה לגודלי המאגרים, זמן הריצה של האלגוריתם התארך בצורה משמעותית.
לאחר מכן ניסינו לבצע אלגוריתמים משולבים בקני מידה שונים. כלומר מחזור שלם, יבצע מספר איטרציות, קבוע מראש, במרחב הצופים ולאחר מכן מספר איטרציות אחר במרחב הסרטים. המערכת תבצע מספר מחזורים עד שתופסק על ידי תנאי עצירה כלשהו.
מהבדיקות שעשינו בנושא ראינו שהאיטרציות שבמרחב הצופים משנות יותר מרכזים מאשר האיטרציות במרחב הסרטים ועל כן הוחלט לבצע 4 איטרציות על מרחב הצופים על כל איטרציה אחת במרחב הסרטים.
כמו כן, מכיוון שזמן האלגוריתם עשוי להשתנות על פי הקלט והפרמטרים השונים החלטנו להגביל את זמן הריצה של כל הפעלה ל10 שניות כברירת מחדל. אולם בשלב ההרצה שבדוח חישבנו את האלגוריתם עם 40 שניות כדי לראות תוצאות

מעניינות. תנאי העצירה של האלגוריתם הוגדר להיות הגבלת הזמן במקום נקודות השבת על מנת להתאים לדרישה זו. כך זמן הריצה של האלגוריתם בעבור כל קלט (עד כדי זמן העיבוד של הקלט והפלט, התלוי בנפחו) הוא זהה וידוע מראש.
מכיוון שבכל שלב החלוקה למרכזים מהווה חלוקה תקינה לאשכולות, השיפור בין איטרציה לאיטרציה צפוי לשפר את ערך העלות גם אם לא יגיע לנקודת המינימום המקומית שהאלגוריתם מסוגל לספק והפלט היוצא יהווה פתרון חוקי.

אף כי מתוצרי הבדיקה שלנו, נראה כי ככלל האלגוריתם החדש מספק תוצאות טובות מתוצאות האלגוריתם הנאיבי, מכיוון שערך המשקל אותו האלגוריתם מצמצם הינו המרחק האוקלידי בין האלמנטים והמרכזים ולא ערך העלות שהוגדר לנו בעבודה, לעיתים האלגוריתם מבצע שיפור ביחס לערך המרחק אך ערך העלות דווקא גדל. מקרים אלו קרו לרוב כאשר האלגוריתם האקראי חילק את קבוצת הסרטים למרכזים עבור סרטים בעלי "מעט דמיון". כאשר ביצענו מספר איטרציות על מקרים אלו בתוך האלגוריתם ואפשרנו לחלוקה האקראית להשתנות קיבלנו בחלק מהריצות תוצאות אחרות ואכן נראתה ירידה במהלך השלבים השונים של האלגוריתם.
פתרון לנושא זה בגרסאות עתידיות יהיה לבצע מספר איטרציות על האלגוריתם כולו ולבחור את הערך האופטימלי מבניהן או לבחור את ערך המינימום שהמערכת הגיעה אליו במהלך ריצת האלגוריתם בכל איטרציה.
יחד עם זאת, נראה כי גם בגרסה הנוכחית של האלגוריתם הפרמטר העיקרי שמשפיע על התוצאות הינו ערכם של k המרכזים בשני המרחבים. לכן, בחירה טובה של פרמטרים אלו הוציאה תוצאות טובות מהאלגוריתם הנאיבי גם כאשר זמן ריצת האלגוריתם הוגבל.

תוצאות מעניינות:

שאלה 4 בהגדרות העבודה

עבור פרק זה בחרנו להציג 3 תתי קבוצות של סרטים בעלי אופי מעניין.

כמו כן הוגדרו ערכי ברירת המחדל: $\boldsymbol{k_{movie} = 35; k_{viewer} = 50; time = 40\ sec}$

תת קבוצה 1

ראשית בחרנו בתת הקבוצה בעלת 30 הסרטים הנצפים ביותר. לכל הסרטים בקבוצה זו כמות צופים רבה וכן ולכן חלק גדול מהצופים צפו בהם. ההשערה היא שמכיוון שהרבה צופים ראו את הסרטים האלו הסיכוי למצוא "קליקה" בגרף גדל ולכן הציפייה היא שהאלגוריתם הנאיבי יהיה יעיל במצב זה. אולם לאחר הבדיקה ראינו כי ככל הנראה כמות הצופים הגדולה יוצרת גם פרופילי צפייה שצפו במעט סרטים מתוך הרשימה. נראה כי מסיבה זו ואולי גם מאופייה היחסי של ההסתברות, רבים מהסרטים דווקא נשארו בקבוצות נפרדות בתום חישוב האלגוריתם.

האלגוריתם הנאיבי מחזיר את הרשימה הבאה:

2762 "Sixth Sense, The (1999)"
1210 "Star Wars: Episode VI - Return of the Jedi (1983)"
2858 "American Beauty (1999)"
1197 "Princess Bride, The (1987)"
296 "Pulp Fiction (1994)", 318 "Shawshank Redemption, The (1994)",
527 "Schindler's List (1993)", 593 "Silence of the Lambs, The (1991)",
608 "Fargo (1996)"
2997 "Being John Malkovich (1999)"
1196 "Star Wars: Episode V - The Empire Strikes Back (1980)",
1240 "Terminator, The (1984)"
260 "Star Wars: Episode IV - A New Hope (1977)"
480 "Jurassic Park (1993)", 589 "Terminator 2: Judgment Day (1991)"
858 "Godfather, The (1972)"
1 "Toy Story (1995)", 1265 "Groundhog Day (1993)"
1198 "Raiders of the Lost Ark (1981)"
1617 "L.A. Confidential (1997)"
356 "Forrest Gump (1994)"
2028 "Saving Private Ryan (1998)"
2396 "Shakespeare in Love (1998)"
110 "Braveheart (1995)"
2716 "Ghostbusters (1984)"
2571 "Matrix, The (1999)"
1270 "Back to the Future (1985)"
1580 "Men in Black (1997)"
1097 "E.T. the Extra-Terrestrial (1982)"
2628 "Star Wars: Episode I - The Phantom Menace (1999)"

221.09447311183814

האיחוד בין סרטים שונים לא מפיק תוצאות בעלות הגיון מיוחד, אף כי הגיוני שצופים שצפו ב"חומות של תקוה" צפו גם ב"רשימת שינדלר" בשל היותם סרטים זוכי שבחים שיצאו לאקרנים באותה התקופה. לעומת זאת, הקרבה בין "רשימת שינדלר" ו"פרגו" נשמעת לוגית פחות.

האלגוריתם המשופר מחזיר תוצאות בעלות עלות נמוכה יותר:

1 "Toy Story (1995)", 356 "Forrest Gump (1994)"
110 "Braveheart (1995)", 260 "Star Wars: Episode IV - A New Hope (1977)"
296 "Pulp Fiction (1994)", 318 "Shawshank Redemption, The (1994)"
480 "Jurassic Park (1993)", 1240 "Terminator, The (1984)",
1270 "Back to the Future (1985)", 1580 "Men in Black (1997)"
527 "Schindler's List (1993)"
589 "Terminator 2: Judgment Day (1991)", 2571 "Matrix, The (1999)"
593 "Silence of the Lambs, The (1991)", 608 "Fargo (1996)"
858 "Godfather, The (1972)"
1097 "E.T. the Extra-Terrestrial (1982)"
1196 "Star Wars: Episode V - The Empire Strikes Back (1980)",
1210 "Star Wars: Episode VI - Return of the Jedi (1983)"
1197 "Princess Bride, The (1987)", 2628 "Star Wars: Episode I - The Phantom Menace (1999)"
1198 "Raiders of the Lost Ark (1981)"
1265 "Groundhog Day (1993)"
1617 "L.A. Confidential (1997)"
2028 "Saving Private Ryan (1998)"
2396 "Shakespeare in Love (1998)", 2997 "Being John Malkovich (1999)"
2716 "Ghostbusters (1984)"
2762 "Sixth Sense, The (1999)", 2858 "American Beauty (1999)"
110.0183426095424

אף שהאיחוד בין "צעצוע של סיפור" ו"פורסט גאמפ" נשמע פחות לוגי בשל האופי השונה של הסרטים ניתן לראות ששני סרטי "מלחמת הכוכבים" אוחדו יחד וקיים הגיון באיחוד של הסרטים "פארק היורה" ו "שליחות קטלנית".
מעניין לראות שבשני האלגוריתמים אוחדו יחד "ספרות זולה" ו"חומות של תקווה" כמו גם "שתיקת הכבשים" ו"פארגו".
נציין שהאלגוריתם החדש הציג את אותן התוצאות גם עבור ריצה של 120 שניות במקום 40 כך שככל הנראה הגיע למינימום מקומי.
ניתן לשער שהגרלת הסרטים האקראית היא זו שהפרידה לקבוצות צפיה שונות את הסרטים "מלחמת הכוכבים" 4 ו-5 שככל הנראה הכילו גם צופים זהים רבים, שצפו בשני הסרטים.

תת קבוצה 2

לקבוצה זו בחרנו 30 סרטים בעלי נתוני צפיה נמוכים, סרטים בעלי 15 או 14 צופים (יותר מ-10 צופים לפי דרישות העבודה).
כאן שיערנו שהעבודה שהסרטים מפוזרים במרחב וככל הנראה לא מחזיקים בהרבה צופים משותפים תגדום לך שהאלגוריתם הנאיבי לא ימצא הרבה קבוצות לאחד, או יאחד את רוב הסרטים בשל היותם בלתי תלויים. כמו כן מכיוון שהאלגוריתם המשופר מבצע הגרלה באקראי והמרחב הנפרש בת"ל בין הסרטים השונים, לא יתבצעו שינויים רבים במיקום המרכזים.

האלגוריתם הנאיבי מחזיר את הרשימה הבאה:

751 "Careful (1992)", 2388 "Steam: The Turkish Bath (Hamam) (1997)"
189 "Reckless (1995)", 2079 "Kidnapped (1960)"
1574 "Fall (1997)", 2824 "On the Ropes (1999)", 3283 "Minnie and Moskowitz (1971)",
3374 "Daughters of the Dust (1992)"
3883 "Catfish in Black Bean Sauce (2000)"
124 "Star Maker, The (Uomo delle stelle, L') (1995)",
2101 "Squanto: A Warrior's Tale (1994)"
3611 "Saludos Amigos (1943)"
1656 "Swept from the Sea (1997)", 2008 "This World, Then the Fireworks (1996)"
1153 "Raw Deal (1948)", 1154 "T-Men (1947)"
1520 "Commandments (1997)", 2415 "Violets Are Blue... (1986)"
681 "Clean Slate (Coup de Torchon) (1981)"
2627 "Endurance (1998)"
807 "Rendezvous in Paris (Rendez-vous de Paris, Les) (1995)"
3351 "Two Thousand Maniacs! (1964)"
3581 "Human Traffic (1999)"
1423 "Hearts and Minds (1996)",
1450 "Prisoner of the Mountains (Kavkazsky Plennik) (1996)"
1349 "Nosferatu a Venezia (1986)", 1989 "Prom Night III: The Last Kiss (1989)"
2298 "Strangeland (1998)"
2981 "Brother, Can You Spare a Dime? (1975)"
390.6501304634521

נראה שעל אף התיאוריה מספר האחודים בקבוצה גדול יחסית אך לא כולל קבוצות גדולות. ככל הנראה רוב ההסתברויות לא היו בלתי תלויות
$p(m_1, m_2) \neq p(m_1) \cdot p(m_2)$ כלומר הייתה דווקא קורלציה שלילית בין הרבה מהסרטים שנבחרו כאלמנט m שבחישוב האלגוריתם בכל שלב. אחרת, הינו מצפים לראות קבוצות גדולות יותר מ-3 סרטים.

יחד עם זאת חישוב האלגוריתם המשופר אכן הציג את התוצאות ששיערנו ועל פי נתוני ה-*LOG* שלו נכנס לנקודת קיצון מקומית כבר בשלב הגרלת הנתונים הראשונה:

124 "Star Maker, The (Uomo delle stelle, L') (1995)"
189 "Reckless (1995)"
681 "Clean Slate (Coup de Torchon) (1981)", 1154 "T-Men (1947)"
751 "Careful (1992)", 1153 "Raw Deal (1948)"
807 "Rendezvous in Paris (Rendez-vous de Paris, Les) (1995)",
2824 "On the Ropes (1999)"
1349 "Nosferatu a Venezia (1986)", 2298 "Strangeland (1998)",
3581 "Human Traffic (1999)"
1423 "Hearts and Minds (1996)", 2627 "Endurance (1998)"
1450 "Prisoner of the Mountains (Kavkazsky Plennik) (1996)",
3283 "Minnie and Moskowitz (1971)",
3883 "Catfish in Black Bean Sauce (2000)"
1520 "Commandments (1997)", 2388 "Steam: The Turkish Bath (Hamam) (1997)"
1574 "Fall (1997)"
1656 "Swept from the Sea (1997)"
1989 "Prom Night III: The Last Kiss (1989)", 3351 "Two Thousand Maniacs! (1964)"
2008 "This World, Then the Fireworks (1996)", 3611 "Saludos Amigos (1943)"
2079 "Kidnapped (1960)"
2101 "Squanto: A Warrior's Tale (1994)"
2415 "Violets Are Blue... (1986)"
2981 "Brother, Can You Spare a Dime? (1975)"
3374 "Daughters of the Dust (1992)"
129.16817530353427

האלגוריתם מחלק את הקבוצות באקראי בתחילת הריצה ומייצר את אותה כמות של קבוצות אך מקבל ערך נמוך יותר. דבר זה מעיד על המשמעות של אקראיות הבחירה עבור שני המודלים, שכן רוב האחודים שנעשו באלגוריתם השני היו עשויים להיעשות באלגוריתם הראשון אם סדר הבחירה היה אחר. כאשר הגיע האלגוריתם לסרט "Commandments (1997)" 1520 הסרט "Steam: The Turkish Bath (Hamam) 2388 (1997)" כבר נבחר לאשכול אחר ולא יכל להיות מצורף לאשכול החדש.
כך שילובים שערכם לא מוסיף לסכמה הסופית או מוסיף מעט, כמו איחוד של ערכים בלתי תלויים או של ערכים בעלי קורלציה חיובית נמוכה, מונעים את האיחוד של ערכים שהקורלציה בינהם חיובית ממש או חיובית בערכים גדולים מאיחוד של ערכים קודמים.

תת קבוצה 3

עבור קבוצה זו בחרנו 30 אלמנטים באקראי מכל המאגר:
נרצה להראות את השוני בין ערכי k שונים.
נדגיש שכל הריצות נעשו על אותו הסאבסט ועל מנת למנוע שינויים באלגוריתם האקראי וכן לאפשר שחזור תוצאות ההכנסה נעשתה על פי מספרו הסידורי של אינדקס הסרט.

ראשית תוצאות הבקרה שלנו הינן האלגוריתם הנאיבי:

2797 "Big (1988)", 2918 "Ferris Bueller's Day Off (1986)"
1545 "Ponette (1996)"
1197 "Princess Bride, The (1987)"
48 "Pocahontas (1995)",
594 "Snow White and the Seven Dwarfs (1937)",
595 "Beauty and the Beast (1991)", 661 "James and the Giant Peach (1996)",
914 "My Fair Lady (1964)", 919 "Wizard of Oz, The (1939)", 938 "Gigi (1958)",
2018 "Bambi (1942)", 2687 "Tarzan (1999)"
2398 "Miracle on 34th Street (1947)"
1193 "One Flew Over the Cuckoo's Nest (1975)"
1035 "Sound of Music, The (1965)"
1270 "Back to the Future (1985)"
2355 "Bug's Life, A (1998)"
2321 "Pleasantville (1998)"
1097 "E.T. the Extra-Terrestrial (1982)"
720 "Wallace & Gromit: The Best of Aardman Animation (1996)",
745 "Close Shave, A (1995)"
3105 "Awakenings (1990)"
3408 "Erin Brockovich (2000)"
1287 "Ben-Hur (1959)"
2804 "Christmas Story, A (1983)"
2340 "Meet Joe Black (1998)"
2791 "Airplane! (1980)"
527 "Schindler's List (1993)"
1721 "Titanic (1997)"
598.1172637993851

נשים לב שסרטי ילדים רבים אוחדו עם הסרט "פוקהונטס".

כעת נציג את תוצאות האלגוריתם המשופר עבור ערכי k שונים ועבור 40 שניות ריצה של האלגוריתם:

א. עבור ערכי ברירת המחדל $k_{movie} = 35; k_{viewer} = 50$ האלגוריתם המשופר מחזיר:

48 "Pocahontas (1995)", 661 "James and the Giant Peach (1996)",
2687 "Tarzan (1999)"
527 "Schindler's List (1993)"
594 "Snow White and the Seven Dwarfs (1937)", 2018 "Bambi (1942)"
595 "Beauty and the Beast (1991)"
720 "Wallace & Gromit: The Best of Aardman Animation (1996)",
745 "Close Shave, A (1995)"
914 "My Fair Lady (1964)", 1035 "Sound of Music, The (1965)"
919 "Wizard of Oz, The (1939)"
938 "Gigi (1958)", 2340 "Meet Joe Black (1998)"
1097 "E.T. the Extra-Terrestrial (1982)"
1193 "One Flew Over the Cuckoo's Nest (1975)"
1197 "Princess Bride, The (1987)"
1270 "Back to the Future (1985)", 2804 "Christmas Story, A (1983)",
3408 "Erin Brockovich (2000)"
1287 "Ben-Hur (1959)"
1545 "Ponette (1996)"
1721 "Titanic (1997)"
2321 "Pleasantville (1998)"
2355 "Bug's Life, A (1998)"
2398 "Miracle on 34th Street (1947)"
2791 "Airplane! (1980)"
2797 "Big (1988)", 2918 "Ferris Bueller's Day Off (1986)"
3105 "Awakenings (1990)"
111.17794766796641

ב. עבור ערכי $k_{movie} = 10; k_{viewer} = 20$:

48 "Pocahontas (1995)", 661 "James and the Giant Peach (1996)",
720 "Wallace & Gromit: The Best of Aardman Animation (1996)",
745 "Close Shave, A (1995)", 2687 "Tarzan (1999)"
527 "Schindler's List (1993)", 1097 "E.T. the Extra-Terrestrial (1982)",
1270 "Back to the Future (1985)"
594 "Snow White and the Seven Dwarfs (1937)", 595 "Beauty and the Beast (1991)",
2018 "Bambi (1942)"
914 "My Fair Lady (1964)", 938 "Gigi (1958)", 1287 "Ben-Hur (1959)",
1545 "Ponette (1996)", 2340 "Meet Joe Black (1998)",
2398 "Miracle on 34th Street (1947)"
919 "Wizard of Oz, The (1939)", 1197 "Princess Bride, The (1987)"
1035 "Sound of Music, The (1965)"
1193 "One Flew Over the Cuckoo's Nest (1975)", 2355 "Bug's Life, A (1998)"
1721 "Titanic (1997)", 3408 "Erin Brockovich (2000)"
2321 "Pleasantville (1998)", 2804 "Christmas Story, A (1983)",
3105 "Awakenings (1990)"
2791 "Airplane! (1980)", 2797 "Big (1988)", 2918 "Ferris Bueller's Day Off (1986)"
295.29745474647035

ג. עבור $k_{movie} = 10; k_{viewer} = 50$

48 "Pocahontas (1995)", 661 "James and the Giant Peach (1996)",
720 "Wallace & Gromit: The Best of Aardman Animation (1996)",
745 "Close Shave, A (1995)"
527 "Schindler's List (1993)", 1270 "Back to the Future (1985)"
594 "Snow White and the Seven Dwarfs (1937)", 595 "Beauty and the Beast (1991)",
919 "Wizard of Oz, The (1939)", 1035 "Sound of Music, The (1965)",
2018 "Bambi (1942)"
914 "My Fair Lady (1964)", 938 "Gigi (1958)", 1287 "Ben-Hur (1959)",
1545 "Ponette (1996)", 2340 "Meet Joe Black (1998)",
2398 "Miracle on 34th Street (1947)", 2687 "Tarzan (1999)"
1097 "E.T. the Extra-Terrestrial (1982)"
1193 "One Flew Over the Cuckoo's Nest (1975)", 2355 "Bug's Life, A (1998)"
1197 "Princess Bride, The (1987)", 1721 "Titanic (1997)"
2321 "Pleasantville (1998)", 2804 "Christmas Story, A (1983)",
3105 "Awakenings (1990)"
2791 "Airplane! (1980)", 2797 "Big (1988)", 2918 "Ferris Bueller's Day Off (1986)"
3408 "Erin Brockovich (2000)"
352.0970260755678

ד. עבור $k_{movie} = 35; k_{viewer} = 20$

48 "Pocahontas (1995)", 661 "James and the Giant Peach (1996)",
2687 "Tarzan (1999)"
527 "Schindler's List (1993)"
594 "Snow White and the Seven Dwarfs (1937)", 2018 "Bambi (1942)"
595 "Beauty and the Beast (1991)"
720 "Wallace & Gromit: The Best of Aardman Animation (1996)",
745 "Close Shave, A (1995)"
914 "My Fair Lady (1964)", 1035 "Sound of Music, The (1965)"
919 "Wizard of Oz, The (1939)"
938 "Gigi (1958)", 2340 "Meet Joe Black (1998)"
1097 "E.T. the Extra-Terrestrial (1982)", 1270 "Back to the Future (1985)"
1193 "One Flew Over the Cuckoo's Nest (1975)"
1197 "Princess Bride, The (1987)"
1287 "Ben-Hur (1959)"
1545 "Ponette (1996)"
1721 "Titanic (1997)"
2321 "Pleasantville (1998)"
2355 "Bug's Life, A (1998)"
2398 "Miracle on 34th Street (1947)"
2791 "Airplane! (1980)"
2797 "Big (1988)", 2918 "Ferris Bueller's Day Off (1986)"
2804 "Christmas Story, A (1983)", 3408 "Erin Brockovich (2000)"
3105 "Awakenings (1990)"
103.08144629266351

ה. עבור $k_{movie} = 5; k_{viewer} = 100$

48 "Pocahontas (1995)", 594 "Snow White and the Seven Dwarfs (1937)",
595 "Beauty and the Beast (1991)", 661 "James and the Giant Peach (1996)",
720 "Wallace & Gromit: The Best of Aardman Animation (1996)",
745 "Close Shave, A (1995)", 914 "My Fair Lady (1964)", 938 "Gigi (1958)",
1035 "Sound of Music, The (1965)", 1287 "Ben-Hur (1959)",
1545 "Ponette (1996)", 2018 "Bambi (1942)", 2340 "Meet Joe Black (1998)",
2398 "Miracle on 34th Street (1947)", 2687 "Tarzan (1999)"
527 "Schindler's List (1993)", 1270 "Back to the Future (1985)",
2791 "Airplane! (1980)", 2918 "Ferris Bueller's Day Off (1986)"
919 "Wizard of Oz, The (1939)", 1097 "E.T. the Extra-Terrestrial (1982)",
1197 "Princess Bride, The (1987)", 2797 "Big (1988)"
1193 "One Flew Over the Cuckoo's Nest (1975)", 1721 "Titanic (1997)",

2355 "Bug's Life, A (1998)"
2321 "Pleasantville (1998)", 2804 "Christmas Story, A (1983)",
3105 "Awakenings (1990)", 3408 "Erin Brockovich (2000)"
965.74244459382

ا. $\boldsymbol{k_{movie} = 35; k_{viewer} = 3}$

48 "Pocahontas (1995)", 2398 "Miracle on 34th Street (1947)", 2687 "Tarzan (1999)"
527 "Schindler's List (1993)", 1097 "E.T. the Extra-Terrestrial (1982)",
1197 "Princess Bride, The (1987)"
594 "Snow White and the Seven Dwarfs (1937)"
595 "Beauty and the Beast (1991)"
661 "James and the Giant Peach (1996)", 720 "Wallace & Gromit: The Best of Aardman Animation (1996)"
745 "Close Shave, A (1995)"
914 "My Fair Lady (1964)", 1035 "Sound of Music, The (1965)", 2018 "Bambi (1942)"
919 "Wizard of Oz, The (1939)"
938 "Gigi (1958)", 2340 "Meet Joe Black (1998)"
1193 "One Flew Over the Cuckoo's Nest (1975)", 1721 "Titanic (1997)",
2355 "Bug's Life, A (1998)"
1270 "Back to the Future (1985)"
1287 "Ben-Hur (1959)"
1545 "Ponette (1996)"
2321 "Pleasantville (1998)"
2791 "Airplane! (1980)"
2797 "Big (1988)", 2804 "Christmas Story, A (1983)",
2918 "Ferris Bueller's Day Off (1986)"
3105 "Awakenings (1990)"
3408 "Erin Brockovich (2000)"
142.15543010032226

מסקנות:
ראשית נשים לב שבבחירת נתונים "גרועה" כפי שניתן לראות בהרצה ה'. אכן ניתן ליצור מצבים בהם האלגוריתם הנאיבי טוב מהמשופר. יחד עם זאת ברוב המקרים כאשר הפרמטרים נמצאים בתווך "סביר" האלגוריתם מציג ערכים טובים מהאלגוריתם הנאיבי.
כמו כן, חשוב להזכיר שזמן הריצה של האלגוריתם חסום ב-40 שניות לשלב אלגוריתם k המרכזים הכפול. לכן, מכיוון שהגדלה של כמות מרכזי הצופים דורשת איטרציה יותר ארוכה על כל שלב הקצאה ושלב עדכון, האלגוריתם מבצע פחות איטרציות כאלו אם ערכי k גדולים יותר. כך ניתן להסביר את השיפור בביצועי האלגוריתם בין הרצה א' ל-ד' וכן בין הרצה ג' ל-ב'. נציין שנתוני הלוג של מצב ד' הראו במפורש כי אחת האיטרציות של פונקציית העלות הייתה זהה לפונקציית העלות ב-א' והשתפרה על ידי איטרציות עדכון של המרכזים.
נדגיש שככל שהמאגר מגוון יותר ויש יותר שוני בין נתוני הצפייה של הצופים השונים, כך חשוב יותר ערך k גבוה לצופים אולם במאגר הנוכחי כמות הסרטים קטנה ולכן ככל הנראה כמות הפרופילים של צופים שונים קטנה גם היא. לפיכך איחוד של צופים שונים לאותם המרכזים משפיע לרעה על פונקציית העלות רק כאשר ערך k מרכזי הצופים קטן במיוחד. בהרצה ו' ניתן לראות שאכן איחוד של כמות גדולה של צופים לכמות מרכזים קטנה גורעת מפונקציית העלות.
בנוסף, פרט חשוב בהרצה הנוכחית הינו שכמות הסרטים ברשימה הינו 30. לפיכך, כאשר אנו מספקים k בעבור הסרטים שגדול מערך זה כבר בשלב ההגרלה יתמלאו פחות מרכזי סרטים מהערך k. כך רוב הסרטים מתחילים מנותקים זה מזה ומאוחדים רק על ידי איחוד של פרופילי צופים שונים. מכיוון שהמאגר מספיק קטן כדי לאפשר יצירה של כמות מרכזים השווה לכמות הסרטים אנו יודעים כי איחודים בין מרכזים אלו התרחשו בשל ההגרלה האקראית בתחילה או בשל קירבה קטנה מספיק בין שני סרטים בלבד. כך הקבוצות המאוחדות לרוב יקטינו את ערכי העלות ולא יגדילו אותן.
אם כך, ניתן לסכם שכאשר מאגר המידע מאפשר זאת, הגדלת ערך k שאינה גוררת את הגדלת זמני הריצה במידה ניכרת, הן עבור הצופים והן עבור הסרטים, אכן משפרת את תוצרי האלגוריתם ומאפשרת הקטנה טובה יותר של ערכי העלות. אולם, האטה של האלגוריתם במידה ניכרת, עשויה להקטין את כמות האיטרציות ולכן עשויה לגרוע מתהליך השיפור ולהביא לתוצאה טובה פחות.

פרט מעניין נוסף בהרצה זו הינה העובדה שברוב האיטרציות אוחדו סרטי הילדים "במבי" ו"שלגיה ושבעת הגמדים" יחד. כמו גם הסרטים "ביג" ו"שמתי ברז למורה" ("Ferris Bueller's Day Off"). סרטים אלו אוחדו גם באלגוריתם המשופר וגם באלגוריתם הנאיבי ולכן בהכרח בעלי קורלציה חיובית. לפיכך, אפשר להצביע על קשר מסוים בין שיטת חישוב k המרכזים ו*Correlation Clustering* וכן בינם ובין המינימיזציה של פונקציית העלות.

השוואה סטטיסטית:

שאלה 5 בהגדרות העבודה

נתונים אלו נלקחו מאלגוריתם ברירת המחדל בו : $k_{movie} = 35; k_{viewer} = 50$.
נציג השוואה בין האלגוריתם הנאיבי והאלגוריתם המשופר עבור זמן ריצה של 40 שניות:

cost for algorithm 2	***cost for algorithm 1***	***Random dataset***
3038.283	*4241.474*	***1***
2862.624	*3242.282*	***2***
2489.675	*5776.577*	***3***
4022.207	*4706.173*	***4***
3434.738	*4432.061*	***5***
2317.126	*4455.225*	***6***
4346.961	*3735.492*	***7***
2759.004	*6009.247*	***8***
3053.999	*5457.078*	***9***
2995.412	*4738.671*	***10***
2418.786	*3953.326*	***11***
2347.167	*4399.237*	***12***
2002.547	*6152.773*	***13***
2395.929	*3773.454*	***14***
3643.373	*5890.576*	***15***
3610.822	*5276.273*	***16***
2270.425	*5292.749*	***17***
1961.901	*4368.979*	***18***
2830.425	*3635.692*	***19***
2109.12	*5036.984*	***20***
2845.526	***4728.71615***	***mean***
677.5446	***855.3958401***	***Standard deviation***

מתוצאות אלו ניתן לראות כי ככלל האלגוריתם המשופר מוציא תוצאות טובות מאלו של האלגוריתם הנאיבי, כאשר העלות הממוצעת של האלגוריתם המשופר הינה כ-60 אחוזים מהעלות של האלגוריתם הנאיבי. אף כי קיימות תתי קבוצות בהן האלגוריתם הנאיבי מוצא תוצאות טובות מאלו של האלגוריתם המשופר כמו בתת קבוצה 7, קיימות גם תתי קבוצות בהן האלגוריתם המשופר מוצא תוצאות טובות בהרבה כמו בתת קבוצה 13 בה התוצאה היא רק כ-30 אחוז מתוצאת האלגוריתם הנאיבי.

באופן כללי, נראה שערכי העלות האבסולוטיים של האלגוריתם המשופר קרובים יותר לערך הממוצע שלהם כפי שמציג הנתון של סטיית התקן אך במקביל סטיית

התקן באחוזים היא כ-24 אחוז מהממוצע, בעוד עבור האלגוריתם הנאיבי היא רק כ-18 אחוזים. לפיכך קרבתם היחסית של ערכי האלגוריתם הנאיבי טובה משל המשופר כך שערכי האלגוריתם הנאיבי יהיו לרוב יותר קרובים יחסית למוצע שלהם.

אם כך, תוצאות ההתפלגות הסטטיסטית מעידות על סיכויים גבוהים לקבלת ערך טוב מהאלגוריתם המשופר יותר מאשר מתוצאות האלגוריתם הנאיבי. זאת, אף כי קיימות פלקטואציות בעלות אמפליטודה יחסית גדולה יותר בין הערכים שיתקבלו מאלגוריתם זה וכן כי לעיתים עשויה להתקבל תוצאה פחות טובה מתוצאת האלגוריתם הנאיבי.

נדגיש שנתונים אלו התקבלו בעבור ערכי k נתונים מראש אך עשויים להיות ערכי k עבורם התוצאות יהיו טובות יותר וכן טובות פחות בהרצת האלגוריתם המשופר. כמו כן הגבלת הריצה ל-40 שניות, ככל הנראה, הגדילה את ערכי העלות עבור הרצות מסוימות אולם הקטינה אותן עבור הרצות בהן ההגרלה הראשונית גרמה לקשר הפוך בין הקטנת המרחק בין האלמנטים והמרכזים ושווי פונקציית הערך. תופעה שציינו בסעיף האתגרים.
על פי נתוני התצפית שלנו נראה כי התופעה הראשונה נפוצה יותר אולם גם את התופעה השנייה ניתן לפתור על ידי ההצעות שהצגנו בסעיף האתגרים – על ידי הפעלה של האלגוריתם עם מספר פרמטרים שונים וחיפוש אחר הפתרון האופטימלי מבניהם. בנוסף, מכיוון שאין קשר חישובי בין התוצאות עבור k שונים ניתן להפעיל חישוב זה במקביל על מספר מעבדים שונים.
כך, על ידי הפעלה ארוכה יותר של האלגוריתם המשופר וכן מקבול של הפעולות אנו משערים שניתן יהיה להגיע לתוצאות טובות אף יותר.

השיפור בין האלגוריתמים

שאלה 2 בהגדרות העבודה

ניתן לראות על סמך סך הדוגמאות שמעל שאכן קיים שיפור בין האלגוריתם הנאיבי ובין האלגוריתם המשופר. ננסה לשער מדוע נגרם השיפור.
ראשית, האלגוריתם הנאיבי, כפי שנטען במסמך העבודה, מתחשב רק בערכים בעלי קורלציה חיובית בין האלמנט הנבחר והאלמנטים המצורפים אליו, אולם לא מתחשב בקורלציה השלילית שעשויה להיווצר בין הרכיבים הפנימיים של אותה הקבוצה. אם שני סרטים m_1, m_2 מקיימים קורלציה חיובית עם סרט m אך הקורלציה השלילית שלהם קטנה מחיבור הקורלציות החיוביות עם סרט m, איחוד הקבוצה עשוי לגרום להגדלה של ערך העלות במקום להקטנתו.
לעומת זאת, החלוקה המרחבית של המרכזים באלגוריתם המשופר ממקמת במרחב את האלמנטים בהתאם לנתוני הצפייה. כך ככל שערך הקורלציה בין

סרטים מסוימים יהיה גדול יותר נצפה שהפרישה המרחבית של סרטים אלו במרחב מרכזי הצופים תהיה קרובה יותר אוקלידית וכן ככל שערך הקורלציה קטן יותר הסרטים יהיו רחוקים יותר. כך מתחשב האלגוריתם המשופר הן בערכים החיובים והן בשליליים. כמו כן, איחוד של קבוצות שלמות נעשה לרוב כאשר כל איברי הקבוצה נמצאים בקרבה פיזית ולכן בין כל איבר בקבוצה צפויה קורלציה גבוהה ולא רק בין רכיבים שלה.

זאת אף כי קיימים מקרי קצה בהם ההגרלה הראשונית יוצרת איחוד של אלמנטים "שונים" מדי ולכן "הורסת" את פעילות האלגוריתם. תופעה זו עשויה להיגרם מערכי k קטנים מדי עבור תת הקבוצה המחושבת ובשל ערכים אקראיים שלא קיבלו זמן ריצה מספק כדי לבצע מספיק איטרציות על מנת להגיע לנקודת מינימום. כמו כן, לעיתים ההגרלה האקראית מפרידה בין אלמנטים בעלי פרופילי צפייה דומים או מאחדת בין פרופילים שונים אך בגלל טיב האלגוריתם – המתוכנן למצוא את המרכז הקרוב ביותר, שמושפע בעצמו מערכם – מיקום המרכז המקורי נשאר המרכז ה"טוב ביותר" עבורם. כך, הגרלה אקראית כלשהי עשויה לגרום ל"רלקסציה" של האלגוריתם סביב נקודת קיצון מקומית ולא גלובלית.

לסיכום, אף כי קיימים אתגרים באלגוריתם שבחרנו לממש ולדעתנו ניתן לשפר את תוצאותיו אף יותר מהתוצר הסופי שהגשנו במספר דרכים, בחלקם דנו בשלב האתגרים, התוצר שסיפקנו אכן עונה על מטרתנו ומביא תוצאות טובות יותר מתוצאות האלגוריתם הנאיבי. כמו כן הבחירה בתחום k מרכזים אפשרה לנו להתעמק בנושאי הלמידה הבלתי מונחית וכך לרכוש כלים וידע להמשך דרכינו.

קבצי ההגשה:

הגרסה הסופית של הקוד שלנו כתובה בשפה Scheme לטובתה ביקשנו שתותקן בשרת cicero5 התוכנית chezscheme9.5.

שורת ההפעלה הכתובה בקובץ בשם moviecluster מפעילה את קובץ הschemes scripts/executions/from-data-set.scm שבתהליך ריצתו משתמש בקבצים מהתיקיות: algorithms, fasl, libs, scripts.
הקבצים בתיקיות אלו משתמשים במיקומם היחסי ולכן האלגוריתם דורש שמירה על סדר התיקיות הנ"ל.

כמו כן בתיקיית ההגשה נמצאים שלושת קבצי ה-subset בהם השתמשנו ליצירת התוצאות המעניינות וכן תיקיית subsets בה נמצאים 20 ה-subset*random* בהם השתמשנו כדי ליצור את ההתפלגות הסטטיסטית.