BookBuster

פרוייקט מחט בערימת דאטה

מגישים:

	id	mail	CS id
Maryna Romanchuk	323255844	Maryna.Romanchuk@mail.huji.ac.il	marynar
David Pitts	313371080	David.Pitts@mail.huji.ac.il	Deven14423
Harel Yacovian	311319990	Harel.Yacovian@mail.huji.ac.il	harelyac

רקע

תעשיית הקולנוע היא תעשייה המוערכת בעשרות מליארדי דולרים ($136 מיליארד נכון ל-2018 ע"פ ויקיפדיה).
קהל יעד בכל הגילאים לוקח חלק בצריכה של התכנים הללו, דרך סרטים בבתי קולנוע, תוכניות טלוויזיה, סדרות, אפליקציות תוכן ייעודיות שיוצרות תוכן עצמאית (כגון נטפליקס) ומקורות נוספים.
תהליך יצירת סרט דורש משאבים רבים והוא מהווה סיכון במהותו: ניתן רק לנחש ולנסות להעריך אם סרט יצליח, או לחלופין היוצרים יספגו הפסדים. שלב ראשוני בתהליך הוא ההחלטה אם הרעיון הראשוני, הסיפור, שווה את ההשקעה הנדרשת להפוך לסרט. ההחלטה מתבססת על ניסיון אנושי ושיקול דעת של אנשים עם ותק מקצועי, ופחות על נתונים מדעיים והחלטות מנומקות היטב.
ע"פ מחקר של איגוד המו"לים בבריטניה (מצורף PDF), סרטים שמבוססים על מקורות קודמים- כגון ספר או קומיקס, ירוויחו 53% יותר (בעולם) לעומת סרטים אחרים. זאת משום שסרטים המבוססים על סרטים פופולריים ייהנו מהפופולריות הקיימת וקהל היעד המבוסס של הספר. בנוסף סרט מבוסס על ספר יטה להיות עשיר יותר מבחינת תוכן, בעל עלילה מפותחת יותר שתמשוך את קהל היעד. על כן לסרט המבוסס על ספר יש יתרון על פני סרטים שלא מבוססים על ספר.

על כן עולה השאלה המתבקשת- כיצד ניתן לקבוע על סמך הספר אם כדאי להפוך אותו לסרט או לא?
שאלת המחקר הזו מתעלמת מהשלבים הבאים בתהליך קבלת ההחלטות ביצירת הסרט ומתמקדת בהחלטה אם לתת 'אור ירוק' לסיפור עצמו.

מטרות

מלכתחילה רצינו להצליח לענות אם ספר שטרם הפך לסרט- כדאי להפוך אותו לסרט או לא? ניתן לענות על השאלה הזו ע"י למידה של דוגמאות ספרים שידוע לגביהם עד כה אם הפכו לסרטים או לא.
לכן בפרויקט התמקדנו במציאת דאטה רלוונטי, ניתוח שלו ומציאת פיצ'רים מובהקים סטטיסטית מתוכו. לבסוף נוכל לבצע למידה ולענות על השאלה לגבי ספר שטרם הפך לסרט.

<u>ראינו לנכון לבחון את המאפיינים הבאים:</u>

- כמות דירוגי משתמשים לספר, והרייטינג שהוא קיבל. (אינטואיטיבית רייטינג נשמע כמו מאפיין עיקרי!)
- האם ספרים שהפכו לסרטים מדורגים שונה מספרים שלא הפכו?
- הגורמים מאחורי הקלעים: המוציא לאור והסופרים. מי מהם כותב ספרים שנוטים להפוך לספרים?
- האם יש קשר לאורך הספר? (ספוילר: לא!)
- אולי יש קשר למיקום שבו הספר יצא לאור?
- עד כמה ספר מקורי? האם ספרים שהפכו לסרטים מאוד מקוריים ושונים מאלו שלא הפכו לסרטים?
- האם יש ז'אנרים פופולריים יותר שספרים מתוכם הופכים לסרטים?

הדאטה

את הדאטה אספנו בעצמנו מתוך אתר 'goodreads' שכן נדרשנו לפיצ'רים ספציפיים ולא היה בנמצא דאטה מוכן שכלל את הנתונים שרצינו לבחון. עבור כל פיצ'ר ביצענו scraping דרך ה-api של goodreads וניגשנו לעמודי הספרים, מהם דלינו את המידע. התבססנו על רשימות מתוייגות של goodreads לגבי ספרים שהפכו לסרטים וכאלה שלא על מנת לקבל כמות דאטה מספיקה ללמידה של ספרים שמתוייגים תחת התיוג 'האם הפך לסרט?'.

קיבלנו 1800 ספרים שהפכו לסרטים, 30,000 כאלה שלא.
מתוך הדטא דלינו רק ספרים שכמות הדירוגים שלהם גדולה מ-100. נותרנו עם 1580 ספרים שהפכו לסרטים, וגם כאלה שלא. מבחינת הסופרים- אספנו דאטה על 30,000 סופרים- רבים מהם ללא מובהקות סטטיסטית מבחינת כמות החומרים שיצאו לאור. לכן הדאטה עליהם היה דליל ולא מייצג ומשום כך השארנו את הסופרים הפופלרים ביותר בסדר

יורד מתוך הספרים שלא הפכו לסרטים בלבד, כך שבסכימת כל העבודות שלהם מגיעים למספר מצטבר של 1580 ספרים שלא הפכו לספרים.
המידע הבסיסי שהוצאנו בשלב זה: כותרת הספר, הסופר, קישור לעמוד הסופר, הרייטינג של הספר, כמות המצביעים וקישור לעמוד הספר. ספרים רבים היו ללא מדרגים כלל או עם דירוג 0- מחקנו את כלל הספרים הללו מתוך הדאטה.
בנוסף עבור כל אחד מהספרים ניגשנו בנפרד לעמוד הספר והסופר ומהם אספנו את הז'אנרים העיקריים של כל ספר, ופרטים אישיים לגבי מחברי הספרים והוצאת הספרים.

הפיצ'רים

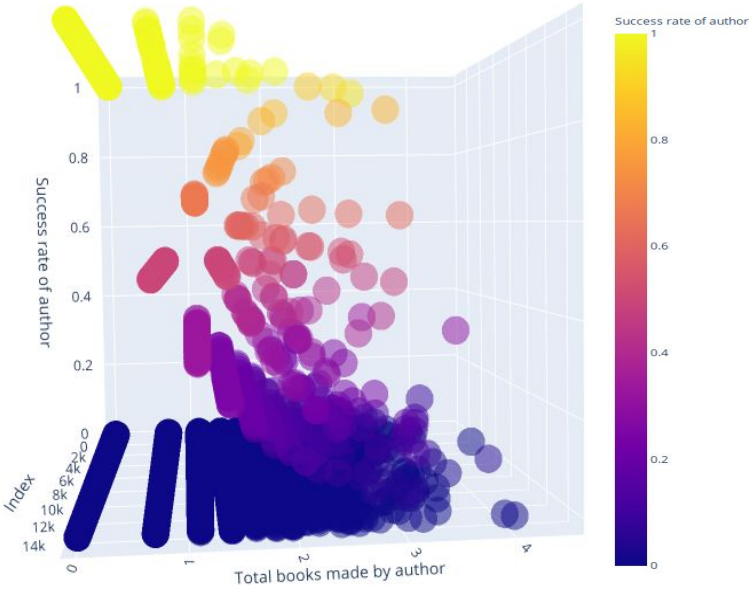
להלן פירוט של הדאטה שאספנו והפיצ'רים שהתקבלו מתוכו, כתלות במסקנות שהסקנו לגבי הדאטה.
בקובץ feature_vector_from_book_ind.py בנינו את וקטור הפיצ'רים שכולל את הפיצ'רים הבאים:

פיצ'רים 1-0: הסופר

לכל ספר יש סופר, וישנם סופרים מצליחים שהספרים שלהם הפכו לסרטים מוצלחים. האם יש סופרים שהספרים שלהם טובים יותר עבור סרטים מסופרים אחרים?
משום שאין משמעות לשם של הסופר חשבנו שהגיוני יהיה להשתמש ב-dummy variables עבור השמות של הסופרים. הדבר יגרור וקטור פיצ'רים עם sparsity גבוה.
אספנו עבור כל ספר מהדאטה את הסופר של הספר. לכל סופר בדקנו כמה ספרים מפרי עטו הפכו לסרטים- כך קיבלנו ערך שברי המעיד על 'הצלחה' של הסופר במובן שהספרים שלו הופכים לסרטים:

$$\frac{\textit{books turned into movies}}{\textit{total books made by author}} = \textit{Author success rate}$$

הגרף הבא מתאר את הקשר בין ההצלחה של סופר לבין כמות הספרים מפרי עטו של הסופר:



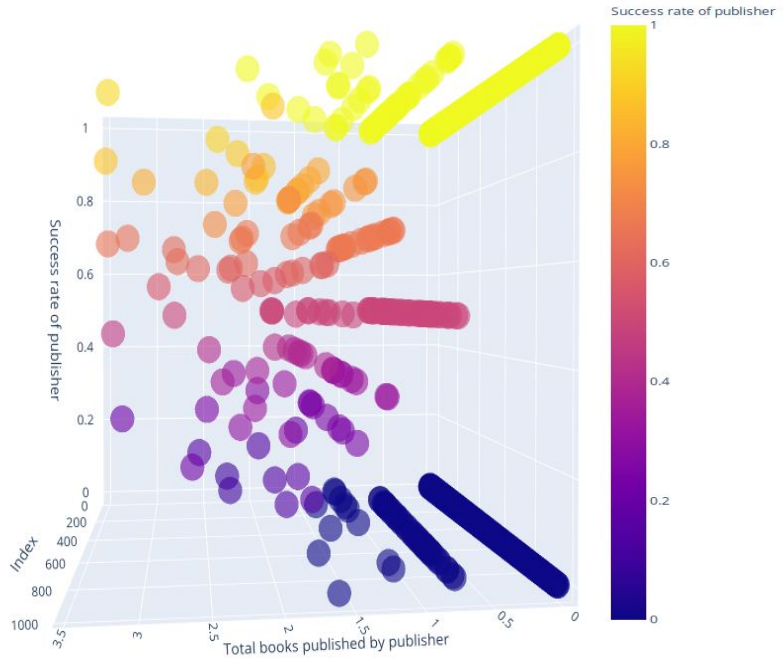
ציר ה-x המציין את מספר הספרים שהסופר הוציא נמדד בסקאלה לוגריתמית. מתוך הגרף ניתן להבחין שמעטים הם הסופרים שספריהם הופכים לספרים בסבירות גבוהה (בצבע צהוב). סופר שיש לו ספר בודד שהפך לסרט- לא בהכרח מעיד על כך שגם הספרים הנוספים שלו יהפו לסרט.(החלק הימני התחתון של הגרף שם יש סופרים עם הרבה ספרים ומעט מהם שהפכו לסרטים). נראה שהסופרים עם ההצלחה המירבית במובן של סרטים פרסמו ספרים בודדים, ובאופן כללי פחות מסופרים אחרים. בנוסף ניתן לראות שסופרים שסרטיהם לא הפכו לסרטים- לא בהכרח מוציאים מעט ספרים. לכן כמות הספרים של הסופר לא מעידה על כך שכולם יהפכו לסרטים.

מתוך כך הסקנו את הפיצ'רים:
פיצ'ר 0#- כמות הספרים של הסופר של הספר
פיצ'ר 1#- אחוזי הצלחה של הסופר.

הפיצ'ר של אחוזי ההצלחה נורמל סביב 0.5: סופרים עם אחוזי הצלחה מעל 0.5 קיבלו ערך חיובי, וסופרים עם אחוזי הצלחה מתחת ל0.5 קיבלו ערך שלילי. בכך ניתנה חשיבות גדולה יותר לסופרים שלפחות מחצית מספריהם הפכו לספרים.

פיצ'רים 2-3: ההוצאה לאור

כל ספר יצא לאור תחת הוצאה לאור אשר ערכה אותו, הייתה אחראית לכריכה, לתמחור ולפרסום של הספר. האם יש הוצאות ספרים שהופכות את הספרים שלהן ליותר אטרקטיביים ולכן טובים יותר כבסיס לסרט? אולי יש הוצאות לאור שמזהות סופרים שנוטים לכתוב ספרים שמתאימים לעיבוד לסרטים?



הוצאה לאור זהו פיצ'ר קטגוריאלי בבסיסו, בדומה לפיצ'ר של הסופרים. גם במקרה זה התלבטנו אם להמיר את השמות בעזרת dummy variables לוקטורים אורתוגונלים, אך בחרנו להשתמש בשיטה דומה לפיצ'ר הקודם במקום זאת, מתוך הדמיון שבהתפלגויות של שני הפיצ'רים.
בגרף האחרון ניתן לראות שישנם מוציאים לאור שרוב הספרים שלהם הופכים לסרטים (הצבעים הצהובים). כמות ספרים גדולה יותר של המוציא לאור לא בהכרח משקפת שכל הספרים הללו יהפכו לסרטים, אך בכל זאת יש הוצאות לאור בולטות יותר לעומת היתר.

מתוך כך הסקנו את הפיצ'רים:
פיצ'ר 2#- כמות הספרים של ההוצאה לאור של הספר
פיצ'ר 3#- אחוזי הצלחה במובן של ספר שהפך לסרט של ההוצאה לאור.

$$\frac{Publisher\ amount\ of\ books\ made\ into\ movies}{Total\ amount\ of\ publisher\ books\ published} = Success\ rate\ of\ publisher$$

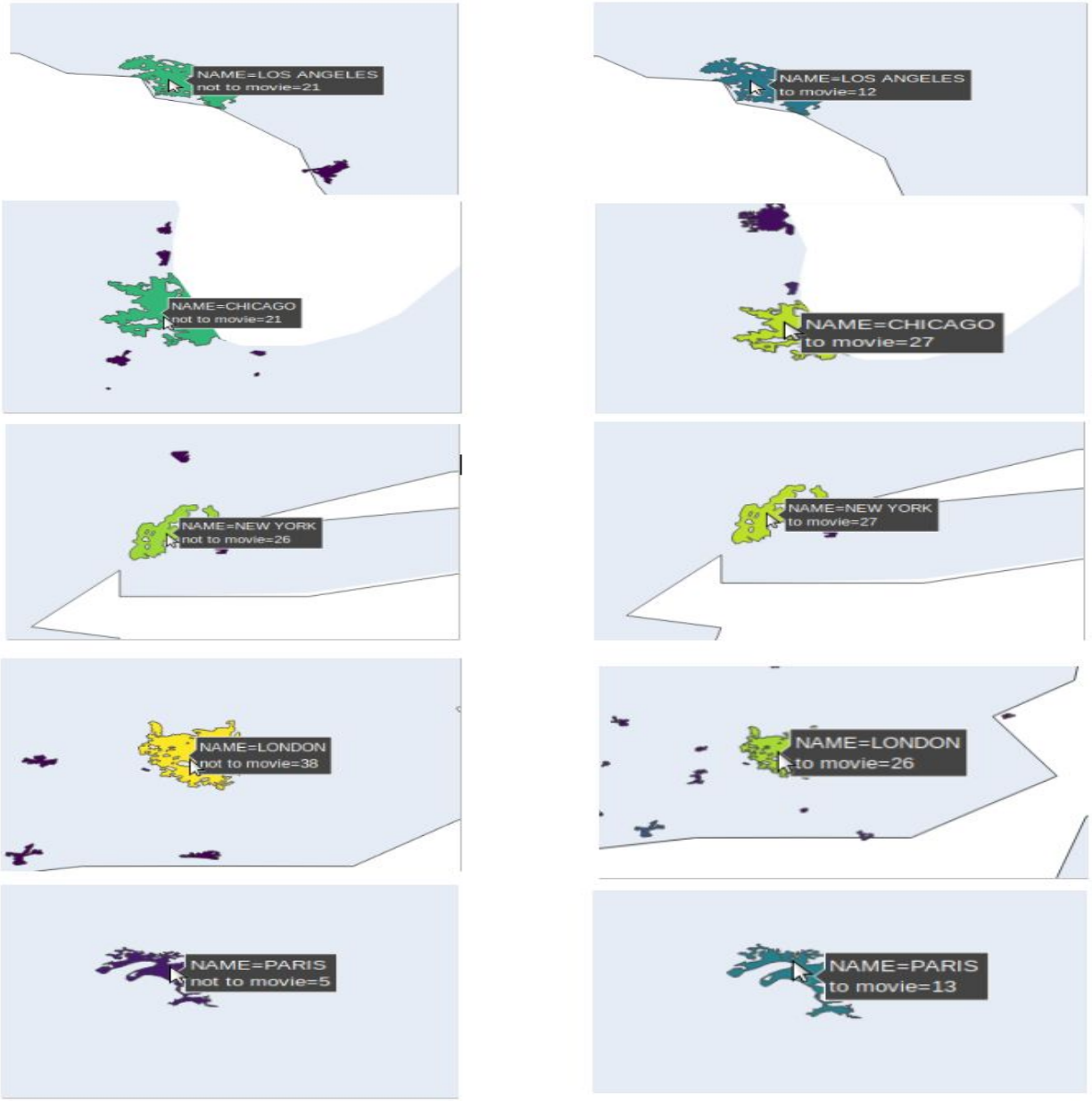
גם כאן הפיצ'ר נורמל בדומה לפיצ'ר של הסופרים.

פיצ'רים 4-5: hometown

האם יש קשר בין עיר שבה עובד הספר לבין הפיכתו לסרט?
תחילה חשדנו שאולי יש ארצות שנוטות יותר מאחרות להפוך ספרים לסרטים (ארה"ב אולי?) בדומה למוציאים לאור ולסופרים. מבחינה מעשית נתקלנו במכשולים שלא הצלחנו להתגבר עליהם, ולכן הסתפקנו בערים בלבד. מעבר לכך הדאטה שאספנו ככה"נ לא מייצג את העולם כולו ומוטה במידה מסויימת.
יש ערים שבהם אין הבדל משמעותי בכמות הספרים בין אלו שהפכו לסרטים ואלו שלא (ניו-יורק למשל).

ערים אחרות דווקא היו עם הבדלים בולטים יותר בין שתי הקטגוריות: לונדון, פריז, לוס אנג'לס, שיקגו.

בגרפים הבאים השוואה בין כמות הספרים שהפכו לסרטים ואלו שלא בין ערים בולטות מתוך הדאטה:



מתוך כך הסקנו את הפיצ'רים:
פיצ'ר 4#- כמות הספרים בעיר של הספר
פיצ'ר 5#- אחוזי הצלחה במובן של ספר שהפך לסרט בעיר של הסרט.

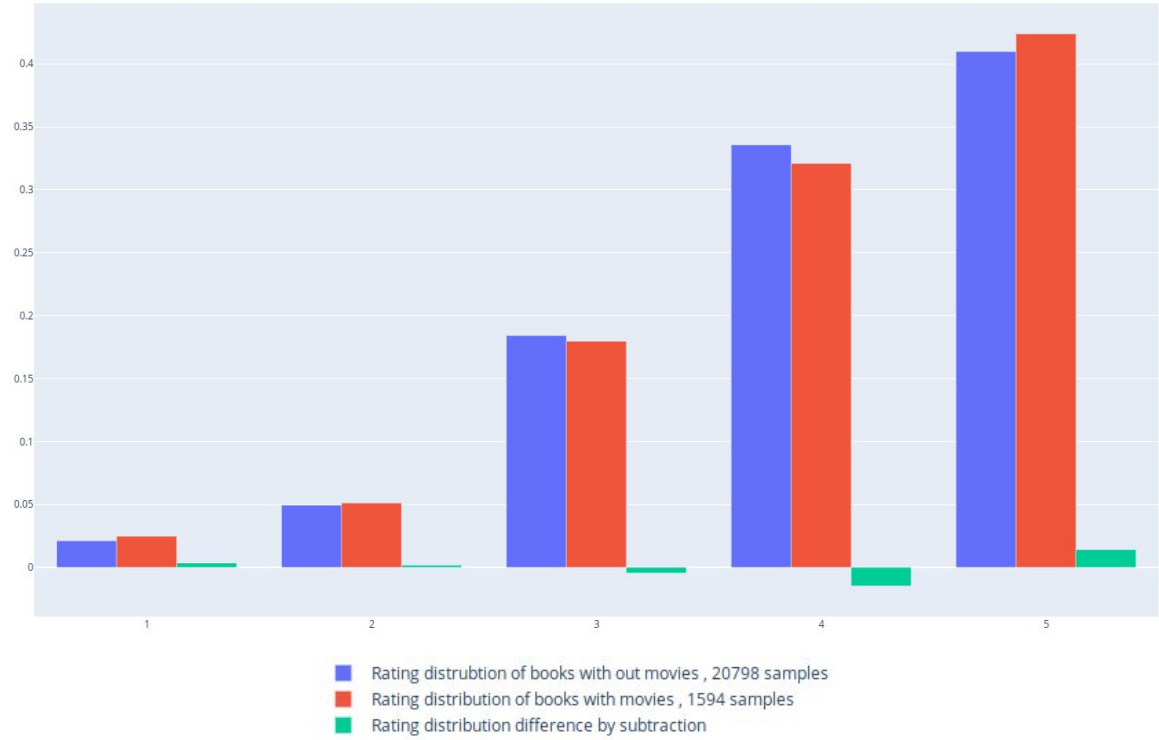
$$\frac{\textit{Book that were made into movie from the current city}}{\textit{total books from the current city}} = \textit{City success rate}$$

Publication date- דאטה שלא השתמשנו בו:
תאריך הפרסום המקורי של הספר היה חסר בעמודים של ספרים של מהדורות שונות. לעיתים תאריך הפרסום היה תאריך המהדורה הנוכחית ולא תאריך פרסום הספר. מצאנו שהדאטה הנ"ל לא מספיק ברור על מנת להסתמך עליו ולהסיק ממנו מסקנות, ולכן לא השתמשנו בו כפיצ'ר.

פיצ'רים 6-7: rating distribution

חלק עיקרי בפופולריות של ספר קשורה במדרגים של הספר. משתמשים שמדרגים את הספר בטווח שבין 1 ל-5. Goodreads סיפק דאטה לגבי כמות הצבעות לכל אחד מהדירוגים הללו. אם כן- שאלה מתבקשת היא האם יש הבדלים מבחינת התפלגות ההצבעות בין הדירוגים השונים? האם יש יותר דירוגים של 5 כוכבים לספרים שהפכו לסרטים? האם יש פחות דירוגים של מעט כוכבים? (2-3 כוכבים).

לשם כך סמכנו את התפלגות ההצבעות במספרים שלמים עבור כל הספרים שהפכו לסרטים ואלה שלא הפכו לסרטים בנפרד. אחר כך ביצענו נרמול לכל וקטור המבטא את הסכימה של כלל ההצבעות, ומכך התקבל מרחב הסתברות המייצג את ההסתברות שמשתמש רנדומלי יצביע 5 כוכבים עבור ספר לסרט, 4 כוכבים, 3 כוכבים, 2 כוכבים ,או כוכב יחיד. הסתברויות הללו נסכמות ל-1 ומגדירות מרחב הסתברות עבור כל אחת מהקבוצות (ספרים עם/בלי סרטים).



בגרף בצבע ירוק מופיע ההבדל בהתפלגות בין שתי הקבוצות. עבור הדירוגים הנמוכים (1 כוכב- 3 כוכבים) ההבדל בהתפלגות זניח. לעומת זאת ניתן לראות כי יש 'חילוף' במסת התפלגות בין הדירוגים 4 ו-5.
ספרים שהפכו לסרטים- ישנו חלק שסביר שיצביעו להם דירוג 5 כוכבים, והוא זהה בגודלו לחלק שככה"נ יצביעו לספרים שלא הפכו לסרטים 4 כוכבים. כלומר לספרים שהופכים לסרטים נוטים להצביע יותר בדירוג 5, ופחות ל4 , ועבור ספרים שלא הפכו לסרטים נוטים להצביע יותר ל4 ופחות ל5!

מידול הפיצ'ר התבצע באופן הבא: בחירת סף לכל אחד מהדירוגים 4 ו-5. ספר עם כמות דירוגים של 5 מעל הסף סביר יותר שיהפוך לסרט. ספר עם דירוגים של 4 מעל הסף- פחות סביר שיהפוך לספר. לכן בחרנו לתת חיזוק לדירוגים מעל הסף של 5, ו'חיזוק הפוך' לדירוגים מעל הסף של 4.
הסף נקבע ע"י מיצוע ההסתברות בין שתי הקבוצות (הפכו לסרט/לא הפכו לסרט). עבור דירוג 5 כוכבים נקבע המשתנה Thershold_dist_5, ועבור דירוג 4 כוכבים נקבע המשתנה Theshold_dist_4.
מתן ציון בוקטור הפיצ'רים התבצע ע"פ הנוסחאות הבאות:

distance_5=book_prob_for_5_stars-threshold_dist_5
distance_4=threshold_dist_5-book_prob_for_4_stars

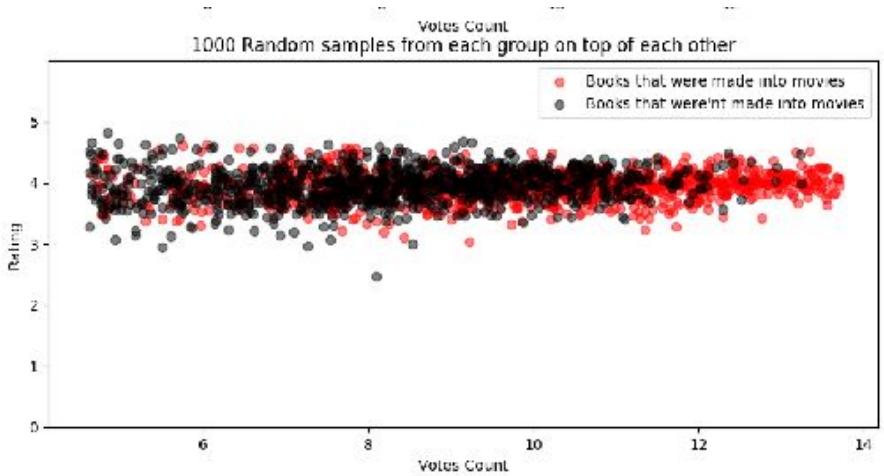
העליונה מחזקת מרחק חיובי מהממוצע (יותר דירוגים מעל הסף), והשנייה מחזקת מרחק שלילי מהממוצע (פחות דירוגים מתחת לסף). שני הערכים מבטאים 'קנס' או 'פרס' בהתאם לכמה הדירוגים של הספר תואמים לאלו שהסקנו מתוך הדאטה.

הערכים עבור הספים שקיבלנו :
0.416 - סף עבור 5 כוכבים, 0.328 - סף עבור 4 כוכבים.

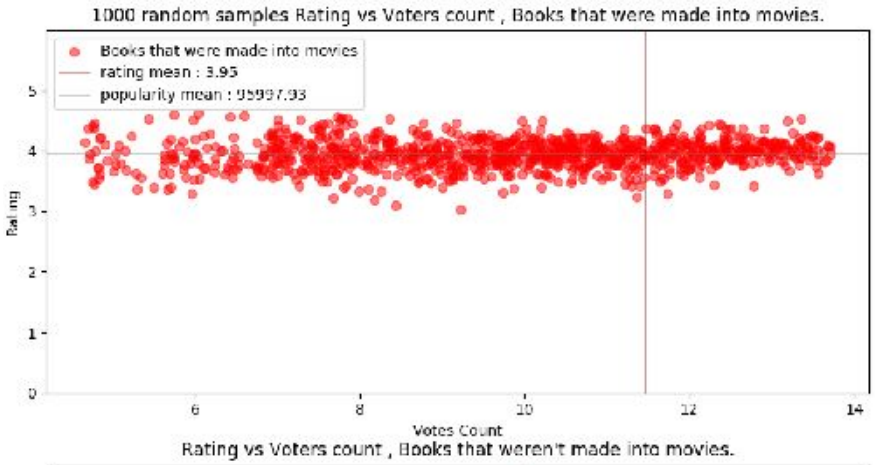
פיצ'ר 8- 'אוריגינליות'- פירוט מורחב מאוחר יותר בדו"ח

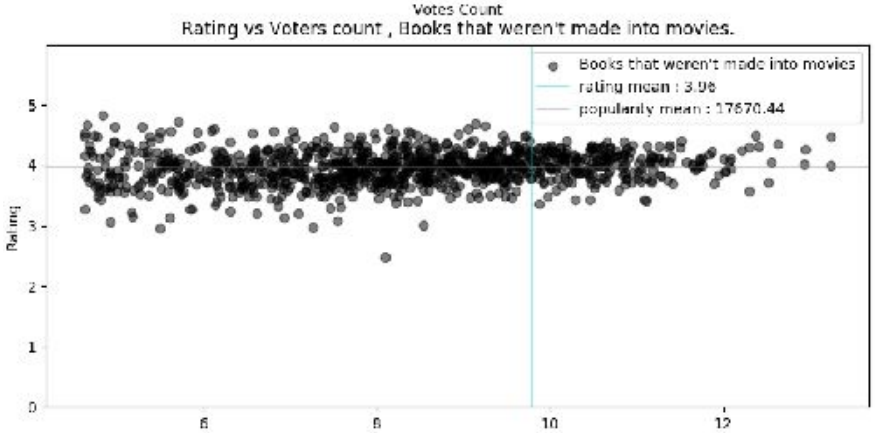
פיצ'ר 9- פופולריות
ספרים מוצלחים נודעו בכך שהפכו לסרטים. האם זה גם הכלל כאשר מסתכלים על התמונה הגדולה ולא על יחידים?
כדי לוודא זאת בחנו את ההבדלים בין ספרים שהפכו לסרטים לאלה שלא כתלות ברייטינג ובפופולריות שלהם.
אינטואיטיבית נראה שיש קשר בין הרייטינג של ספר לבין הסיכויים שלו להפוך לסרט. האם בהכרח ספרים פופולריים הם טובים יותר (מבחינת רייטינג) לעומת ספרים פחות פופולריים?
כדי להבין את הדאטה בצורה ברורה יותר בחרנו רנדומית מדגם של 1,000 ספרים, מתוך אלו שהפכו לספרים וכאלה שלא. בדקנו אכן שדגימת המדגם באופן רנדומלי אינה מעוותת את הדאטה, והגענו למסקנות דומות במספר מדגמים חלקיים שלקחנו. היתרון במדגם החלקי הוא בגרף שניתן להסיק ממנו מסקנות באופן מובהק וברור יותר.
בגרפים מופיע הקשר בין כמות המדרגים של הספר לבין הדירוג הממוצע של הספר, כתלות אם הספר הפך לסרט או לא. הממוצע במספר המדרגים של סרטים שהפכו לסרטים גבוה (בערך פי 5) ממוצע המדרגים בספרים שלא הפכו לסרטים. לעומת זאת הרייטינג בשתי הקבוצות זהה: ~3.95 ללא תלות אם הספר הפך לסרט או לא.
המסקנה: <u>ספרים שהופכים לסרטים נוטים להיות פופולריים יותר!</u>
אבל החל מסף מסויים אין חשיבות לפופולריות של הספר- כלומר אין מקבץ גדול יותר של ספרים משמעותית פופולריים יותר שהפכו לסרטים מיתר הספרים שהפכו לסרטים. בנוסף קיימים outliers בדאטה- ספרים מאוד פופולריים שמטים את הממוצע של שתי הקבוצות. משום שיש ספרים כאלו בשתי הקבוצות הם מטים באופן דומה את שני הממוצעים.
ייתכן גם שספרים שהפכו לסרטים יהיו פופולריים יותר דווקא בגלל הסרט, שיתרום לפרסום הספר. אך באופן דומה סרט שמבוסס על ספר, מקווה להנות מהקהל הקיים של הספר ומהפופולריות שלו, ולכן ככה"נ קיומם של הספר והסרט ביחד מזינים זה את זה.

בגרפים בעמוד הבא: 1,000 דגימות רנדומיות מכל אחת מהקבוצות (ספרים שהפכו לסרטים וכאלו שלא).
הגרף הבא מציג את ההבדלים המפורטים מעלה של שתי הקבוצות יחד:



ויזואליזציה ברורה יותר של כל קבוצה בנפרד:



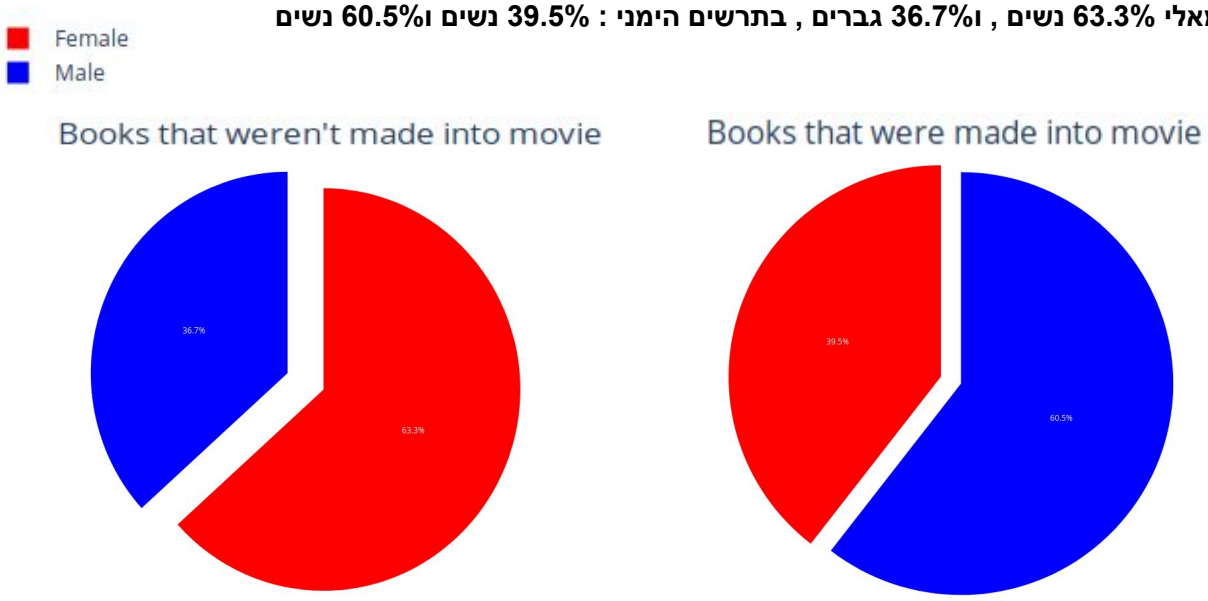


פיצ'רים 10-11: מגדר (male,female) הסופר

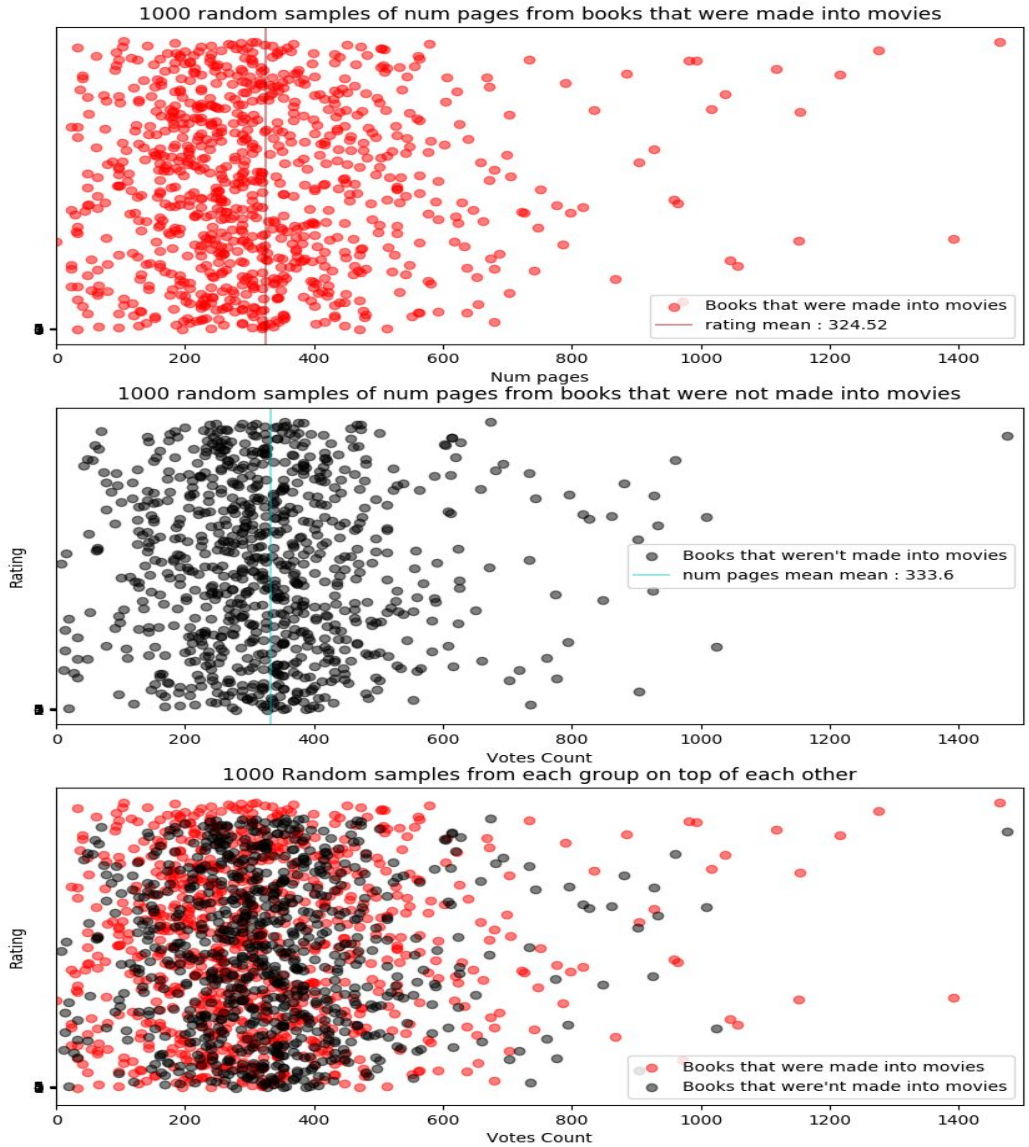
לא פעם ישנה הטייה מגדרית בתחומים שונים סביבנו. רצינו לבדוק אם הטייה זו באה לידי ביטוי גם בסופרים וספרים שהופכים לסרטים. חלקנו הופתעו שיש לפיצ'ר הזה מובהקות סטטיסטית, חלקנו לא הופתעו לראות שהתוצאות נוטות לכיוון המגדר הגברי: יש יותר ספרים שנכתבו ע"י נשים שלא עובדו לסרטים מאשר כאלו שנכתבו ע"י גברים. ולמרות זאת ישנה הטייה ויש יותר ספרים שנכתבו ע"י גברים שהפכו לסרטים! למרות שמלכתחילה היו פחות סרטים כאלו.

הפיצ'ר שיצרנו הינו פיצ'ר אורתוגונלי ב-2 מימדים.

בתרשים השמאלי 63.3% נשים , ו36.7% גברים , בתרשים הימני : 39.5% נשים ו60.5% נשים



דאטה נוסף שלא הוליד פיצ'ר: מספר העמודים בספר
חישוב של ממוצע עמודים בספרים שהפכו לספרים לעומת אלו שלא הפכו לספרים. הממוצעים וההפרשים ביניהם משתנים כתלות בכל קבוצת ספרים רנדומלית בגודל 100 שלקחנו מתוך הדאטה. הסקנו מכך שהשונות גדולה מידי ואין קשר ישיר בין אורך הספר לאם יהפוך לסרט או לא. חוסר המובהקות הסטטיסטית באה לידי ביטוי גם בגרף:



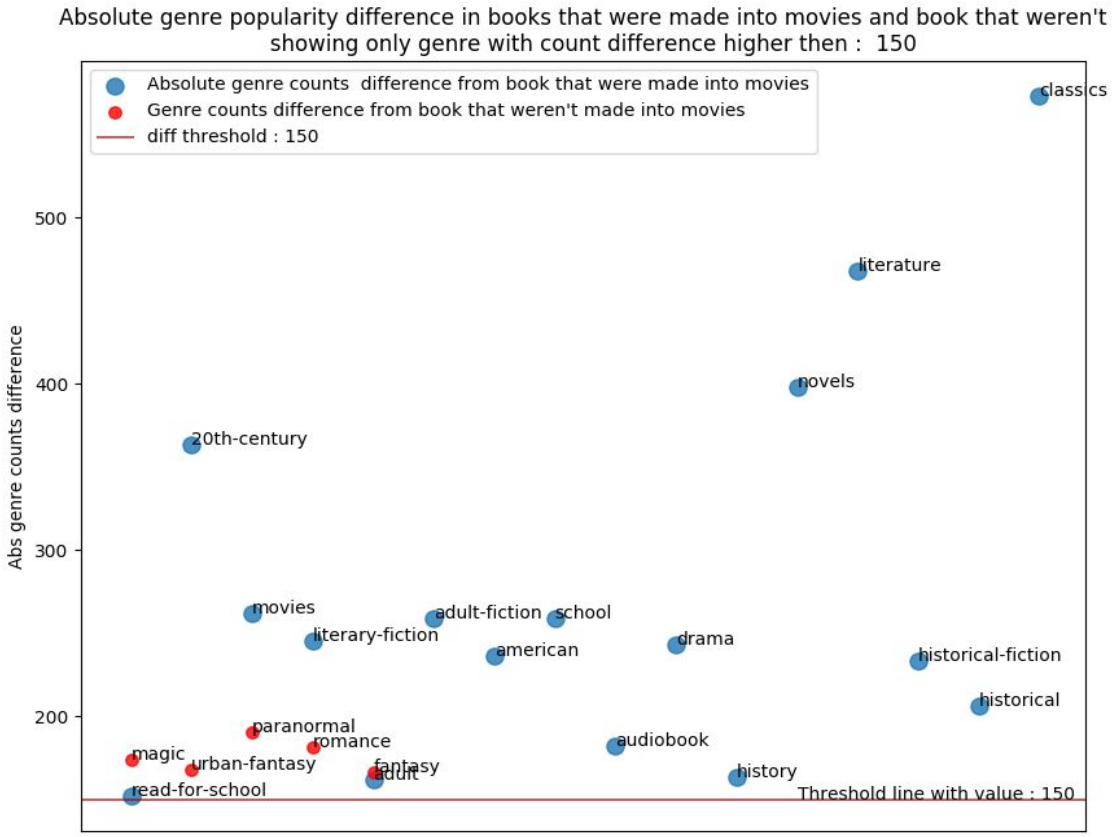
על כן הפיצ'ר לא מצא את דרכו לוקטור הפיצ'רים שלנו.

פיצ'ר 8 (פיצ'ר האורתוגונליות), ו-393 הפיצ'רים האחרונים של הז'אנרים
אינטואיטיבית- הסרטים המפורסמים שמבוססים על ספרים שאנחנו מכירים הם בסגנון דומה ל"הארי פוטר', או ל"משחקי הכס'. האם אכן יש ז'אנרים בולטים המהווים גורם משמעותי בהחלטה לעבד ספר לסרט?

בדקנו מהם הז'אנרים הפופולריים ביותר המופיעים בספרים- שהפכו לסרטים וכאלה שלא הפכו לסרטים. עבור כל ז'אנר החסרנו בין שתי הקבוצות. הערכים שהתקבלו הינם חיוביים או שליליים- בהתאם לקבוצה בה הז'אנרים נפוצים יותר(בסרטים שהפכו לסרטים או כאלה שלא). חיפשנו את הז'אנרים שבהם יש הבדל בין שתי הקבוצות, כך שנוכל לטעון שהז'אנר אופייני יותר עבור ספרים שהפכו לסרטים או עבור כאלו שלא.
אכן יש ז'אנרים המובהקים יחסית לכאן ולכאן- ולכן החלטנו על הוספת הפיצ'ר של הז'אנרים לוקטור הפיצ'רים. משום שהז'אנר קטגוריאלי נדרשנו ליצור dummy-variables, ולכן הגדלנו את וקטור הפיצ'רים במספר הז'אנרים.

בוקטור הפיצ'רים התחשבנו בקבוצת ז'אנרים גדולה: כל ז'אנר שההפרש בו בין שתי הקבוצות גדול מ-50, סה"כ 393 ז'אנרים. הדאטה לגבי הז'אנרים נלקח מתוך Hashtags שgoodreads מספק. כל ספר שייך לכמה מהתיוגים הנ"ל, ע"פ הצבעות של משתמשים לכל תיוג. כלומר המתמשים מחליטים על הז'אנרים של ספר. לכן במובן זה נכון יותר לומר שמידלנו Hashtags נפוצים , הכוללים ביניהם ז'אנרים מסויימים. בנוסף היו מספר תיוגים, כמו למשל movies, שהיטו את וקטור הפיצ'רים (שכן הכילו בתוכם את התיוג, אם ספר הפך לסרט או לא) ולכן הסרנו אותם מהדאטה.

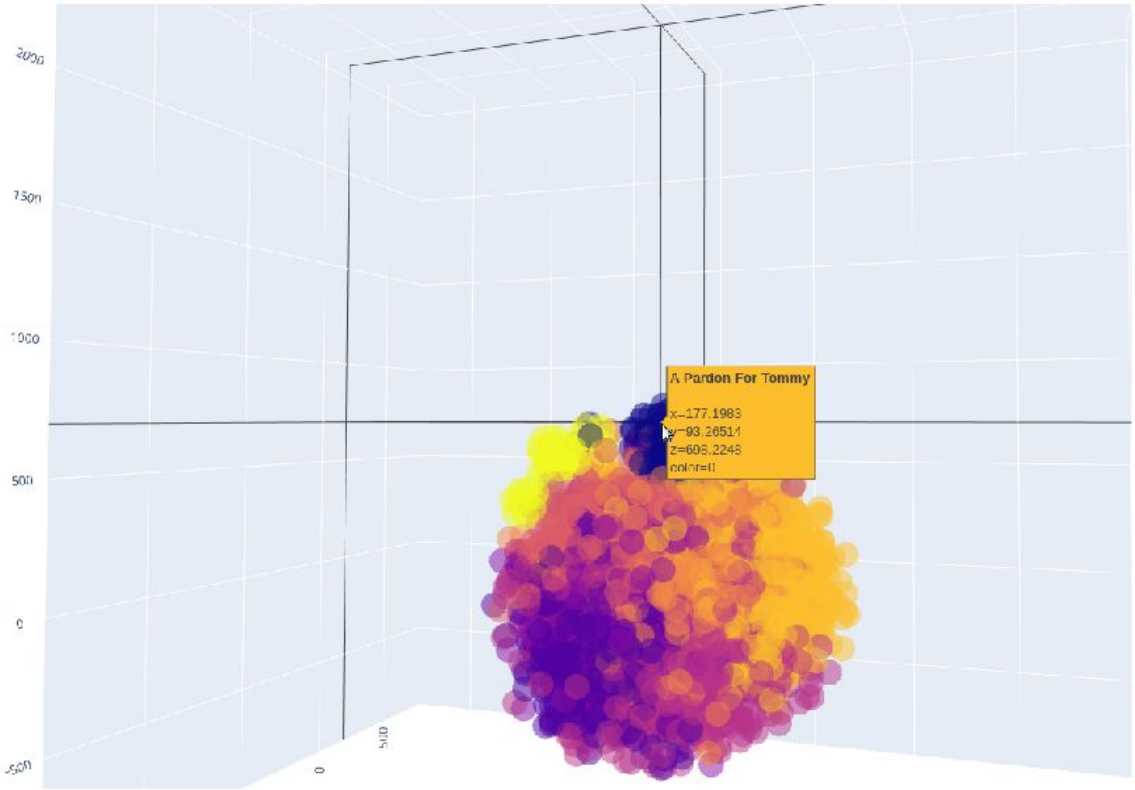
הגרף הבא מייצג קבוצת ז'אנרים קטנה יותר: הז'אנרים המשמעותיים ביותר עם הפרש של 150 בין כמות הספרים בשתי הקבוצות. זאת כדי שהגרף יהיה מובן יותר וימחיש את המסקנות לגבי הדאטה.
ניתן לראות בבירור ז'אנרים שספרים מתוכם נוטים להפוך לסרטים, ולהפך- ז'אנרים שככה"נ יתקבלו פחות ספרים.



כל ספר קיבל תיוג של 1 בוקטור המתאים לז'אנרים שהספר שייך אליהם ע"פ הדאטה.

לכל ספר בgoodreads יש בנוסף summary שמתאר את העלילה של הסרט. קיווינו שמתוך קטע הטקסט הנ"ל נוכל להסיק ז'אנרים נוספים ע"י שימוש ב- plot embedding של הספרים. זאת ע"י ביצוע clustering על הplot embeddings ובדיקה אם ה-clustering מבטא דפוסים נוספים שלא הצלחנו לחשוב עליהם.

מספר בעיות שמנעו מאיתנו לממש רעיון זה:
הראשונה- נדרש להוריד מימד מ- 300 מימדים ל - 3 על מנת להציג כל ספר במרחב תלת מימדי. הורדמת במימד פגמה בדיוק. בעיה נוספת היא במטריקה לפיה מודדים מרחק. מטריקת המרחק פחות מדוייקת במרחב של 300 מימדים. בנוסף משום שהמטריקה האוקלידית עדיפה עבור וקטורים שאינם sparse מאשר מטריקת cosine, בחרנו בה בנסיון לבצע K-Means clustering. התוצאה שהתקבלה חסרת מובהקות סטטיסטית מספקת- או לכל הפחות אנחנו לא הצלחנו למדל מתוך כך פיצ'רים מתאימים נוספים. ובכל זאת התוצאה שהתקבלה ראויה לציון:



הצביר הכתום מכיל רוב של ספרי פחד/ערפדים. הצביר הסגול מכיל רוב של ספרי פנטזיה- זה מצביע על כך שיש קשר בין הספרים. משום הבעיות שציינו קודם- התקבל הרבה רעש בתוך הקלאסטרים. אמנם בתוך כל קלאסטר יש רוב לז'אנר כלשהו (מבדיקה ידנית), אך רוב זה אינו מספיק גדול על מנת להסתמך עליו כפיצ'ר ואינו מספיק ברור.
בנוסף מעניין לשים לב לכך שנוצר צביר הצהוב - אשר מכיל ספרים בשפות שונות מאנגלית, וזוהי תופעת לוואי לכך שהword2vec יוצר word embedding למילים באנגלית בלבד, לכן מתעלם מתוכן שאינו באנגלית ולכן ממפה את כלל הספרים הללו לאותה הקטגוריה- אותו אזור במרחב המייצג ספרים ללא מידע על הplot embedding שלהם.

ואמנם לא השתמשנו בכלי האחרון כדי לבחון התאמה לפי ז'אנרים, אבל מצאנו שאולי זו דרך לבחון עד כמה ספרים 'אוריגינליים'. איך הגענו למסקנה הזו? מחיפושים ברחבי האינטרנט לגבי **מה אנשים מהתחום אומרים על סרטים מבוססי ספרים?**

גילינו ששני הדברים המרכזיים שהופכים ספר לסרט הם:

- כמה קל לבצע את האדפטציה מבחינת תוכן הספר. ספרים בעלי אלמנטים על טבעיים דורשים תקציבים ליצירת אפקטים וסצנות מורכבות יותר שלא זמינות במציאות היום-יומית לעומת סרטי דרמת נעורים למשל.
- מקוריות של סרט. להגדיר מקוריות של סרט זו בעיה קשה ולא טריוויאלית כלל. אין מדד שניתן להסיק ישירות מהנתונים היבשים של פופולריות, התפלגות דירוגים, פרטי הסופר או הפיצ'רים האחרים שמידלנו.

התקווה הייתה לראות את המאפיינים הללו באים לידי ביטוי בדאטה עצמו. לשם כך פנינו לכלים של עיבוד שפה טבעית על מנת להסיק יותר מהתקצירים של כל ספר.

מבחינת העיבוד של הדאטה- בשלב הראשון תרגמנו את התקציר של ספר כך שיהיה בר השוואה לתקצירים של ספרים אחרים. כלומר בנינו מרחב שניתן להטיל עליו תקציר של עלילה של ספר.
בשלב השני בדקנו ייצוג של העלילה במרחב ומדדנו את מרחק האוקלידי מהספר הנוכחי לכל יתר הספרים האחרים במרחב. במידה והמרחק הממוצע של הספרים הקרובים ביותר לספר זה (נניח 10 הקרובים) הוא קטן מיתר הספרים הדאטה, אזי נסיק שהספר לא מקורי ביחס ליתר הספרים בדאטה. אם המרחק הממוצע שלו משאר הספרים גדול יחסית- אזי ככל הנראה העלילה של הספר, תחת הנחת המרחב עליו הטלנו את העלילה, מייצגת עלילה מקורית!

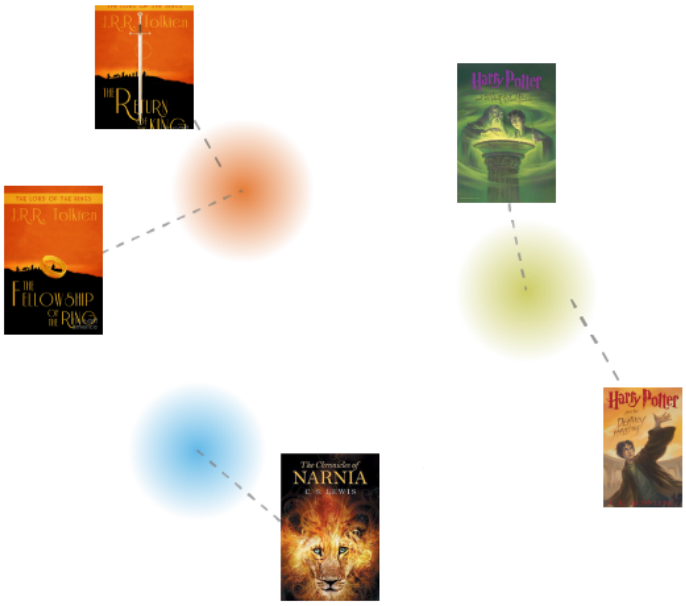
מהיכן נלקחה העלילה?

הטקסט של הספר עבור כל ספר בדאטה סט היה יוצר דאטה בגודל שלא יכולנו להתמודד איתו מבחינת זמנים, גם לו היינו מצליחים לאסוף את הדאטה המתאים. בנוסף לא כל הספרים זמינים בפורמט txt. על מנת להתגבר על כך לקבל שערוך לעלילת הספר הוצאנו את תיאור הספר מהאתר goodreads עבור כל הספרים שנכללים בדאטה שאספנו

"הטלת עלילת הספרים למרחב":

יצירת מרחב עבור כלל העלילות התבצע ע"י אלגוריתם Word2vec. תחילה סיננו את המילים המשמעותיות ביותר מתוך העלילה ובכך צמצמנו את כמות המילים בכל עלילה לכ50 מילים. עבור קבוצות המילים שנוצרו עברו כל ספר, הוצאנו את ה-word embedding , וסכמנו את כל הembeddings. לאחר מכן חילקנו את ה"plot embedding" בכמות המילים שקיימות במילון word2vec- כאלה שיש להם embedding- בשביל למפות ספרים לטווח זהה.

ציפינו שהמרחב שיתקבל יראה דומה לזה שבתמונה להמחשה, אך פחות Sparse.



ייצוג של עלילת ספר ביחס לספרים אחרים במרחב:

בחרנו ספר פופולרי ואהוב (הארי פוטר) ובדקנו אם מטריקת המרחק האוקלידי מייצגת באמת שוני ודמיון בין תיאורי ספרים שונים כאשר מפעילים אותה על Plot embeddings שונים ביחס לספר הנ"ל? קיבלנו תשובה חיובית למדי עם מעט רעש.

בדוגמאות שבדקנו 9/10 הPlot embeddings הקרובים היו באמת ספרים דומים מאוד בתיאורים לספר המקורי , לדוגמא ניתן לראות בתרשים הבא את המרחק מהספרים השונים שבחרנו אל הארי פוטר. וכפי שניתן לראות הספרים הרחוקים מאוד שונים מאוד בתוכן שלהם מ'הארי פוטר'. הספר הכי קרוב אליו הוא ספר אחר מהסדרה של 'הארי פוטר', וזה לא מפתיע! ספרים קרובים נוספים הם עם מאפיינים דומים מעט: ספרי פנטזיה אהובים עם דמויות מורכבות ועולמות חדשים: 'נרניה', 'שר הטבעות'.

תוצאות אלו הביאו לנו אור ירוק להמשיך עם הכיוון וכעת לנסות למדל מקוריות של סרט!

<u>שימוש במרחב על מנת למדל מקוריות כפיצ'ר:</u>

משמצאנו דרך לבחון שוני בין שני ספרים, אנחנו נדרשים להבין מהו סף שהחל ממנו ספר נחשב למקורי. לשם כך חיפשנו סף ממוצע מקוריות עבור ספר בקבוצת האימון. עבור כל ספר נחשב את מרחק מהמקוריות של הספר עצמו אל הסף שנקבע. אם הערך המתקבל הוא שלילי אזי הספר אינו מקורי במיוחד, אם הערך חיובי אזי נסיק שהספר מקורי ביחס לממוצע הספרים המקוריים בדאטה.

<u>חישוב מקוריות עבור ספר יחיד:</u> המרחק של הספר משאר הספרים בדאטה , ולקיחת ממוצע המרחק מעשרת הספרים הקרובים ביותר במרחב אל הנקודה בה הPlot embedding של הספר הנוכחי נמצא.

<u>חישוב סף מקוריות:</u> חישבנו מקוריות כפי שהוגדר למעלה לכל ספר אחר בדאטה. סכמנו את כולם וחילקנו במספר הספרים. הערך שהתקבל הוגדר להיות סף המקוריות. הסף שהתקבל: 0.4292828985483242

מידול הפיצ'ר הסופי של מקוריות:
עבור כל ספר , חישוב ממוצע המרחקים של הPlot embeddings של הספר עצמו מעשרת הספרים הקרובים אליו.
לאחר מכן חיסור המרחק שהתקבל מסף המקוריות והוספת הערך לוקטור הפי'צרים. נקבל ערך חיובי אם הספר מקורי ('פרס') ושלילי אם לא ('קנס'). בכך נאפשר לספרים מקוריים סיכוי יותר גבוה להצלחה מאשר לספרים שאינם מקוריים, בהתאם לצורך של הבמאים.

בניית המודל

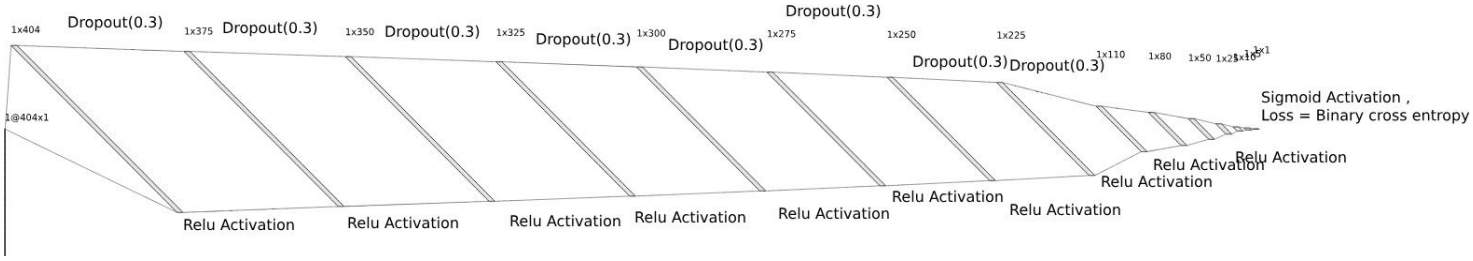
פרדיקציה:
חיברנו את כול הפיצרים לפיצ'ר וקטור יחיד באורך 404. כל עמודה בוקטור מבטאת פיצ'ר שונה שהסקנו לגביו מובהקות סטטיסטית מהדאטה בשביל להשפיע על תוצאות הפרידקציה.

במקום 0 : כמות הספרים הטוטאלית שהוצאה ע"י הסופר
במקום 1 : $(AuthorSuccesRate - 0.5) * 2 * TotalBooks$
במקום 2 : כמו הספרים הטוטאלית שהוצאה ע"י הPublisher
במקום 3: $(PublisherSuccessRate - 0.5) * 2 * TotalBooksPublished$, עקב אותו פירוט כמו שכתוב למעלה עבור
מיקום 4: כמות הספרים הטוטאלית שהוצאה מן הhometown של הסופר
מיקום 5 : $(HometownSuccessRate - .5) * 2 * TotalBooks$ נרמול זהה לסיבות מעל , למטרת פרס וקנס
מיקום 6 : מרחק של ההסתברות להופעה של 5 כוכבים בהתפלגות ההצבעות עבור הספר הנוכחי מן הסף שקבענו.
מיקום 7 : מרחק של ההסתברות להופעה של 4 כוכבים בהתפלגות ההצבעות עבור הספר הנוכחי מן הסף שקבענו.
מיקום 8 : ממוצע מרחק מן עשרת הPlot embeddings הקרובים ביותר (עבור הספר הנוכחי) פחות הסף שקבענו.
מיקום 9 : כמות האנשים שדירגו את הספר הנוכחי
מיקום 10 : האם Author הינו גבר?
מיקום 11 האם Author הינו אישה?
מיקום 12-404 : וקטור אורתוגונלי המעיד אם הספר שייך לז'אנר או לא. מכיל ז'אנרים שהחלטנו לגביהם שהם מובהקים סטטיסטית.

וקטור זה הינה הInput עבור רשת נוירוניים Fully connected.

השתמשנו ב-3160 ספרים- כאשר הדאטה מחולק לחצי חצי עבור ספרים שהפכו ועובר ספרים לא הפכו לסרטים.
סט אימון: 2860 ספרים- כאשר חצי מהם הפכו וסרט לסרטים וחצי מהם לא הפכו לסרטים.
סט טסט: 300 ספרים- כאשר 150 מהם הפכו ו150 מהם לא הפכו לסרטים.

ארכיטקטורה הרשת:



המספרים בשכבות:
1 , 5 , 10, 25 , 50 , 80 , 110 , 225 , 250 , 275 , 300 , 325 , 350 , 375 , 404

פונקציות האקטיבציה בכל שכבה חוץ מן האחרונה הינה Relu, ובשכבה האחרונה הפעלנו Sigmoid על מנת לקבל פרדיקציה הסתברותית בין 0 ל1 , אם ההסתברות הייתה גדולה מ0.5 , קבענו שכדאי להפוך את הספר לסרט , ואם ההסתברות הייתה קטנה מ0.5 קבענו שלא כדאי.

בנוסף בין כול 8 השכבות הראשונות הופעלה שכבת Dropout עם הסתברות 0.3
האופטימייזר : ADAMS
Epochs size : 2500
Batch size : 200

ההגעה לארכיטקטורה זאת הייתה תהליך ארוך להבין מה עובד הכי טוב עבור הדאטה שלנו והבעיה הספציפית.
בעיות שנתקלנו בהם:

1. דיוק נמוך מאוד בשלבים הראשונים: נבע בעיקר עקב יחס הגדול בין הפופולריות לבין שאר הפיצרים - אחוזי הצלחה למשל של סופר היו בין -1 ל1 כפול מספר הספרים שעשה, ערך שיכול להתגמד יחסית לעומת כמות דירוגים- ערך שעשוי להגיע למיליונים. לכן בחרנו להכפיל את שאר הפיצ'רים בקבוע. שיחקנו עם קבוע זה עד שהגענו ליחס האידיאלי שמצליח לבטא אותם בסדר גודל נכון מספיק בשביל לתת לכל אחד מהם את המשקל המתאים לו- הקבוע הוא 700 אחרי המון trail and error.
2. אובר פיט של המודל: תמיד אחוז הדיוק על טסט האימונים היה קטן ב%6~ מאחוס הדיוק על האימון. פתרנו בעיה זו ע"י trail and error של אילו שכבות Dropout כדאי לשים והיכן, ועם איזו הסתברות.
אחרי המון נסיונות גילינו שבעיקר כדאי לשים אותם בשכבות הגדולות והראשונות. כששמנו אותם בשכבות האחרונות המודל לא למד בהסתברות גבוהה בכל אימון.
כמות הDropouts האידיאלית שהגענו אליה (8) יצרה הפרש של פחות מ%1~ בין הTraining לבין הTest
3. דיוק נמוך: זאת משום שהרשת אינה עמוקה מספיק. הדיוק עלה משמעותית כאשר הגדלנו את העומק מ3 שכבות ל15 שכבות , אחרי שמצאנו את הערכים האידאליים לDropout.
4. כמות הEpchos: המודל הראה ביצועים טובים יותר כאשר העלנו את כמות הEpchos לאלפים, במיוחד כאשר הגדלנו את הרשת להיות עמוקה יותר.
5. גודל הבאצ': כאשר לקחנו באצ' קטן מידי המודל הגיע לדיוק נמוך יותר משמעותית באופן מהיר מאוד (חושדים שעקב Overfit אם לומדים על כמות דוגמאות קטנה מידי).

במהלך מציאת הפרמטרים הטובים ביותר השתמשנו ב-Confusion Matrix. זה עזר לנו להבין איפה אנחנו צודקים ואיפה אנחנו טועים וככה לתת למודל חיזוק במקומות מסויימים - לדוגמא ראינו שהמודל צודק יותר בספרים שהפכו לסרטים אך טועה בספרים שלא הפכו לסרטים, ולכן בחרנו להקטין את הפופולריות בפיצ'ר וקטור בחצי מכיוון שאנחנו יודעים שספרים פופולריים נוטים להפוך לסרטים יותר. בכך הורדנו את הפעפוע של ערך זה ברשת, מכיוון שהיא צודקת באופן יחסית טוב לגבי סרטים שהפכו לסרטים. בכך ניסינו להטוות את הרשת כך שתשתפר גם בספרים שלא הפכו לסרטים ע"י כך שניתן לפיצרים נוספים להתבטא בצורה יותר חזקה.

לבסוף הגענו לדיוק של 97.3% על Test Set בגודל 300, המחולק ל150 ספרים שעובדו ו150 ספרים של עובדו .
הדיוק נמדד ע"פ Accuracy, כאשר מספר labels שהוציא המודל תאמו לlabels מהדאטה.

הסתייגות קלה מן התוצאות :
טענו בתחילה שננסה למצוא האם כדאי להפוך ספר לסרט. התוצאות שלנו יותר ממדלות **האם ספר כבר הפך לסרט** , מכיוון שהדאטה שלקחנו על סרטים שהפכו כבר לסרטים אולי מוטה עקב השפעת הסרטים שנוצרו חזרה על הספרים. הגיוני שפרסום הסרט יגדיל על קהל היעד של הספר גם כן, וישפיע על הפופולריות של הספר . מבחינה מעשית התקשינו להשיג דאטה מהימן לגבי Timestamps של ספרים: מתי הם יצאו לאור ומתי הם הפכו לסרטים. יש ספרים שהפכו אפילו מספר פעמים לסרטים- וחלק מהסרטים הללו לא הצליחו! לכן ככל הנראה קיים Bias בData שאנו

מודעים אליו, אך מקווים שבכל זאת הצלחנו להגיע לתוצאות טובות עם הדאטה שעבדנו איתו, ואולי גם עבור ספרים שטרם הפכו לסרטים באמת נקבל דיוק גבוה בעתיד הרחוק שנבדוק את התוצאות של המודל.

מה הפיצ'ר הכי משמעותי עבור הפרידקציה של המודל?

על מנת להבין זאת איפסנו בכל פעם עמודה שונה בפיצ'ר וקטורים של הTest_data- למשל את העמודה של הפופולריות- ובדקנו כמה תוצאות המודל הושפעו מהאיפוס.

הפופולריות העלתה ממצא מפתיע: הפופלריות כלל לא משפיעה על תוצאות המודל, ואפילו הדיוק ירד ל97% מ- 97.3%. למרות שציפינו שהפופולריות תהיה מרכיב עיקרי במודל!

רוב הפיצ'רים הורידו את הדיוק באופן זהה, אך הפיצ'רים שהשפיעו באופן המשמעותי ביותר הם:

- כמות הספרים שיצאו לאור ע"י הPublisher הורידה משמעותית את אחוזי הדיוק ל- 89.6% - מפתיע!
- השינוי המשמעותי ביותר (בפער מיתר הפיצ'רים) שגרם לירידה ל-59.6% באחוזי ההצלחה הוא הפיצ'ר של המכפלה בין כמות הספרים הכוללת שהוציא סופר באחוזי ההצלחה שלו.
- כמות הספרים הכוללת שהוציא הסופר לפני המכפלה באחוז ההצלחה המנורמל - מורידה את הפרידקציה ל 58.3%. כלומר זהו הפיצ'ר המשמעותי ביותר שקבע את הפרידקציה של המודל.

לסיכום, הגורם המשפיע ביותר על הפיכת ספר לסרט ע"פ המודל שלנו הוא בעיקר הסופר עצמו. יתר הפיצ'רים משקטים ומשפרים את המודל במעט סביב פיצ'ר זה.

קשיים במהלך הדרך:

- לקשר בין ספר לסרט המתבסס על הספר. לא מצאנו כלל דאטה עם הנתונים הללו, ונדרשנו להבין כיצד לאסוף דאטה מתוייג נכון, וכיצד לאסוף את יתר הדאטה הנדרש מלבד הנתונים הבסיסיים. התהליך היה ארוך ומורכב, גם מבחינה תכנותית וגם מבחינת הזמן שנדרש כדי לגשת לapi של goodreads.
- בגישה לapi של goodreads קיבלנו גישה רק לדאטה של ספר ספציפי יחיד. המידע החסר לגבי הסרטים לא היה בנמצא, וכשניסינו להשיגו ממקומות אחרים נתקלנו בהרבה רעש וטעויות! את הקישור בין הדאטה שלנו לתיוג של סרטים עשינו זאת ע"י Scraping לרשימות שמשתמשים הרכיבו עבור ספרים שהפכו לסרטים, ועבור ספרים שהיו רוצים שיהפכו לסרטים.
- הapi הגביל את הגישות- גישה אחת כל שנייה. משום כך התקשינו לאסוף דאטה נוסף מזה שמצאנו בעמוד הספר. היינו רוצים אולי למצוא משתמשים בולטים שהתגובות הטובות שלהם מתאימות יותר לספרים שהופכים לסרטים- ובכך למצוא 'חוזי עתיד' ולהשתמש בהם על מנת לנבא על ספרים שכדאי שיהפכו לסרטים. ייתכן ויש מידע נוסף בתגובות של יוזרים שיכולנו להסיק- אך משום הגישה המוגבלת לapi לא יכולנו לעשות זאת. לכן צמצמנו את המאמצים לבניית מודל מספק ומושקע.
- מציאת פיצ'רים משמעותיים עבור הלמידה: לא ברור מלכתחילה מה גורם לספר להיות פוטנציאלית סרט שכדאי ליצור. ניסינו לבחון פיצ'רים אינטואיטיבים שהיינו מצפים שישפיעו, ולמצוא אם גם בדאטה יש התאמה לצפי שלנו.
- בניית הפיצ'רים הייתה כרוכה במציאת מאפיינים, הגדרת ספים ומשחקים למיניהם שדרשו ניסיון וטעייה. כל הפיצ'רים דרשו עיבוד. פיצ'רים קטגוריאלים דרשו שימוש בdummy variables (עבור הז'אנרים). פיצ'רים אחרים דרשו נרמול, מיצוע וחישובים נוספים על מנת להדגיש את ההבדלים והמספרים בעלי החשיבות עבור המודל הסופי. את כל הפיצ'רים השתדלנו להביא לאותם סדרי גודל על מנת שלא יהיו הטיות בדאטה.
- אימון המודל דרש משחקים נוספים וניסיון של להבין מה עובד ומה לא- חלק בפרוייקט שלא ניתן להסיק בו מסקנות חד משמעיות מדאטה גלוי שמוצג בבהירות בגרף , והיה trail & error רב.
- בעולם אידאלי מקביל היינו רוצים לאפשר למשתמש להכניס פרטים יבשים על ספר שטרם הפך לסרט- ולקבל חיזוי לגביו אם הוא יהפוך לסרט או לא. בעולם שלנו הסתפקנו במודל המאומן שהתקבל מהפרוייקט.
- בעולם אידאלי אחר לו יכולנו לבצע אנליזה לעלילה של הספר, אולי להשוות אותה מול תסריטים ולהסיק בכלים של עיבוד שפה טבעית יותר על הטון והשפה שבין המילים בספר- אולי היינו מבינים טוב יותר מה האופי של ספרים שהופכים לסרטים.

לסיכום, היה מעניין ולמדנו הרבה:)
קישור לעמוד הgithub שמכיל את הקוד, הגרפים ויתר השלל:
https://github.cs.huji.ac.il/deven14423/BookBuster
(מכיוון שהדאטה והכל ביחד יותר מ100 מגה!)