דוח מיני פרויקט – יעל יבלונקה ומאיה סגלין

בפרויקט שלנו חיפשנו itemsets מבדילים בין בתי הגידול human ו plant .
לhuman היו במקור 252 טרנזקציות ול plant 74 , על מנת לקבל תוצאות מיטביות לקחנו מhuman רק 74 טרנזקציות. (אם היינו משאירים את זה ללא שינוי ב human היה פי 3 יותר דאטה- ובצורה הזאת לא ניתן להשוות תדירות COG ב human לתדירות בplant)

הבעיה:
כמות המידע עליה עובדים גדולה מאוד ולכן זמני הריצה היו ארוכים במיוחד . בנוסף, בגלל גודל המידע בשלב ה mine של האלגוריתם הרקורסיה הייתה עמוקה מידי ולכן נאלצנו להעלות את ערך ה min_sup ל 350 מה שגרם לתוצאות לא מספיק טובות – קיבלנו item שמופיע בהמון טרנזקציות אך אינו מבדיל.
לאחר מציאת הitem הראשון האלגוריתם מוחק את כל הטרנזקציות בהם הוא מופיע – מכיוון שהרצנו עם ערך min_sup גבוהה כמות הטרנזקציות שנמחקו הייתה גדולה מאוד והאלגוריתם לא הצליח למצוא עוד אייטם שיעמוד בתנאים.

הפתרון:
על מנת להתגבר על בעיית גודל הדאטה החלטנו לחלק את הדאטה ולהריץ את האלגוריתם באופן הבא :
ראשית חילקנו אתה הדאטה על פי בתי הגידול, מכיוון שעדיין מדובר במידע רב חילקנו כל בית גידול למספר קובצות.
הרצנו את החלק הראשון של האלגוריתם (יצירת העץ וה mine) על כל קבוצה בנפרד כך שמכל קבוצה קיבלנו רשימה של .frequent items
לאחר מכן איחדנו את הרשימות הללו לרשימה אחת של frequent items ללא חזרות, מתוך רשימה זו הוצאנו את עשרת האייטמים בעלי ערך הIG הגבוהה ביותר.
אייטמים אלו מופיעים בטבלה בעמוד הבא.
המטרה שלנו הייתה להוריד את ערך הmin_sup על מנת לקבל תוצאות טובות יותר, החלוקה על פי בתי גידול מעלה את הסיכוי שאייטם שנמצא בתדירות גבוהה יהיה גם מבדיל מכיוון שבודקים תדירות שלו רק בבית גידול מסוים בניגוד לאלגוריתם המקורי בו נבדקת תדירות בכל הדאטה (משמע בשני בתי הגידול).
בעזרת השינוי שעשינו הצלחנו להוריד את ערך הminsupp ל24

Distinguishing itemset	IG score	total appearances	Appearances at label 0	Appearances at label 1	% of appearances in label 0	% of appearances in label 1
{'0466'}	0.29135574137309306	86	21	65	28%	87%
{'1799'}	0.2815404407174116	65	54	11	73%	14%
{'0312'}	0.26958719297131706	65	11	54	15%	72%
{'1481'}	0.25175798604785316	69	55	14	75%	18%
{'1316'}	0.2422284810861871	74	52	12	71%	`16%
{'0470'}	0.2381188868896843	93	27	66	36%	89%
{'0547'}	0.21777979241256917	90	26	64	35%	86%
{'0533'}	0.17862562692244788	113	70	43	95%	58%
{'0634'}	0.16624751268053795	105	67	38	91%	51%
{'0508'}	0.16558569617727903	113	44	71	60%	96%

Label 0 – Human; Label 1 – Plant

** ניתן לראות כי כל ה itemset שיצאו לנו הינם יחידונים, יש לציין כי ריצת האלגוריתם אינה דטרמיניסטית מכיוון שהיא תלויה בדרך בניית העץ בפונקציית הmine אשר משתנה מריצה לריצה ובהתאם לכך משתנים הitemsets .
בריצות שונות כן יצאו itemset מעטים שאינם יחידונים.

<u>נתונים מהקובץ COG INFO TABLE עבור כל item שקיבלנו:</u>

COG**0466**;O;CELLULAR PROCESSES AND SIGNALING; Posttranslational modification , protein turnover, chaperones; ATP-dependent Lon protease, bacterial type;

COG**1799**;D;CELLULAR PROCESSES AND SIGNALING; Cell cycle control, cell division, chromosome partitioning;FtsZ-interacting cell division protein YlmF;

COG**0312**;R;POORLY CHARACTERIZED; General function prediction only; Predicted Zn-dependent protease or its inactivated homolog;

COG**1481**;K;INFORMATION STORAGE AND PROCESSING;Transcription;DNA-binding transcriptional regulator WhiA, involved in cell division;

COG**1316**;M;CELLULAR PROCESSES AND SIGNALING; Cell wall/membrane/envelope biogenesis; Anionic cell wall polymer biosynthesis enzyme, LytR-Cps2A-Psr (LCP) family;

COG**0470**;L;INFORMATION STORAGE AND PROCESSING; Replication, recombination and repair; DNA polymerase III, delta prime subunit;

COG**0547**;E;METABOLISM;Amino acid transport and metabolism; Anthranilate phosphoribosyl transferase;

COG**0533**;J;INFORMATION STORAGE AND PROCESSING; Translation, ribosomal structure and biogenesis; tRNA A37 threonylcarbamoyltransferase TsaD;

COG**0634**;F;METABOLISM;Nucleotide transport and metabolism; Hypoxanthine-guanine phosphoribosyl transferase;

COG**0508**;C;METABOLISM;Energy production and conversion; Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide acyltransferase (E2) component;

נביט בitem {1316} – ע"פ הקובץ TABLE COG INFO :

COG**1316**;M;CELLULAR PROCESSES AND SIGNALING;Cell wall/membrane/envelope biogenesis;Anionic cell wall polymer biosynthesis enzyme, LytR-Cps2A-Psr (LCP) family;

נשתמש במקורות **5** ו- **6** על מנת להבין את תפקידם של האנזימים מסוג LytR-CpsA-Psr (LCP) ולנתח את ההבדל במספר ההופעות של cog זה בטרנזקציות בסביבות השונות.

סוג האנזימים LytR-CpsA-Psr (LCP) משתתפים בבניית דופן החיידק, יש להם תפקיד בשמירה על מעטפת תאי חיידקים גראם חיוביים והשפעתם על גורמי ארסיות שונים וכן עמידות לאנטיביוטיקה של פתוגנים אנושיים . הם מעבירים בדרך כלל את הקצה המפחית של גליקופולימרים של דופן התא (CWGPs) של חיידקים גראם חיוביים מחומר ביניים של CWGP הנושא ליפידים, לעמוד השדרה פפטידוגליקן (PGN), בדרך כלל באמצעות קישור פוספודיסטר.
ראשי התיבות "LCP" נובעים משלושה חלבונים שזוהו בתחילה כמכילים תחום LytR - LytR (מדכא ליטי, כיום TagU5), CpsA (מווסת ביטוי רב-סוכר כמוסה), ו- Psr (מדכא סינתזה PBP 5).

ניתן לראות ש cog זה מופיע ב16% מהטרנזקציות שבדקנו בצמחים לעומת 71% אצל בני האדם – כלומר נמצאו יותר נציגים של חיידקי גראם חיוביים בבני אדם מאשר בצמחים, זאת בגלל הסביבה האנאירובית שחיידקי הגראם החיוביים צריכים על מנת להתקיים. בבני אדם ניתן למצוא סביבה אנאירובית גדולה יותר מאשר בצמחים (מערכת העיכול – המעיים) ולכן נמצאו יותר נציגים בבני אדם.

מקורות שהסתמכנו עליהם בכתיבת דו"ח זה:

1. https://en.wikipedia.org/wiki/LCP_family
2. https://www.ncbi.nlm.nih.gov/books/NBK26824/#:~:text=The%20central%20components%20of%20the,with%20regulatory%20subunits%20called%20cyclins.&text=Thus%2C%20activation%20of%20S%2Dphase,cyclin%2DCdk%20complexes%20triggers%20mitosis.
3. http://networks.systemsbiology.net/function/35631/
4. http://www.sbg.bio.ic.ac.uk/~phunkee/html/old/COG_classes.html
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7831098/
6. https://www.technologynetworks.com/immunology/articles/gram-positive-vs-gram-negative-323007#:~:text=Gram%20positive%20bacteria%20have%20a,have%20an%20outer%20lipid%20membrane

הבדלים בזמני הריצה כתלות בערך min supp: