ב"ה

**פרקטיקום במדעי הנתונים - השערות מחקר ותוכנית עבודה כללית**

1. הפרויקט- הסקת מסקנות מתוך מאגר מתכונים חברתי.

מספר רשומות- 39,040

מאפיינים- עבור כל מתכון קיימים הנתונים הבאים:

* דירוג ממוצע והתפלגות הציונים שקיבל בקרב הגולשים (1-5)
* שם המתכון
* תאריך פרסום המתכון
* תיאור המתכון
* הוראות הכנה
* מצרכים
* ערכים תזונתיים
* קטגוריה ותת קטגוריה
* זמן הכנה

מידע חסר: מאפיינים אישיים של מדרגי המתכונים (גיל, מין, ארץ מוצא...)

1. משתתפים:

אוהד פישר – 312166655

אריאל זייד – 307967901

גל עוז – 204455067

נועה ברוכי – 312423338

1. השערות מחקר:
2. אנשים שדירגו או שאהבו מספר מסוים של מתכונים זהים, בסבירות גבוה יאהבו מתכונים נוספים זהים.

דרך פעולה אפשרית: שימוש באלגוריתם collaborative filtering על מנת לחזות את המתכונים שמשתמש מסוים יאהב בסבירות גבוהה לפי דמיון למשתמשים אחרים.

1. מתכונים שזמן הכנתם קצר (עד 20 דק') ידורגו גבוה יותר ממתכונים שזמן הכנתם ארוך (מעל 5 שעות).

דרך פעולה אפשרית: חלוקת המתכונים לשתי קבוצות. חישוב ממוצע הדירוגים עבור כל קבוצה והשוואה ביניהם.---חלוקה לקלסטרים, לפי ציר אחד.

הכנו שתי קבוצות של מתכונים כאשר אחת היא זמן קטן מ20 דק ואחת זמן גדול מ700 דק וערכי הרשימה מורכב מדירוג הגולשים של כל מתכון. לכל קבוצה נבדק ממוצע הדירוגים ונמצא שההשערה הופרכה, ממוצע זמן קטן עמד על 4.33 וממוצע זמן גדול עמד על 4.38 כך שדירוג המתכונים עם זמן הכנה ארוך דורג גבוהה יותר.

מסקנות אפשריות: הדירוג אינו מושפע ממשך הכנת המתכון, או סטטיסטית שהממוצע מתפלג נורמאלי תמיד ללא קשר למשך זמן הכנה.

\*\*אולי שווה לבדוק לפי כמות הגולשים שדירגו הכי גבוהה.

**\*\*\*לעשות בדיקת נורמאליות לנתונים, בדיקות אי מובהקות ושונות לנתונים.**

1. קיים קשר חיובי בין מספר המתכונים שמשתמש דירג לבין רמת הקושי הממוצעת של המתכונים שדירג, כך שככל שהמשתמש דירג יותר מתכונים, רמת הקושי הממוצעת של המתכונים שדירג תהיה גבוהה יותר.

----נדרש מזהה ליוזר ומתכונים שהוא דירג.

דרך פעולה אפשרית: נשמח להכוונה. בדיקת אזור מתכונים מסוים אמצע, לבצע רגרסיה.

כלומר להסתכל על מתכונים עם זמן הכנה סביר לדוגמא בין 15 דק' ל4 שעות. כדי שמתכונים קיצוניים לא יפריעו.

1. מתכונים המשתייכים לקטגורית האוכל האסייתי יכילו יותר מרכיבים בממוצע מאשר מתכונים המשתייכים לקטגוריית האוכל האירופאי.

הצעה חלופית להשערה זו: בהינתן תיאור הרכיבים או המתכון קלסיפקציה של המתכון, מטקסט לייצוג ומייצוג לקלסיפקציה

דרך פעולה אפשרית: חישוב ממוצע המרכיבים בכל אחת מהקטגוריות והשוואה.

1. מתכונים המשתייכים לקטגוריה מסוימת יכילו בממוצע מעל 15% מצרכים דומים.

דרך פעולה אפשרית: בהינתן קטגוריה מסוימת, נחלץ את רשימת המצרכים, נשתמש באלגוריתם tf-idf על מנת לסנן מילות קישור ומצרכים נפוצים (מלח, פלפל...). לאחר מכן, נבצע ספירה של המצרכים שחוזרים על עצמם, נחלק במספר המתכונים. לבסוף נחשב את הממוצע של האחוז שהתקבל.

1. ביצוע svd ושימוש בפירוק של המשתמשים לביצוע קלסטר על המשתמשים, סוג של איפיון המשתמשים.
2. ז.קשר בין זמן מתכון ושיוך המתכון למועד.

מבחינת תהליך עבודה עדיף להתחיל בדיקות סטטיסטיות קלות על הנתונים ואפילו קצת קלסטרים על הנתונים כדי לקבל אינפורמציה עליהם.