**תרגיל בית 1 – הגורם האנושי**

**מאת: גיל קפלן, יעקב פרידמן והלל טריפ**

עשינו סריקה על אתר בשם "all recipe" - עם הקישור הבא: <https://www.allrecipes.com/>. זה אתר שבו אנשים מעלים מתכונים ואנשים מבקרים ומדרגים אותם. לקחנו מידע על מתכונים שונים. יצרנו קובץ נתונים לאחר שסרקנו נתונים מהאתר בעזרת פייתון וספריות שיש בו. בפרט, הנה בשדות שלקחנו:

url, title, rating, num\_rating, num\_reviews, num\_photos, ingredients\_cnt

תיאור הפיצ’רים: קישור לעמוד של המתכון, כותרת, דירוג המתכון, מספר הדירוגים עבור המתכון, כמות הביקורות שאנשים כתבו על אותו מתכון, כמות התמונות עבור אותו מתכון באתר, כמות המרכיבים עבור אותו מתכון.

1. תהליך האיסוף: לקחנו כמה לינקים לאתר שכל אחד הוא לעמוד של קטגוריה אחרת של מתכוננים. בכל עמוד לקחנו את הקישורים למתכונים ספציפיים (לכל מתכון עמוד משלנו). לאחר מכן, עברנו על כל עמוד של מתכון שונה עד שהצלחנו לאסוף מידע מתאים לשדות למעלה עבור 100 רשומות עם ה-7 משתנים שבחרנו.

המקור של הנתונים באתר זה מגיע מאנשים שמעלים מתכונים באתר הזו.

1. בדקנו שמותר לנו לסרוק עם הקישור הבא - <https://www.allrecipes.com/robots.txt>

ראינו שעבור הקישורים שרצינו לסרוק שלא היה רשום disallow ובפרט הבנו שיש לנו אישור לסרוק מאותם עמודים.