אחזור מידע באינטרנט- תרגיל 3

לאה אורלין -[redacted]
בר אלוני-[redacted]

<u>הסבר לפונקציית ה ProductSearch:</u>

1. **מיון ראשוני:** בשלב הראשון ערכנו מיון של הביקורות הרלוונטיות ביותר לפי 2 הפונקציות שממישנו בתחילת התרגיל- vector space ו- language model. החזרנו מכל מודל את 2k הביקורות הרלוונטיות ביותר.
2. **חישוב הציון לכל מוצר (בכ"א מהמודלים בנפרד):** כעת, עברנו על הביקורות שקיבלנו מכל מודל ומכל ביקורת חילצנו את מזהה המוצר עליו נכתבה הביקורת. בנוסף לכך, חילצנו עבור כל ביקורת את הנתונים הבאים שחשבנו שהם רלוונטים להמחשה של מידת שביעות הרצון מהמוצר: score, ולהבנה של הדיוק שהביקורת מספקת : helpfulness numerator, helpfulness denominator.

חישבנו את הציון של כל מוצר בצורה הבאה:
כל מוצר יקבל את הציון שניתן לו ע"י המבקר (score). אך יכול לקבל גם "בונוס" בתנאי הבא:

a. אם אחוז ניכר של המבקרים ציינו את הביקורת כ-helpful:
   i. אם הציון הוא גבוה, ואנשים ציינו את הביקורת כיעילה - נרצה שהמוצר יקבל ציון עוד יותר גבוה, לכן הציון הסופי יחושב בצורה הבאה: $s = s + s \cdot (n/d)$, כאשר:
      1. s : score
      2. n : helpfulness numerator
      3. d : helpfulness denominator
   ii. אם הציון הוא נמוך, ואנשים ציינו את הביקורת כיעילה, כנראה שהמוצר אכן ראוי לציון נמוך ולכן נרצה לתת למוצר ציון שקרוב מאוד לסכום המקורי. לכן הציון הסופי יחושב בצורה הבאה: $s = s + s \cdot (1 - n/d)$ עבור אותם ערכי s,n,d.
   iii. נבחין שבחרנו לפרש את ה-score כמשתנה בינארי:
      לביקורות בעלות ציון (score) גבוה:
      1. אם הן סומנו כמועילות נוסיף להן בונוס יחסית גדול.
      2. אם הן סומנו כפחות מועילות - נוסיף להן בונוס קטן יחסית.
      לביקורות בעלות ציון (score) נמוך:
      1. אם הן סומנו כמועילות - נוסיף להן בונוס קטן יחסית (נפרש זאת כסימן שהמוצר אכן שווה את הציון שניתן לו)
      2. אם הן סומנו כפחות מועילות - נפרש זאת כמחאה על הציון הנמוך שניתן למוצר שלא בצדק ונוסיף להן בונוס יחסית גדול.

b. נשים לב גם כן, שככל שהמוצר הגיע מביקורת שנחשבת לפחות רלוונטית במיון בשלב הראשון, כך נרצה לתת למוצר שלה ציון נמוך יותר באופן יחסי, לכן הוספנו משקל לציון של המוצר הנסקר ביביקורת לפי מיקום הביקורת בשלב המיון הראשוני.

3. **מיזוג ציוני המוצרים:** כעת מיזגנו את הציונים שקיבלנו לכל מוצר, לפי 2 שיטות ה-review search שציינו לעיל (vector space ו language model).מהרשימה הממוזגת שאבנו את K המוצרים בעלי הציון הגבוה ביותר, וכך קיבלנו מ**הביקורות הרלוונטיות ביותר** את המוצרים בעלי **הציון הגבוה ביותר**.