פתרון שאלה 1:

הקדמה:

באלגוריתם Rocchio, הרעיון המרכזי הוא למדוד את הדמיון בין השאילתה החדשה לבין מרכזי הכובד (centroid) של הקטגוריות השונות. השאילתה משויכת לקטגוריה שמרכז הכובד שלה הוא הקרוב ביותר לשאילתה במרחק קוסינוסי:

$$\vec{q}_{opt} = \arg\max_{\vec{q}} [\cos(\vec{q}, \vec{\mu}(C_r)) - \cos(\vec{q}, \vec{\mu}(C_{nr}))]$$

(נציין שיש מרחקים נוספים כמו אוקלידי, מנהטן, jaccard ועוד, תלוי במטרת בסוג הסיווג שמעניין אותנו)

בחישוב רב מימדי אנו מחשבים את קוסינוס הזווית שבין השאילתה לבין וקטור ה-centroid של המסמכים הרלוונטיים עבור הקטגוריה הנבדקת (הוקטור הזה הוא וקטור התכונות [לפי גודל המימד] כאשר כל תכונה היא ממוצע רכיב אותה התכונה בין כל המסמכים הרלוונטיים) ולאחר מכן אנחנו מחסירים את התוצאה הנ"ל בקוסינוס הזווית שבין השאילתה לוקטור ה-centroid של המסמכים הלא רלוונטיים עבור הקטגוריה הנבדקת. לאחר חישוב זה עבור כל הקטגוריות הקיימות אנחנו נשייך את השאילתה למרכז הכובד הקרוב ביותר לשאילתה.

ניגש לפתרון:

על מנת להמחיש טעות בסיווג ניעזר ברמז המנחה אותנו להמחיש את הבעיה באמצעות מקרה חד מימדי (כלומר, במקרה שבו יש רק תכונה אחת, כך שאין לנו ווקטור לחישוב זווית הקוסינוס בינו לבין המסמכים הרלוונטיים והלא רלוונטיים ולכן החישוב הוא ממוצע רכיב התכונה עבור כל מסמך ומסמך).

(לפשטות ההמחשה ניקח מקרה פרטי שבו יש 3 מסמכים רלוונטיים ו-3 מסמכים לא רלוונטיים לכל קטגוריה)

טעות עשויה להתרחש כאשר מרכז הכובד של קטגוריה מסוימת קרוב באותה המידה למרכז הכובד של קטגוריה אחרת, או במקרה הקיצוני יותר (המקרה הרגיל) שבו ישנם כמה מימדים, יכול להיות אפילו שהמסמך המיועד לסיווג יהיה קרוב יותר למרכז הכובד של הקטגוריה הלא נכונה.

מקרה כזה קל להמחיש באמצעות בדיקת מקרה חד מימדי (שיש בו רק תכונה אחת).

נניח מרחב חד-ממדי שבו ישנן שתי קטגוריות:

קטגוריה A מכילה את המסמכים: {1,2,3}

קטגוריה B מכילה את המסמכים: {7,8,9}

חישוב מרכזי הכובד:

- מרכז כובד של קטגוריה A:
  $(1 + 2 + 3)/2 = 6$
- מרכז כובד של קטגוריה B:
  $(7 + 8 + 9)/3 = 8$

כעת, נניח שיש לנו מסמך x=5 שאנחנו מעוניינים לסווג לקטגוריה.

- מרחק לקטגוריה A:

  $|5 - 2| = 3$

- מרחק לקטגוריה B:

  $|5 - 8| = 3$

במקרה זה, Rocchio עשוי לטעות בתוצאה מכיוון שהמרחקים זהים (ההפרש בין המסמכים הרלוונטיים ללא רלוונטיים בכל קטגוריה שווה ל-0) ולכן הוא יבחר בקטגוריה כלשהי באופן אקראי ומשום כך הוא עלול לבחור בקטגוריה הלא נכונה.

מהדוגמה ניתן גם להבין שהבעיה רק הולכת וגדלה במקרים רב מימדיים.