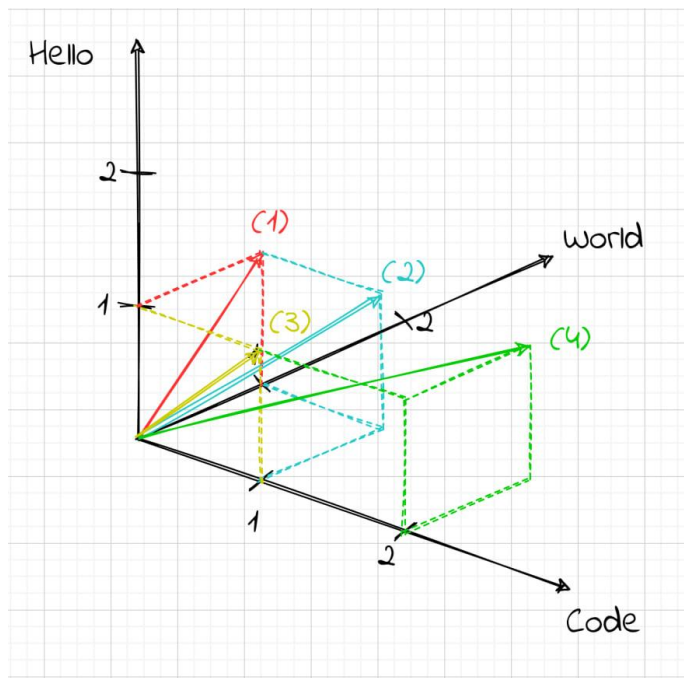


# מבוא לאיחזור מידע

מעבדה 3 - ווקטורизציה של מסמכים



# מטלות המעבדה

בהמשך למבני הנתונים שנבנו במעבדה הקודמת.

1. יש לבנות מטריצת ייצוג ל BOW - יצירת count vector עבור כל מסמך וכלל המילון.
2. יש לבנות פונקציה לחישוב tf-idf למושג יחיד (על פי הנלמד בכיתה) בעזרת ספריות סטנדרטיות ופנדס.
3. יש לבנות מטריצת ייצוג ל tf-idf.
4. יש לממש פונקציית חישוב דמיון - הנותנת ציון למידת הקירבה בין שאילתה ועבור כל מסמך (בהתאם לזווית) cosine sim