פרויקט סיום קורס למידה חישובית וכריית מידע

הפרויקט עוסק בסיווג טקסטים לקהילות רפואיות.

מאגר הנתונים נבנה על ידי והוא תוצאה של scraping על רשת חברתית בנושא בריאות בשם "כמוני". זוהי רשת הבנויה מ-41 קהילות העוסקות בנושאים רפואיים שונים. המטרה בפרויקט הייתה לקבל פוסטים ותגובות מהאתר ולסווג בעזרת סוגים שונים של מודלים שנלמדו לאורך הסמסטר.
הרעיון המיוחד הוא להשתמש בטקסט עצמו כפיצ'רים ולא ב-Metadata, וזה דורש ניסוי וטעיה כדי לבחור את מספר הפיצ'רים האופטימלי לכל מסווג. זה נבחר על ידי חישוב מדד tf-idf לכל המילים בקורפוס ובחירת X המילים עם הדירוג הגבוה ביותר.
כמובן שבוצעה גם parameter tuning כדי לטייב את המודלים.

מבחינת מבנה הפרויקט –

- בתיקיה הראשית יש 5 מחברות שהן כל הפרויקט במלואו, הקבצים שניתן להריץ מחדש ולקבל את התוצאות שהוצגו.
- בתיקיה data יש קובץ מסד הנתונים של האתר כולו וכן קובץ של stopwords בעברית.
- התיקיה notebooks for presentation מכילה את אותן מחברות, אבל מטעמי נוחות מחקתי שם הרבה מהפלטים ולפעמים חלק מתאי הקוד, כדי להציג בכיתה בצורה מתומצתת ולהתמקד במסקנות היחודיות לכל מחברת. יש סיכוי שהיא לא תרוץ כי מחקתי תאי קוד ואולי היו שם הגדרות שנעלמו, לכן כדי להתרשם מהפרויקט ומסקנותיו זו תיקייה טובה מאוד, וזה גם מה שהצגתי בכיתה וניתן לראות בהקלטה. אבל כדי להריץ ולשחזר צריך את המחברות המקוריות.

כל המחברות מבוססות על אותה מתודולוגיה – השוואת כל המסווגים, מספר הפיצ'רים, וטיוב הפרמטרים. הן רק משתנות מבחינת החלק מה-dataset שהן בודקות. להלן ההבדל:

1) המחברת הראשונה מכילה השוואת כל השיטות ביחס ל-2 קהילות רפואיות קטנות ומאוזנות. לכאורה המשימה הכי פשוטה בפרויקט.
2) המחברת השניה מכילה השוואה עבור 2 קהילות קטנות אך לא מאוזנות (אחת יותר תדירה).
3) המחברת השלישית מנסה לטייב את תוצאות מחברת 2 על ידי איזון הקהילות בעזרת oversampling.
4) המחברת הרביעית מנסה לטייב את תוצאות מחברת 2 על ידי איזון הקהילות בעזרת undersampling.
5) המחברת השניה מכילה השוואה עבור 6 קהילות גדולות יותר ולא מאוזנות – לכאורה המשימה הכי קשה בפרויקט.

בסוף כל מחברת יש סיכום ומסקנות של תהליך ההרצה במחברת, ובסוף מחברת 5 סיכום ומסקנות של הפרויקט כולו.

תודה רבה על כל הסמסטר, למדתי ונהניתי מאוד.

רון קינן 203735857