Analytics of big data - Targil 3
הגשה עד 4.8.19

בחרו לכם מאגר נתונים של חוות דעת (review) על מוצר מסוים (מסעדות, סרטים וכדומה, אפשר למצוא כזה ב Kaggle).

עליכם לבנות מודל קיבוץ (clustering) להבנת הדאטה ומודלים לחיזוי הסנטימנט עבור כל חוות דעת.

1. כתבו קוד בספארק אשר חוקר את הנתונים ומציג את המיוחד בהם,
   a. מהו אורך ממוצע במילים של חוות דעת?
   b. איזה מאפיינם להשתמש
   c. האם ישנם בעיות של חוסרים בנתונים,
   d. האם יש נתונים עם חריגות (למשל חוות דעת מאד קצרה)
   e. האם יש חוות דעת שיש בהם סנטימנט שלילי וחיוביי או אירוניות/ציניות? הציגו דוגמאות למקרים כאלו.
2. אפיון למודל: לפי הדאטה שמצאתם, קבעו איזה סוג של בעיות למידה תפתחו
   a. האם זה בעית דירוג של חוות דעת לערכים בין 1 ל -10.
   b. האם זה בעית מיון של קביעת האם חוות הדעת היא חיובית או שלילית.
3. אחרי שעניתם על סעיף קודם, ענו על השאלות הבאות:
   a. איזה שיטת הערכה - evaluation נשתמש כדי להעריך ולשפר את המודל.
   b. מהם המאפיינים שנחלץ מהנתונים בשביל אימון המודל.
   c. תארו איזה מודל תשתמשו והסבירו מה יתרונותיו וחסרונותיו בפתרון הבעיה.
4. כתבו קוד אשר בונה וקטורים פשוטים(וקטור שמכיל 0 או 1 בקורדינטות) עבור כל חוות דעת.
5. חלקו את הנתונים (70-30%) לנתוני אימון ונתוני בדיקה.
6. בנו מודל של classification (למשל naive bayes), העריכו את המודל, בפרט: precision recall, auc
7. הריצו האלגוריתם של קיבוץ (clustering), שנו את מספר המקבצים ונסו להסביר את התוצאות.
8. חזרו על סעיפים 3-7 כאשר הוקטורים הם מסוג TF-IDF.
9. השוו בין התוצאות

הוראות הגשה:

1. יש להשתמש במודלי למידה קיימים בספארק, אין צורך לפתח את המודל בעצמכם.
2. יש להשתמש ב dataframe/dataset.
3. יש לשלוח לבודק את קבצי ה-JUPYTER עם ההסברים והתוצאות בתוך הקובץ.
4. יש להעביר הרצאה של 5-10 דקות של העבודה. ההגנה על הפרויקט יתקיים בשבוע של ה-4 באוגוסט

קריטריוני הערכה:

1. ביצוע המשימה.
2. קוד נקי וברור
3. מודליות
4. יצירתיות
5. הבנת הנתונים (על ידי הצגת תוצאות ניתוח של הדאטה)
6. הבנת האלגוריתם ושימושם

7. הסבר על הביצועים של המודלים.