Analytics of big data - Targil 3

מטרת התרגיל זה למשש את האלגוריתמים שלמדנו בחלק האחרון של הסמסטר עם מימוש בספארק.

1. מצאו דאטה של עשרות קבצי טקסט (יכול להיות ספרים או כל דבר אחר המתאים לתרגיל)
2. כתבו קוד אשר מקבל כקלט תיקיה של קבצי טקסט.
   a. עבור כל קובץ, מחלץ את המילים (tokens), הופך אותם למילים באותיות קטנות, מוריד סמנים (גרש וכו).
   b. הורידו stop words - אפשר להשתמש בספריה קיימת.
   c. בנו טבלה של inverted index עבור המילים שחילצתם בסעיף הקודם.
   d. בנו טבלה של ה-tf עבור המילים והמסמכים הנל.
   e. בנו טבלה עם כל המילים מסעיף א׳ ומספר המופעים שלה, וה-idf שלה.
   f. בנו טבלה של tfidf בעזרת הסעיפים הקודמים.
3. עבור על מסמך מצאו את חמשת המסמכים הקרובים אליו על ידי שימוש ב cos similarity (שתממשו). רשמו את חמשת התוצאות הטובות ביותר של קירובים בין המסמכים.
4. כתבו פונקציה שמקבלת שאילתה (משפט), מבצעת וקטוריזציה ומחזירה את עשרת המסמכים הקרובים ביותר (יש לעשות את אותו תהליך שעשיתם לעיל למסמכים). הריצו 10 דוגמאות שונות והציגו את התוצאות עם הסבר.
5. חלקו את המסמכים למספר קטן של קבוצות לפי נושא, כדאי שזה יהיה מספר קטן של נושאים ביחס למספר המסמכים (למשל ספורט, חדשות וכו).
6. ממשו את האלגוריתם של kmean, והריצו אותו על מספר הקבוצות שחלקתם בסעיף-5, בדקו את התוצאות והסבירו.

הוראות הגשה:

1. יש להשתמש בספארק, יש לפתח את המודלים הנ״ל 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-9 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.