שאלה 3)

פלט מערכת (סעיפים 2 ו3 )

insert file with word list , data and lables

2 )the 10 most discriminative word features :

3142 : graphics

3199 : algorithm

2538 : edges

2810 : hello

2864 : software

2951 : obtain

3111 : multi

3150 : polygon

3212 : displays

3070 : pub

3)learing data acorrding to traindata.txt and trainlabels.txt

accuracy of train data :0.9632422243166824

evaluating the classifier according to files: testdata.txt and testlabes.txt

classifier success to classify at accuracy of :0.7609618104667609

סעיף 2 : לפי דעתי המלים האלו בהחלט יכולים להיות בעלי הבדלים משמעותיים במידה ומדובר בשני מאמרים שונים (ז"א לדוגמא אם מאמר אחד הוא מאמר במדעי המחשב מאמר והשני מאמר במדעי החברה סבירות גבוהה שיהיה הבדלים משמעותיים ביניהם , ואכן יש מכנה משותף לרוב המילים רובם הגדול בא מסביבת מדעי המחשב (גרפיקה , אלגוריתם תוכנה ,פוליגון) ז"א כנראה שמאמר השני בהחלט שייך לנושא אחר שיש למדעי החברה .

סעיף ) 3 דיוק המסווג שבנינו הוא ברמה של הדיוק לנתוני אימון הינה 96.3% ודיוק של המסווג על תוצאות המבחן הינה 76% , צפינו לאחוז יותר של דיוק על הנתונים והן על המבחן (יכול להיות בעיות בתוכנה בין , חוסר סיווג נכון של מילים .)

סעיף 4) הנחה שכל המילים בלתי ותלויות היא בעיתית , ידוע שיש מילות הקשר (גם לנו יצאו כמה מילות הקשר) , אך אם זאת הצלחנו להוכיח שאפשר לבנות מסווג אשר יוכל לסווג מאמרים עם הנחה זו .

סעיף 5) אפשר להרחיב את המודל יש מילים שבד"כ שבאות ביחד (צמדי מילים, פתגמים ..) והכניס אותם כגורם אחד ז"א שבמקום יהיו רק מילים בלתי תליות יהיו משפטים/ביטוים שיש סבירות יותר גבוהה שהם יהיו ביחד מאשר כל מילה בנפרד .