עבודת בית 1

אסיל סמרי – 208654475

גורג גובראן – 211622600

שאלה 1)

מצאו מילה באורך 4 או יותר הכי שימושית בקובץ, שימו לב, אין חשיבות להבדל בין אותיות קטנות או גדולות.

המילה היא

“That” which is present 322 times

שאלה 2)

מצאו BIGRAM הכי שימושית בקובץ.

“the little” which is present 183 times

שאלה 3)

כמה מילים באורך לפחות 3 יש בקובץ?

Count of all words with 3 or more letters is: 12688

Count of all distinct words with 3 or more letters is: 2093

שאלה 4)

בנו SCATTER PLOT של 100 מילים הכי נפוצות, האם אתם רואים תמיכה לחוק ZIPF?

Yes, we can see a zipf-y graph like shown below

