A Needle in a Data Haystack — Introduction to Data Science

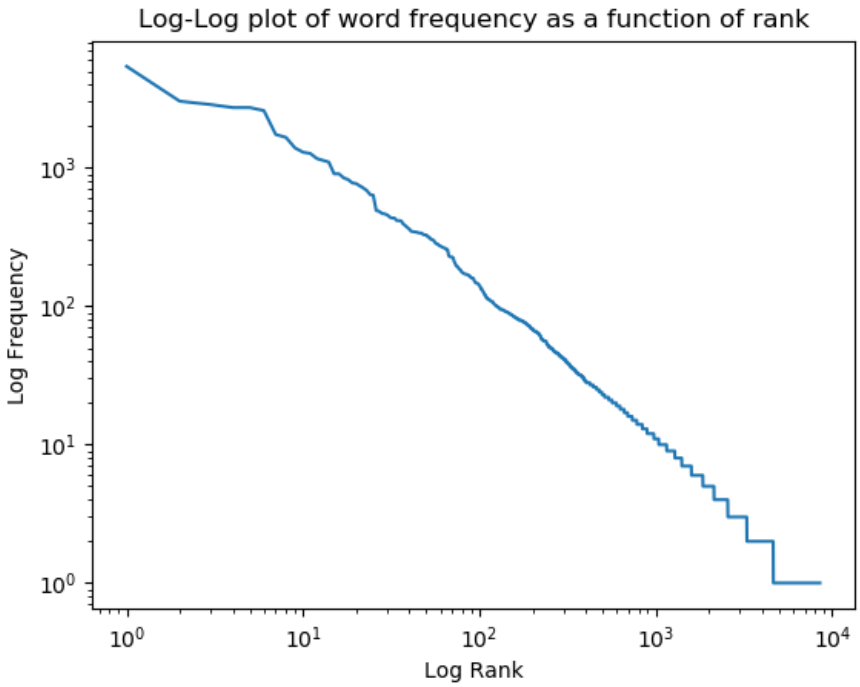
Roy Zohar (209896174) and Yoav Tamir (322291519)

January 6, 2019

Question 3

Problem 1. The book we chose to process is The Adventures of Sherlock Holmes

Problem 2. We found the following graph for the log frequency:

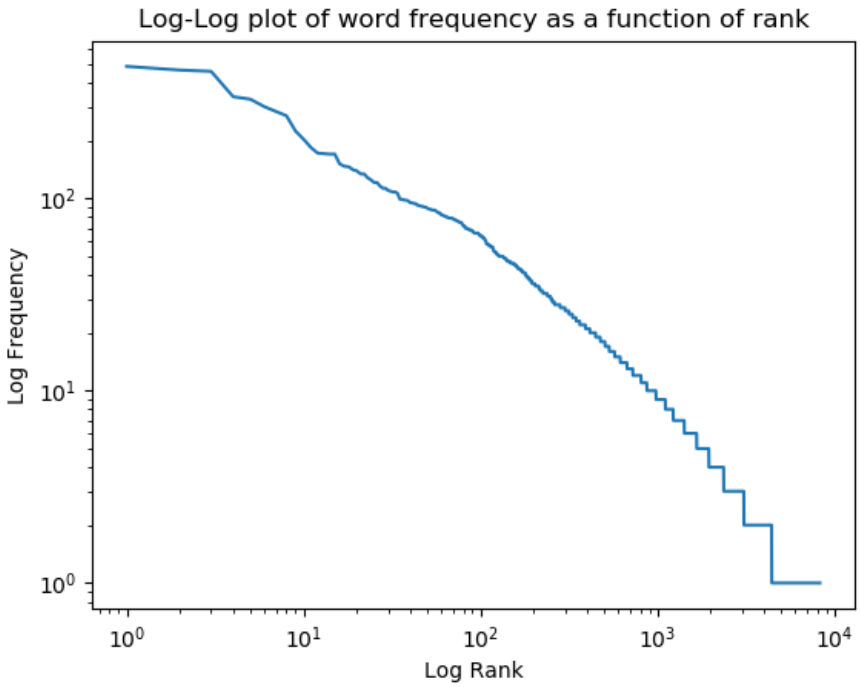


The top 20 most frequent words for this part are as follows:

1. the
2. I

3. and
4. of
5. to
6. a
7. in
8. that
9. was
10. it
11. you
12. he
13. is
14. his
15. have
16. my
17. with
18. had
19. as
20. which

Problem 3. In this part, we removed the stopwords from the text. We got the following graph for the log frequency:

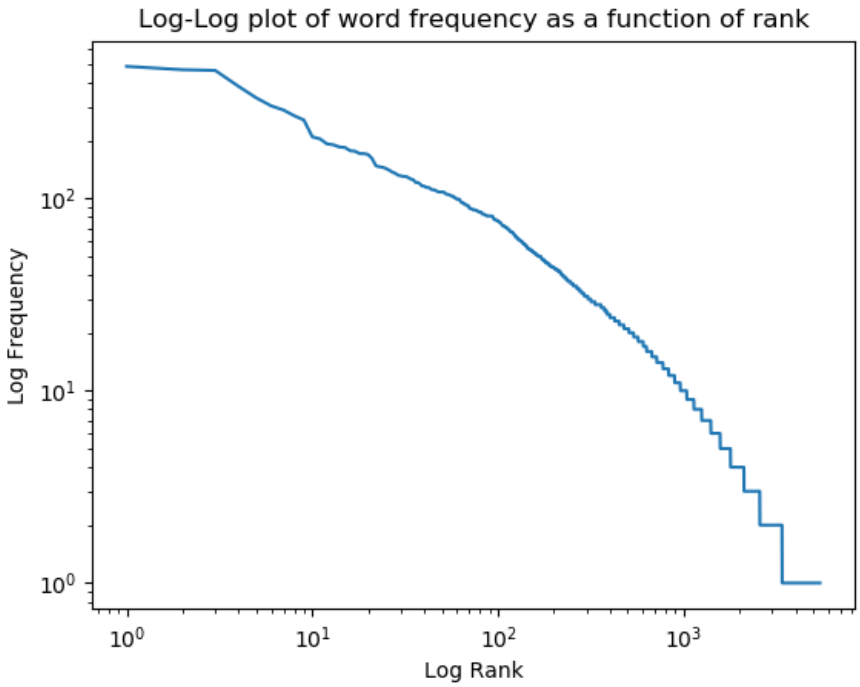


The top 20 most frequent words for this part are as follows:

1. said
2. upon
3. Holmes
4. one
5. would
6. man
7. could
8. little
9. see
10. may
11. us
12. think

13. know
14. shall
15. must
16. time
17. come
18. came
19. door
20. back

Problem 4. In this part, we removed the stopwords from the text and stemmed the text. We got the following graph for the log frequency:

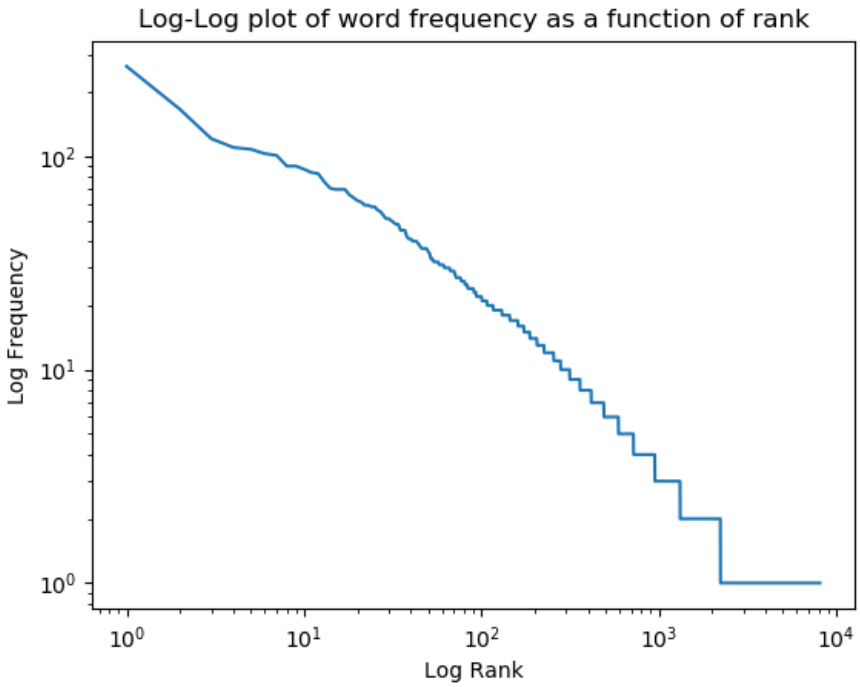


The top 20 most frequent words for this part are as follows:

1. said
2. upon
3. holm

4. one
5. would
6. man
7. could
8. littl
9. see
10. may
11. come
12. hand
13. know
14. think
15. us
16. look
17. well
18. must
19. shall
20. time

Problem 5. For this question, we ran POS-tagging on the original text, and extracted all the noun phrases. We got the following graph for the log frequency:



The top 20 most frequent noun phrases for this part are as follows.

1. Holmes
2. man
3. time
4. door
5. room
6. matter
7. way
8. hand
9. house
10. nothing
11. case
12. face

13. Sherlock Holmes
14. Well
15. Watson
16. father
17. morning
18. Mr. Holmes
19. day
20. hands

Problem 6. One example where POS-tagging failed is "wherever Sir George Burnwell". The pos tagging classified wherever as a noun, when the context implied it was a different part of speech