CMI/Aug-Sep 2019/Assignment 3 Sep 2019

Chennai Mathematical Institute

INFORMATION RETRIEVAL DEADLINE: SEP 20, 2019. MAX MARKS: 10.

ROLL NO.: ____________________

NAME: ________________________

A challenge in the implementation of any programming system is the identification of stop words. Using a programming language of your choice, implement an effective term frequency based approach to list appropriate stop words for the short description of the news items dataset available at https://www.kaggle.com/rmisra/news-category-dataset (same as you used in Assignment 2). In a short write-up (two pages or less), explain the following:

- Does Zipf's law apply to this dataset? Elaborate your answer with frequencies for top 5 tokens.
- Can we use the rule of 30 to discover stop words in a dataset? Why or why not? Reason your answer using suitable examples.
- As a general rule, Will top-k terms ordered by descending frequency of occurrence work as effective stop words? Explain with appropriate examples.
- How many stop words do you think will make an effective list for the given dataset? Explain how you arrived at them and reason why they make a list of effective stop words.

Submit your code and report on moodle. Please do not include the data files in your submission. You do not need to demonstrate your implementation.