Assignment 2
ENGR5775- 2018

A CSV file is given. The following are the column description:

1. ID
2. Time stamp
3. Title if the news story
4. Content of the news story
5. Pageview count of the news story

Task: Text data preparation for a pageview prediction task

1. Split the dataset in 70 to 30 partitions (70% training and 30% test)
2. Apply any necessary preprocessing on 3rd and 4th columns (“title” and “original content”) separately.
3. Construct Document-term matrix for both title and content and for training and test partitions.
4. What is the best interpretation of pageview to this prediction?