CS7052-Machine Learning

Workshop 11: Working with text data

You will learn:

- To practice working with text data with several datasets including movie review dataset collected by Stanford researcher Andrew Mass

Open Muller & Guido's book from the reading list. Open your Jupyter notebook.

Follow the instructions on pages 331-351 in chapter 7.

Muller and Guido's book comes with accompanying code, which you can find on https://github.com/amueller/introduction_to_ml_with_python.

You can download the code and then open corresponding file to chapter 7 (07-working-with-text-data.ipynb) in your Jupyter Notebook.

Make sure you understand the meaning of each line of code, make some changes to improve your understanding and answer the following questions:

W11.1. What is feature number 20012 in Movie Reviews dataset explained on page 337?

W11.2. How many stop words are removed from the features on page 341?

W11.3. Was removing the stop words helpful in terms of performance?

W11.4. Looking at the heat map in Figure 7-3 on page 349, do you think bigrams are better choice in our model or trigrams?

W11.5. Name three bigrams that are important features for negative reviews.

W11.6. Name three bigrams that are important features for positive reviews.

W11.7. Name three trigrams that are important features for negative reviews.

W11.8. Name three trigrams that are important features for positive reviews.

Show the output to your tutor when you are done.