05-898 Sp18: Data Science for Product Management

HW 3: Classifying Customer Complaints

1. Download data from https://data.consumerfinance.gov/dataset/Consumer-Complaints/s6ew-h6mp
2. Discard the rows where consumer complaint narrative is blank. How many rows does this yield?
3. Identify the bank that the complaint is about by extracting bank name from consumer complaint narrative. (It is OK to assume that any "Organization" name extracted will refer to the banks – determining whether an arbitrary Organization name is in fact a bank or other financial institution is beyond the scope of the assignment. Draw a histogram of number of complaints by bank name. What can you conclude about which banks are causing the most complaints?
4. Create a model to predict the product based on the consumer complaint narrative. What modeling techniques could you use? How accurate are your predictions for each field, and how did you evaluate this?
5. (Extra credit) Answer the questions in #4 for subproduct, issue-and sub-issue

HW3 Resources

- About the dataset: https://www.consumerfinance.gov/data-research/consumer-complaints/
- Extracting named entities from text http://www.nltk.org/book/ch07.html
- Learning to classify text can be found at http://www.nltk.org/book/ch06.html