CSCI-LING 5832-001,001B: Natural Language Processing
Homework 1

Instructor: Katharina Kann
GSS: Beilei Xiang
Due: September 09, 2020 at 10:00 AM
Turn in electronically via Canvas.

Problem 1 (30 Points)

Given the following short “movie reviews”, each labeled with a genre, either comedy or action:

1) fun, couple, love : Comedy
2) fast, furious, shoot : Action
3) couple, fly, fast, fun : Comedy
4) furious, shoot, fun. : Action
5) fly, fast, shoot, love : Action

Compute the most likely class for: (fast, couple, shoot, fly)

Assume a Naive Bayes classifier and use add-1 smoothing for the likelihoods.

Problem 2 (30 Points)

Write a short summary for three papers discussing the topic “bias in word embeddings”. You can search for papers on scholar.google.com or arxiv.org.

Please submit a separate pdf file containing all paper summaries. Each summary should consist of 2 to 4 sentences.

Problem 3 (30 Points)

For the given training set “hw1_training_sets.txt”, write a python program to perform the following tasks:

1) Count the tokens and types in the training set.
2) Find the 5 most frequent words with their frequency.

Submit your python code as “lastname-firstname-hw1.py” and an output file “output.txt” containing:

- First line for number of tokens.
- Second line for number of types.
- Keep third line blank.
- Lines 4-8: Most frequent words with their frequency, one per line; word and frequency should be space-separated.

Problem 4 (10 Points)

Lemmatize the following sentences:

1) The striped bats were hanging on their feet and ate best fishes.
2) At some point, I realized that I started to use the bike more often, not only to get to work but also to catch up with friends and to head out for coffee on weekends.

If you have any doubts, please post your questions on Piazza.