CS 6120 Natural Language Processing
Northeastern University, Spring 2019

Assignment 1

Ruosen Li: li.ruos@husky.neu.edu

1 Naive Bayes for Text Categorization

Priors:

$$|V| = 14 \quad \lambda = 0.1$$
$$P(vegetable) = \frac{2}{8} \quad P(flower) = \frac{3}{8} \quad P(fruit) = \frac{3}{8}$$

Conditional Probabilities:

$$P(banana|vegetable) = \frac{2+0.1}{8+1.4} = \frac{21}{94} \qquad P(carrot|vegetable) = \frac{1+0.1}{8+1.4} = \frac{11}{94}$$
$$P(cucumber|vegetable) = \frac{1+0.1}{8+1.4} = \frac{11}{94} \qquad P(pea|vegetable) = \frac{2+0.1}{8+1.4} = \frac{21}{94}$$
$$P(potato|vegetable) = \frac{1+0.1}{8+1.4} = \frac{11}{94} \qquad P(basket|vegetable) = \frac{1+0.1}{8+1.4} = \frac{11}{94}$$
$$P(others|vegetable) = \frac{0+0.1}{8+1.4} = \frac{1}{94}$$
$$P(lotus|flower) = \frac{2+0.1}{13+1.4} = \frac{7}{48} \qquad P(pea|flower) = \frac{2+0.1}{13+1.4} = \frac{7}{48}$$
$$P(rose|flower) = \frac{3+0.1}{13+1.4} = \frac{31}{144} \qquad P(lily|flower) = \frac{2+0.1}{13+1.4} = \frac{7}{48}$$
$$P(hibiscus|flower) = \frac{3+0.1}{13+1.4} = \frac{31}{144} \qquad P(cucumber|flower) = \frac{1+0.1}{13+1.4} = \frac{11}{144}$$
$$P(others|flower) = \frac{0+0.1}{13+1.4} = \frac{1}{144}$$
$$P(hibiscus|fruit) = \frac{2+0.1}{14+1.4} = \frac{3}{22} \qquad P(grape|fruit) = \frac{2+0.1}{14+1.4} = \frac{3}{22}$$
$$P(school|fruit) = \frac{1+0.1}{14+1.4} = \frac{1}{14} \qquad P(mango|fruit) = \frac{2+0.1}{14+1.4} = \frac{3}{22}$$
$$P(apple|fruit) = \frac{3+0.1}{14+1.4} = \frac{31}{154} \qquad P(lotus|fruit) = \frac{1+0.1}{14+1.4} = \frac{1}{14}$$
$$P(lily|fruit) = \frac{1+0.1}{14+1.4} = \frac{1}{14} \qquad P(banana|fruit) = \frac{1+0.1}{14+1.4} = \frac{1}{14}$$
$$P(rose|fruit) = \frac{1+0.1}{14+1.4} = \frac{1}{14} \qquad P(others|fruit) = \frac{0+0.1}{14+1.4} = \frac{1}{154}$$

Note: "others" represents each word in total vocabulary which do not appear in corresponding categories.

Choosing a class:

$$P(vegetable|D1) \propto \frac{2}{8} \times \frac{1}{94} \times \frac{1}{94} \times \frac{1}{94} \times \frac{11}{94} \approx 3.52 \times 10^{-8}$$
$$P(flower|D1) \propto \frac{3}{8} \times \frac{31}{144} \times \frac{7}{48} \times \frac{1}{144} \times \frac{1}{144} \approx 5.67 \times 10^{-7}$$
$$P(fruit|D1) \propto \frac{3}{8} \times \frac{1}{14} \times \frac{1}{14} \times \frac{31}{154} \times \frac{1}{154} \approx 2.50 \times 10^{-6}$$
$$P(vegetable|D2) \propto \frac{2}{8} \times \frac{21}{94} \times \frac{11}{94} \times \frac{1}{94} \times \frac{1}{94} \approx 7.40 \times 10^{-7}$$
$$P(flower|D2) \propto \frac{3}{8} \times \frac{7}{48} \times \frac{1}{144} \times \frac{7}{48} \times \frac{1}{144} \approx 3.85 \times 10^{-7}$$
$$P(fruit|D2) \propto \frac{3}{8} \times \frac{1}{154} \times \frac{1}{154} \times \frac{1}{14} \times \frac{3}{22} \approx 1.54 \times 10^{-7}$$

Results:
According to the results, D1 most likely belongs to fruit category and D2 most likely belongs to vegetable category.

2 Word Sense Disambiguation

2.1 The total number of senses for each open class word

right	15	arm	06	lay	05	cushion	01	parapet	02	raise	27
hand	14	make	49	light	25	quick	06	movement	11	lover	03
see	24	eye	05	fix	12	man	11	arena	04	turn	26
firm	10	rapid	02	step	11	walk	10	empty	04	space	09

2.2 Distinct combinations of senses

The first sentence:	900
The second sentence:	458419500
The third sentence:	72
The forth sentence:	2640
The fifth sentence:	2059200

3 Language Modeling

Code for this problem is under folder "Problem3". Please read "Problem3/README" before running it.

3.1 N-gram Language Model

The count file is saved under folder "Problem3/lm" named "unigram", "bigram", and "trigram" respectively.

3.2 Linear Interpolation Smoothing

The score file is under folder "Problem3/3_2". All first 50 highest score files are French file.

3.3 Add-λ Smoothing

The score file is under folder "Problem3/3_3". All first 50 highest score files are French file.

3.4 Comparison

According to the score files calculated by two smoothing function, interpolation smoothing function is obviously better than add-λ function.

Linear interpolation smoothing function:
Pros:
1. Perform well on data and report in high accuracy.
Cons:
1. Run in relatively slower speed than add-λ smoothing function.
2. Need to find combination of lambdas first by grid search when using linear interpolation smoothing. If you need to calculate lambda based on context, it will cost more time and be hard to implement.

Add-λ smoothing function:
Pros:
1. Run fast than interpolation smoothing function.
2. Easy to implement.
Cons:
1. Bad performance on all data than interpolation smoothing.

4 POS Tagging - HMM

Code for this problem is under folder "Problem3". Please read "README" before running it.

4.1 HMM parameters

The count file is saved under folder "Problem4/4_1" named "word-tag", "tag-unigram", and "tag-bigram" respectively.

4.2 Transition Probabilities

The result is showed in Jupyter notebook as a output result after running the last cell or corresponding function.

4.3 Emission Probabilities

The result is showed in Jupyter notebook as a output result after running the last cell or corresponding function.

4.4 Sentences generation

The result is showed in Jupyter notebook as a output result after running the last cell or corresponding function.

4.5 POS tag parsing

The POS tag file is under folder "Problem4/4_5".