CS440 MP3
Team member: Mingren Feng (mingren3) ECE448, Q3;
Wenyao Jin (wenyaoj2) CS440, Q4
Ziyi Chen (ziyic2) CS440, R3
Date: 2019/04/01

Section I: Image Classification

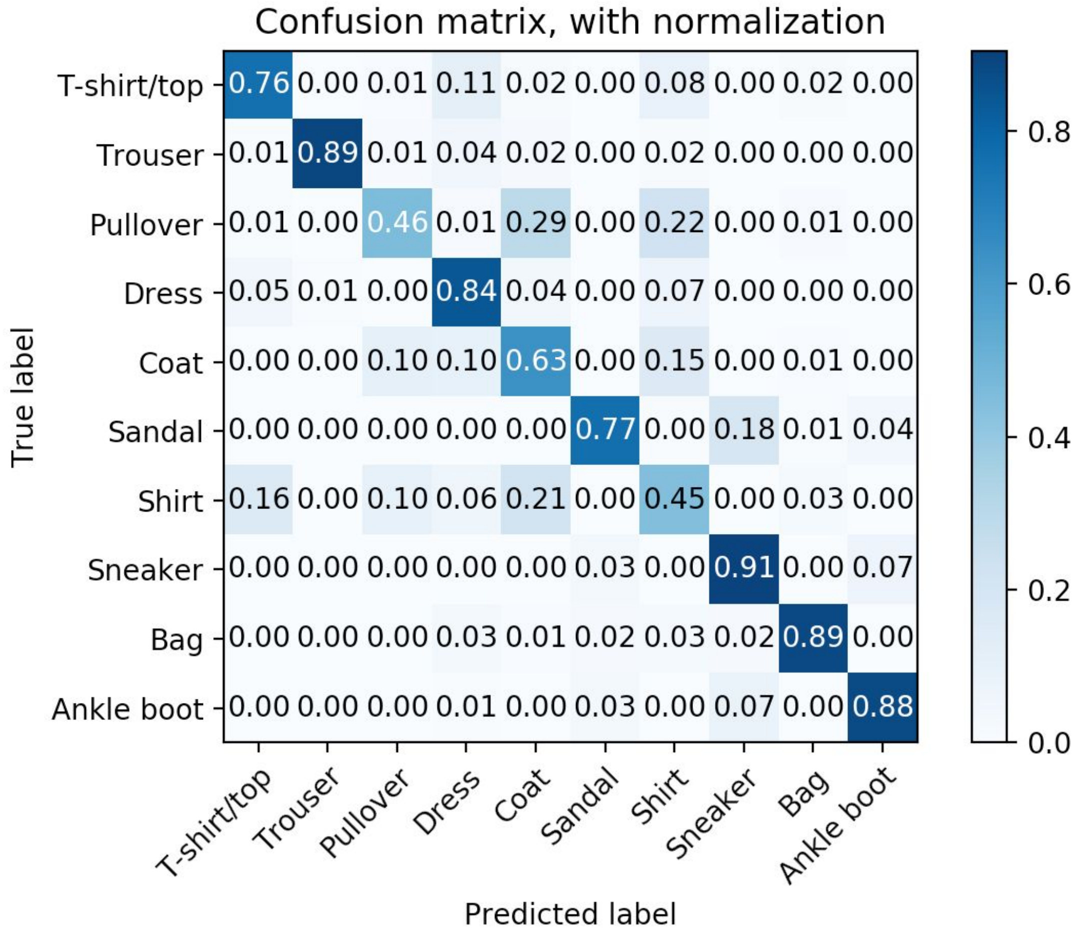
Part 1.1 Naive Bayes

For Laplace smoothing, we find that the smaller k would lead to the higher accuracy. Therefore, we set k=0.1. Average classification rate=0.747

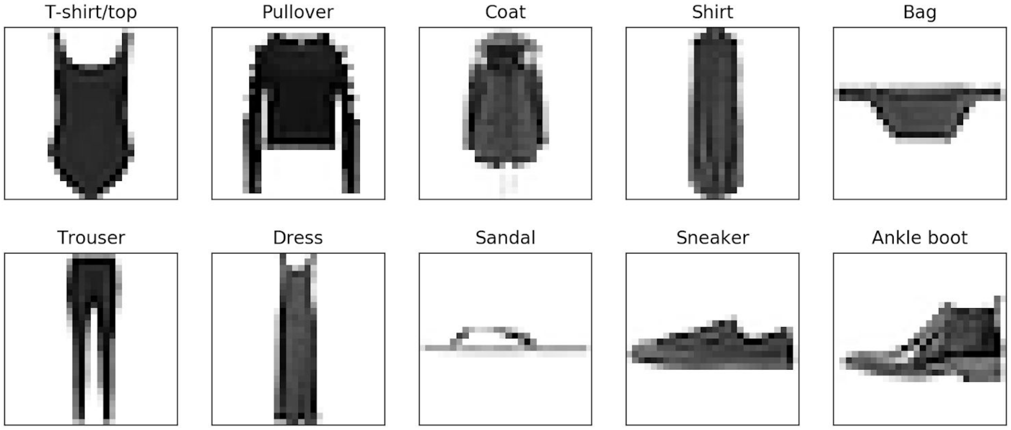
Classification rate for each class

Class	Classification rate
T-shirt/top	0.76
Trouser	0.89
Pullover	0.46
Dress	0.84
Coat	0.63
Sandal	0.77
Shirt	0.45
Sneaker	0.91
Bag	0.89
Ankle boot	0.88

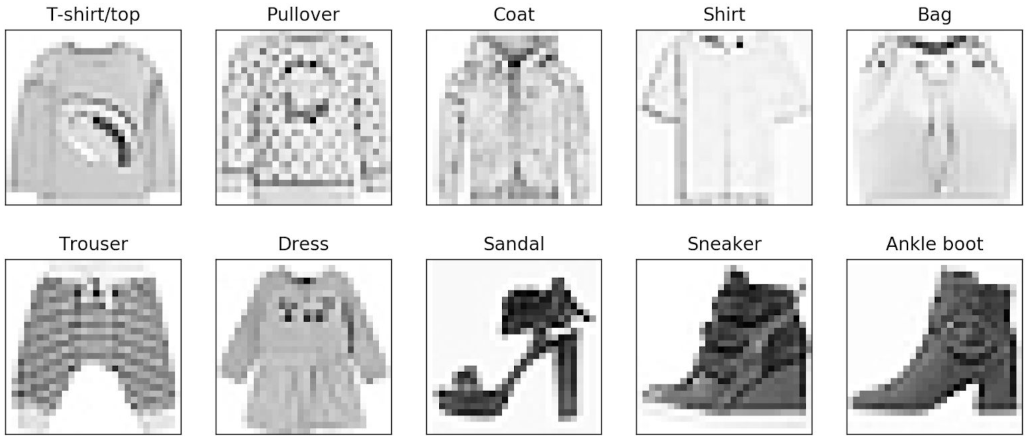
Confusion matrix



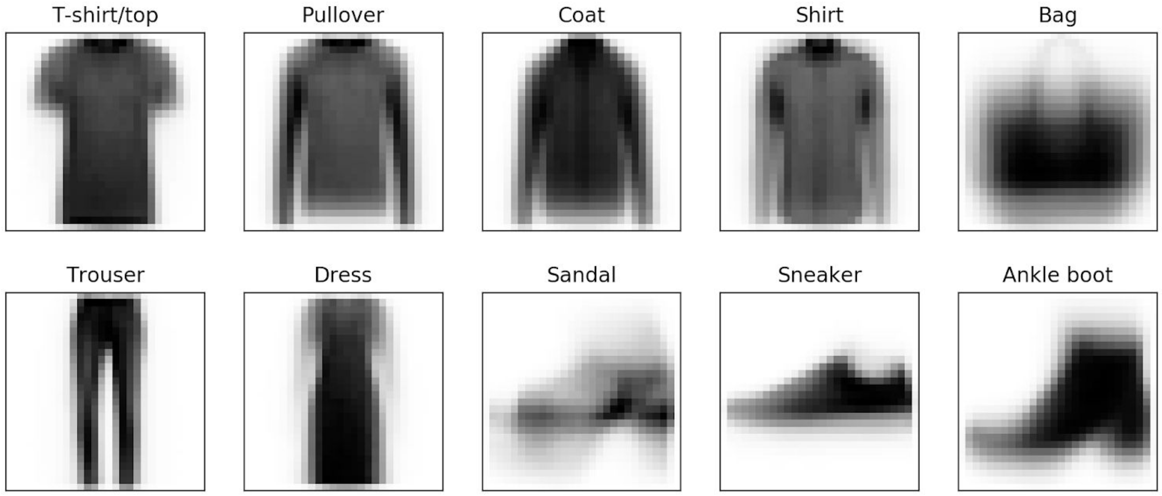
Images in each class with highest posterior probabilities



Images in each class with lowest posterior probabilities



Ten visualization plots for feature likelihoods



Part 1.2 Perceptron

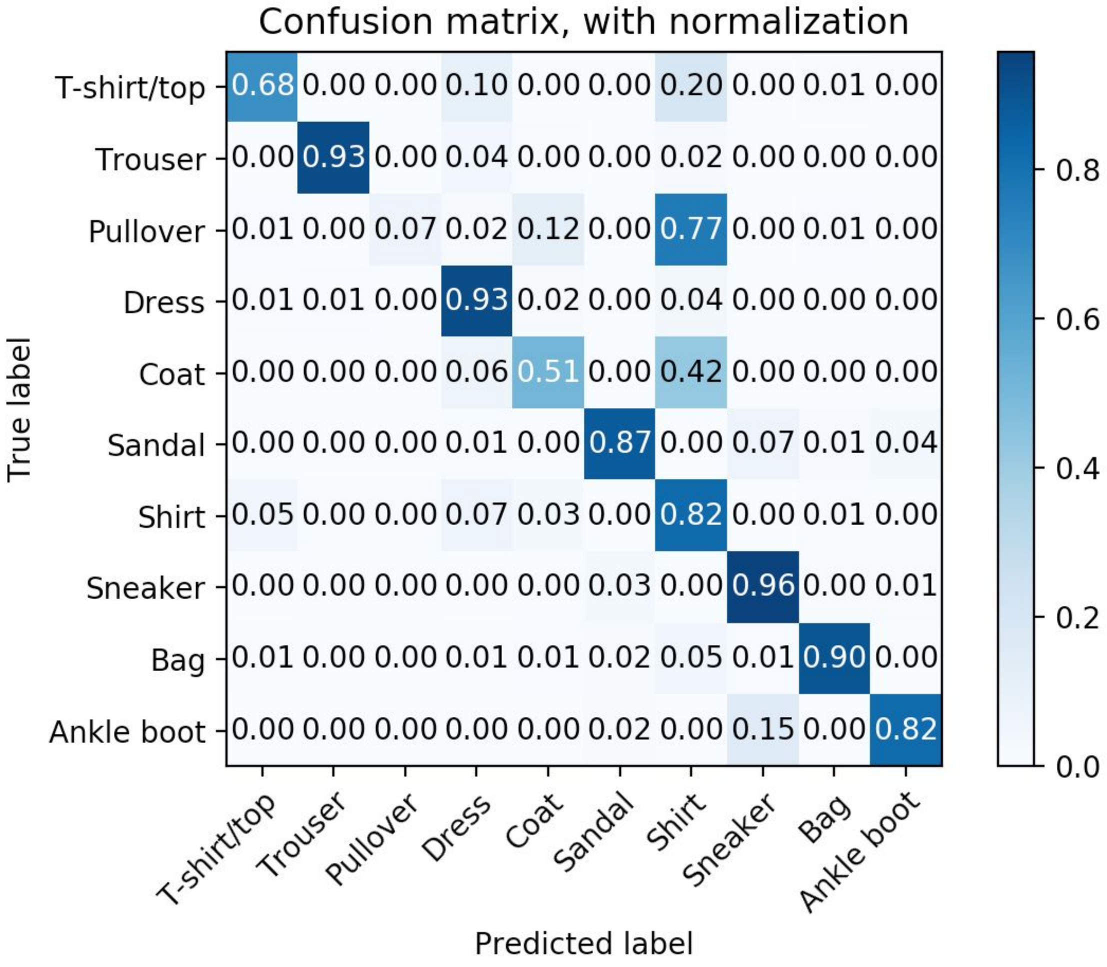
Average classification rate=0.7488

Classification rate for each class

Class	Classification rate
T-shirt/top	0.68
Trouser	0.93

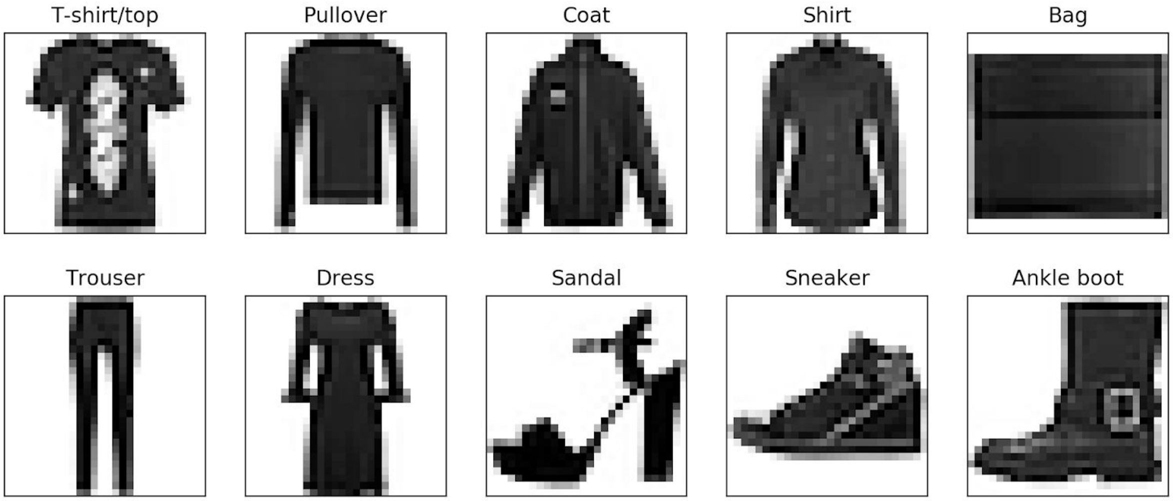
Pullover	0.07
Dress	0.93
Coat	0.51
Sandal	0.87
Shirt	0.82
Sneaker	0.96
Bag	0.90
Ankle boot	0.82

Confusion matrix

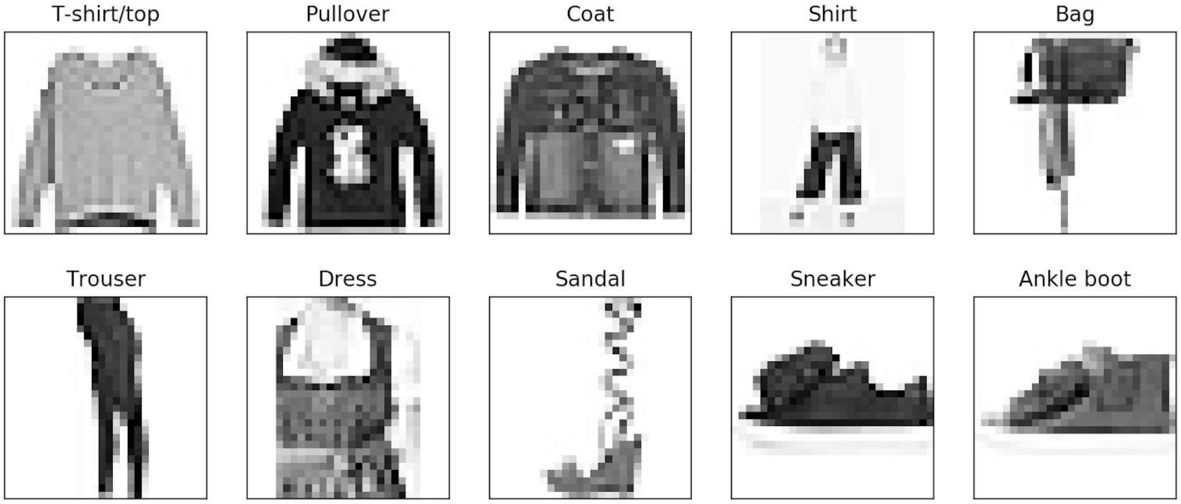


	T-shirt/top	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
T-shirt/top	0.68	0.00	0.00	0.10	0.00	0.00	0.20	0.00	0.01	0.00
Trouser	0.00	0.93	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.00
Pullover	0.01	0.00	0.07	0.02	0.12	0.00	0.77	0.00	0.01	0.00
Dress	0.01	0.01	0.00	0.93	0.02	0.00	0.04	0.00	0.00	0.00
Coat	0.00	0.00	0.00	0.06	0.51	0.00	0.42	0.00	0.00	0.00
Sandal	0.00	0.00	0.00	0.01	0.00	0.87	0.00	0.07	0.01	0.04
Shirt	0.05	0.00	0.00	0.07	0.03	0.00	0.82	0.00	0.01	0.00
Sneaker	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.96	0.00	0.01
Bag	0.01	0.00	0.00	0.01	0.01	0.02	0.05	0.01	0.90	0.00
Ankle boot	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.15	0.00	0.82

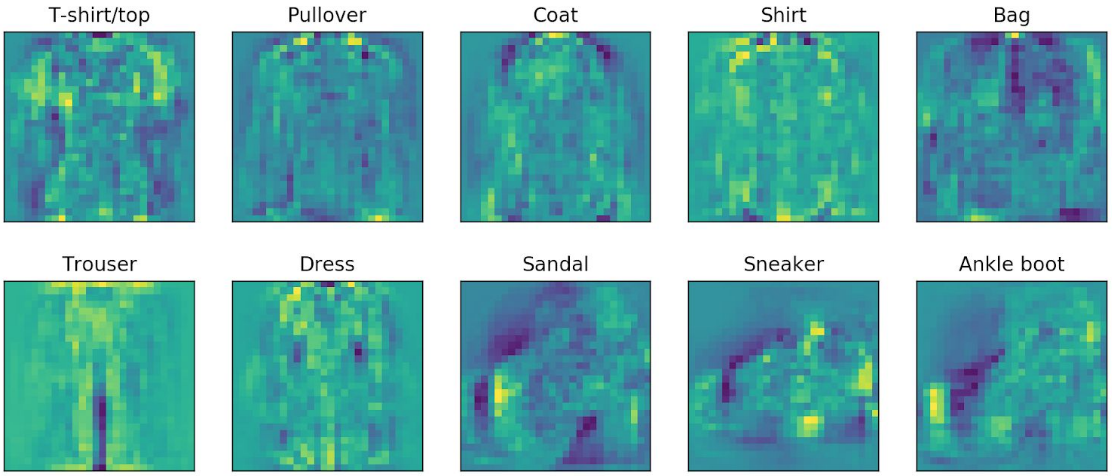
Images in each class with highest posterior probabilities



Images in each class with lowest posterior probabilities

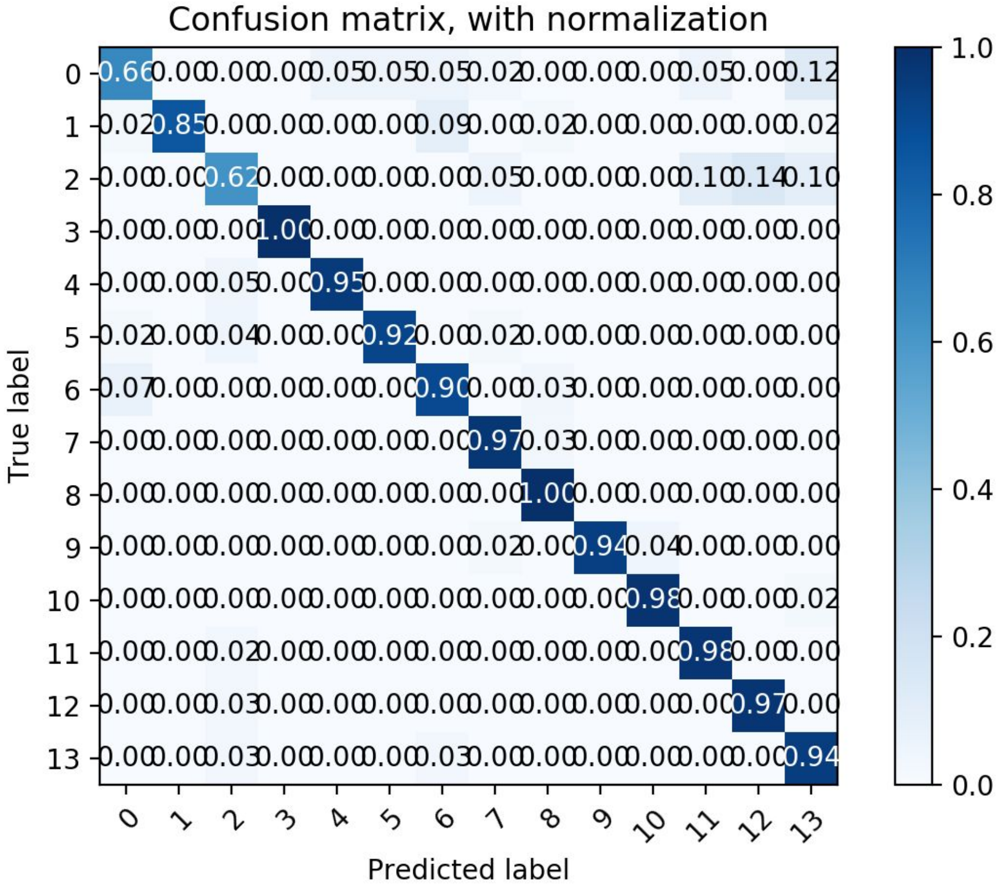


Ten visualization plots for perceptron weight vectors



Section II:Text Classification

Confusion matrix:



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0.66	0.00	0.00	0.00	0.05	0.05	0.05	0.02	0.00	0.00	0.00	0.05	0.00	0.12
1	0.02	0.85	0.00	0.00	0.00	0.00	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.02
2	0.00	0.00	0.62	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.10	0.14	0.10
3	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.05	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.02	0.00	0.04	0.00	0.00	0.92	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
6	0.07	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.03	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.03	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.94	0.04	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.02
11	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00
12	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00
13	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.94

(label should be switched by 1, so 0 means class 1, 13 means class 14)

Precision for all classes : [0.9230769230769231, 1.0, 0.6842105263157895, 1.0, 0.9130434782608695, 0.9361702127659575, 0.8, 0.8918918918918919, 0.7272727272727273, 1.0, 0.9565217391304348, 0.9111111111111111, 0.925, 0.7727272727272727]

Recall for all classes: [0.5853658536585366, 0.8695652173913043, 0.6190476190476191, 1.0, 0.9545454545454546, 0.9166666666666666, 0.9333333333333333, 0.9705882352941176, 1.0, 0.94, 0.9777777777777777, 0.9761904761904762, 0.9736842105263158, 0.9714285714285714]

F1 Score for all classes: [0.7164179104477613, 0.9302325581395349, 0.6500000000000001, 1.0, 0.9333333333333332, 0.9263157894736843, 0.8615384615384616, 0.9295774647887325, 0.8421052631578948, 0.9690721649484536, 0.967032967032967, 0.9425287356321839, 0.9487179487179489, 0.8607594936708862]

Accuracy 0.9048

Featured words

1.
company, based, business, founded, record, records, bergen, systems, services, products, office, buses, located, distribution, national, health, virgin, established, also, regional.
2.
school, high, located, university, college, public, schools, students, education, district, county, founded, one, new, united, established, independent, city, part, catholic.
3.
born, american, known, new, band, writer, best, rock, musician, music, work, also, singer, york, books, author, album, university, series, united.
4.
born, football, played, league, professional, player, plays, footballer, former, national, american, also, hockey, currently, rugby, team, australian, november, world, new.
5.
born, member, district, politician, state, senate, democratic, house, party, served, former, county, since, representatives, republican, united, elected, american, representing, national.
6.
navy, built, war, ship, uss, united, class, aircraft, world, states, launched, service, named, first, designed, royal, commissioned, american, ii, us.
7.
historic, house, built, located, church, building, national, register, places, listed, county, street, united, known, museum, also, states, designed, added, hospital.
8.
river, lake, mountain, located, south, km, north, county, near, tributary, west, range, lies, creek, crater, east, ft, state, flows, pass.
9.
village, district, population, province, located, census, municipality, nepal, state, india, county, km, people, within, 2010, 1991, south, township, central, southern.
10.
family, species, found, genus, moth, gastropod, sea, known, marine, described, tropical, snail, mollusk, endemic, subtropical, habitat, natural, forests, snails, moist.
11.
species, family, plant, genus, native, endemic, flowering, known, found, common, plants, leaves, habitat, tree, grows, name, orchid, south, bulbophyllum, perennial.
12.
album, released, band, records, first, studio, american, songs, music, second, release, recorded, rock, debut, live, tracks, label, albums, new, ep.
13.

film, directed, starring, american, stars, released, written, based, drama, comedy, produced, also, films, silent, first, movie, roles, novel, name, documentary.
14.
published, book, novel, first, journal, written, series, newspaper, american, story, author, new, magazine, fiction, books, peerreviewed, also, science, publication, life.

Prior changes

With prior: 0.9048
When it’s removed: 0.9048
There are some different in precision, but the accuracy is remained that same. Based on the featured words, it is noticeable that different class have very different featured words, since they are quite disjoint, the classification result remains unchanged.

Extra Credit:

For the bigram case, I added a bigram dictionary from two words to the word cont to store the result of the training, and use the formula P(w1..wn)=P(w1)P(w2|w1)..P(wn|wn−1) to get the P(bi|Y).

The Lambda turns out to be 0, since in our case unigram has a very high accuracy and mixed with bigram will decrease its accuracy, so the answer would be 0, with accuracy 0.9048

1. The accuracy of bigram turns out to be 0.71 which is lower than the unigram model. The lower accuracy is probably resulted by the week relation between two adjacent words, or another possibility is that the text isn’t large enough for enough example of bigrams. Since the combination of two words are the square of a single words, so the paragraph needs to be long enough to expose enough bigram cases for classification.
2. For N-gram, the paragraph has to be long and logical, but the 14 classes are rather ambiguous on logic, so it’s unsure whether it’s beneficial to have N-gram. However, like the previous argument, the text need to be large enough for N-gram to make sense, and there are also more train_set required. For the 3000 level, train_set, and around 30 words text, I guess the unigram would be the best fit .

Statement of Contribution:

Wenyao Jin and Mingren Feng worked on Section I

Ziyi Chen worked on Section II