

Московский авиационный институт  
(национальный исследовательский университет)

Факультет информационных технологий и прикладной  
математики

Кафедра вычислительной математики и программирования

Курсовой проект по курсу «Дискретный анализ»

Студент: А. В. Семин  
Преподаватель: С. А. Сорокин  
Группа: М8О-306Б-20  
Дата:  
Оценка:  
Подпись:

Москва, 2023

# Курсовой проект

**Задача:** Реализуйте систему, которая на основе базы вопросов и тегов к ним, будет предлагать варианты тегов, которые подходят к новым вопросам.

## Формат ввода

Формат запуска программы в режиме обучения:

```
./prog learn -input <input file> -output <stats file>
```

Ключ	Значение
<code>- - input</code>	входной файл с вопросами
<code>- - output</code>	выходной файл с рассчитанной статистикой

Формат запуска программы в режиме классификации:

```
./prog classify --stats <stats file> --input <input file> --output <output file>
```

Ключ	Значение
<code>- - stats</code>	файл со статистикой полученной на предыдущем этапе
<code>- - input</code>	входной файл с вопросами
<code>- - output</code>	выходной файл с тегами к вопросам

Формат входных файлов при обучении:

<Количество строк в вопросе [n]>

<Тег 1>,<Тег 2>,...,<Тег m>

<Заголовок вопроса>

<Текст вопроса [n строк]>

Формат входных файлов при запросах:

<Количество строк в вопросе [n]>

<Заголовок вопроса>

<Текст вопроса [n строк]>

Формат выходного файла: для каждого запроса в отдельной строке выводится предполагаемый набор тегов, через запятую.

## Формат вывода

Для каждого запроса в отдельной строке выводится предполагаемый набор тегов, через запятую.

# 1 Описание

Подсчет вероятности подхода тегов к вопросам осуществляется с помощью наивного Байесовского алгоритма.

**Наивный байесовский алгоритм** — это алгоритм классификации, основанный на теореме Байеса с допущением о независимости признаков. Другими словами, НБА предполагает, что наличие какого-либо признака в классе не связано с наличием какого-либо другого признака. Например, фрукт может считаться яблоком, если он красный, круглый и его диаметр составляет порядка 8 сантиметров. Даже если эти признаки зависят друг от друга или от других признаков, в любом случае они вносят независимый вклад в вероятность того, что этот фрукт является яблоком. В связи с таким допущением алгоритм называется «наивным».

Модели на основе НБА достаточно просты и крайне полезны при работе с очень большими наборами данных. При своей простоте НБА способен превзойти даже некоторые сложные алгоритмы классификации.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} - \text{формула Байеса}$$

**Теорема Байеса** (или формула Байеса) — одна из основных теорем элементарной теории вероятностей, которая позволяет определить вероятность события при условии, что произошло другое статистически взаимозависимое с ним событие. Другими словами, по формуле Байеса можно уточнить вероятность какого-либо события, взяв в расчёт как ранее известную информацию, так и данные новых наблюдений. Формула Байеса может быть выведена из основных аксиом теории вероятностей, в частности из условной вероятности. Особенность теоремы Байеса заключается в том, что для её практического применения требуется большое количество расчётов, вычислений, поэтому байесовские оценки стали активно использовать только после революции в компьютерных и сетевых технологиях. На сегодняшний день активно применяется в машинном обучении и технологиях искусственного интеллекта.

Стоит отметить одну из основных отрицательных сторон алгоритма. Если в тестовом наборе данных присутствует некоторое значение категориального признака, которое не встречалось в обучающем наборе данных, то модель присвоит нулевую вероятность этому значению и не сможет сделать прогноз. Это явление известно под названием «нулевая частота» (zero frequency). Данную проблему можно решить с помощью сглаживания. Одним из самых простых методов является сглаживание по Лапласу (Laplace smoothing).

## 2 Исходный код

Файл main.cpp:

```
1 #include <cstring>
2 #include <fstream>
3 #include <unordered_set>
4 #include <cmath>
5 #include "Bayes.h"
6 using namespace std;
7
8 file makeVector(std::string &text) { //
9     std::vector<std::string> words;
10    std::string readingWord;
11
12    for (int i = 0; i < text.size(); i++) {
13        char c = tolower(text[i]);
14        if (c >= 'a' && c <= 'z') {
15            readingWord += c;
16        } else if (readingWord.length() > 0) {
17            words.push_back(readingWord);
18            readingWord = "";
19        }
20        if (i + 1 == text.size() && readingWord.length() > 0) {
21            words.push_back(readingWord);
22        }
23    }
24    return words;
25 }
26
27 std::unordered_map<std::string, long double> softmax(std::unordered_map<std::string,
28     long double>& probs);
29
30 void printError() {
31     std::cerr << "Invalid syntax\n"
32         "    ./prog learn --input <input file> --output <stats file>\n"
33         "    ./prog classify --stats <stats file> --input <input file> --output <
34         output file>\n";
35 }
36
37 void printEmptyError() {
38     std::cerr << "Invalid syntax:\n"
39         "<input file> (or <output file>) field is empty\n";
40 }
41
42 bool checkParamLearn(int argc, char* argv[]) {
43     if (argc == 6 && !strcmp(argv[2], "--input") && !strcmp(argv[4], "--output") &&
44         strcmp(argv[3], argv[5])) {
45         return true;
46     } else {
```

```

44     printError();
45     return false;
46 }
47 }
48
49 bool checkParamClassify(int argc, char* argv[]) {
50     if (argc == 8 && !strcmp(argv[2], "--stats") && !strcmp(argv[4], "--input") && !
        strcmp(argv[6], "--output") && strcmp(argv[3], argv[5]) && strcmp(argv[5], argv
        [7]) && strcmp(argv[3], argv[7])) {
51         return true;
52     } else {
53         printError();
54         return false;
55     }
56 }
57
58 void readText(std::ifstream& inStream, ::string& text, int lines) {
59     for (int i = 0; i < lines + 1; ++i) {
60         std::string cur_line;
61         getline(inStream, cur_line);
62         text += " " + cur_line;
63     }
64 }
65
66 void readFileToDataset(std::ifstream& inStream, std::vector<data>& dataset) {
67     int lines;
68     while (inStream >> lines) {
69         std::string types;
70         std::string text;
71         inStream.ignore();
72         getline(inStream, types);
73         readText(inStream, text, lines);
74         dataset.push_back({makeVector(types), makeVector(text)});
75     }
76 }
77
78 void writeToOutStream(std::ofstream& outStream, std::unordered_map<std::string, long
    double> preds, double threshold) {
79     bool comma = false;
80     for (const auto& pred : preds) {
81         if (pred.second > threshold) {
82             if (comma) {
83                 outStream << ", ";
84             }
85             outStream << pred.first;
86             comma = true;
87         }
88     }
89     outStream << "\n";

```

```

90 }
91
92 bool wrongArgsCount(int argc) {
93     return argc < 6;
94 }
95
96
97 std::unordered_map<std::string, long double> softmax(std::unordered_map<std::string,
98     long double>& probs) {
99     long double max = -1e10;
100     for (const auto& prob : probs) {
101         if (prob.second > max) {
102             max = prob.second;
103         }
104     }
105     long double sum = 0;
106     for (auto& prob : probs) {
107         sum += exp(prob.second - max);
108     }
109     long double constant = max + log(sum);
110     std::unordered_map<std::string, long double> res;
111     for (auto prob : probs) {
112         res[prob.first] = exp(prob.second - constant);
113     }
114     return res;
115 }
116
117
118 void makePredictAndWriteOutput(std::ifstream& inStream, std::ofstream& outStream,
119     BayesClassifier& BC) {
120     int lines;
121     while (inStream >> lines) {
122         std::string text;
123         inStream.ignore();
124         readText(inStream, text, lines);
125         file doc = makeVector(text);
126
127         std::unordered_map<std::string, long double> probas = BC.predict(doc);
128         std::unordered_map<std::string, long double> preds = softmax(probas);
129
130         double threshold = 1. / BC.getTagsCount();
131         writeToOutputStream(outStream, preds, threshold);
132     }
133 }
134
135 int main(int argc, char* argv[]) {
136     std::string inFile;
137     std::string statsFile;

```

```

137     std::string outFile;
138     if (wrongArgsCount(argc)) {
139         printError();
140         return 1;
141     }
142
143     if (!strcmp(argv[1], "learn")) {
144         if (!checkParamLearn(argc, argv)) {
145             return 1;
146         }
147         inFile = argv[3];
148         statsFile = argv[5];
149
150         std::ifstream inStream(inFile);
151         std::ofstream statsStream(statsFile);
152         std::vector<data> dataset;
153         readFileToDataset(inStream, dataset);
154         BayesClassifier BC;
155         BC.initMaps(dataset);
156         BC.saveStats(statsStream);
157
158     } else if (argc >= 6 && !strcmp(argv[1], "classify")) {
159         if (!checkParamClassify(argc, argv)) {
160             return 1;
161         }
162         statsFile = argv[3];
163         inFile = argv[5];
164         outFile = argv[7];
165
166         std::ifstream inStream(inFile);
167         std::ifstream statsStream(statsFile);
168         std::ofstream outStream(outFile);
169
170         BayesClassifier BC;
171         BC.loadStats(statsStream);
172         makePredictAndWriteOutput(inStream, outStream, BC);
173     } else {
174         printError();
175         return 1;
176     }
177 }

```

Файл Bayes.h:

```

1  #pragma once
2  #include <iostream>
3  #include <vector>
4  #include <unordered_map>
5
6  using file = std::vector<std::string>;

```

```

7
8 struct data {
9     file tag;
10    file doc;
11 };
12
13 class BayesClassifier {
14 public:
15     void initMaps(std::vector<data>& dataset) {
16         for (const auto& data : dataset) {
17             file fileTypes = data.tag;
18             file doc = data.doc;
19             for (auto& type : fileTypes) {
20                 for (auto& word : doc) {
21                     wordsInType[type][word]++;
22                     wordsCount[word]++;
23                     fileTypesCount[type]++;
24                 }
25             }
26             totalCount += doc.size();
27         }
28     }
29
30     std::unordered_map<std::string, long double> predict(file &doc) {
31         std::unordered_map<std::string, long double> probabilities;
32         for (auto& data : wordsInType) {
33             std::string curType = data.first;
34             probabilities[curType] = probability(curType, doc);
35         }
36         return probabilities;
37     }
38
39     void saveStats(std::ofstream &out) {
40         out << totalCount << "\n";
41         for (auto& dataType : wordsInType) {
42             std::string tag = dataType.first;
43             auto data = dataType.second;
44             out << tag << " " << data.size() << " ";
45             for (auto& wordsCount : data) {
46                 out << wordsCount.first << " " << wordsCount.second << " ";
47             }
48             out << "\n";
49         }
50     }
51
52     void loadStats(std::ifstream &in) {
53         in >> totalCount;
54         std::string tag;
55         while (in >> tag) {

```



```

56         int amountOfWords;
57         in >> amountOfWords;
58         for (int i = 0; i < amountOfWords; ++i) {
59             std::string word;
60             int count;
61             in >> word >> count;
62             wordsInType[tag][word] = count;
63             wordsCount[word] += count;
64             fileTypeCounts[tag] += count;
65         }
66     }
67 }
68
69 int getTagsCount() {
70     return fileTypeCounts.size();
71 }
72
73 private:
74     long double alpha = 1;
75     std::unordered_map<std::string, int> fileTypeCounts; //-
76     std::unordered_map<std::string, std::unordered_map<std::string, int>> wordsInType;
77     //      <, <, ->>
78     std::unordered_map<std::string, int> wordsCount; //  <, ->
79     int totalCount = 0; //
80
81     long double probability(std::string& fileType, file& doc) { //  $P(\text{type} \mid \text{doc})$  ,
82         doc    fileType
83         long double prob = 0;
84         for (auto word : doc) {
85             prob += log(probability(word, fileType));
86         }
87         prob += log(probability(fileType));
88         return prob;
89     }
90
91     long double probability(std::string &tag) { //  $P(\text{type})$ 
92         return (long double) (fileTypeCounts[tag] + alpha) / (totalCount + alpha *
93             totalCount);
94     }
95
96     long double probability(std::string& word, std::string& fileType) { //  $P(\text{word} \mid$ 
97         type)
98         return (long double) (wordsInType[fileType][word] + alpha) / (wordsCount[word]
99             + alpha * wordsInType.size());
100     }
101 };

```

### 3 Консоль и тестовые файлы

Файл study: (содержит набор данных для обучения)

```
1 1
2 Fish
3 Common info
4 Fish are creatures that live in water and cannot survive on land. Pet fish live in
  small aquariums. Usually domestic fish are small
5 1
6 Cats, fish
7 Dangerous
8 Is it dangerous to leave a cat at home alone with fish? an a cat get into an aquarium
  and eat all the pet fish?
9 1
10 Fish
11 kinds
12 Are pet fish as cockerel, guppy, scalar, catfish the most popular?
13 1
14 Fish
15 Care
16 To keep the fish, you need to have an aquarium with a filter that will supply oxygen
  to the fish. And also periodically feed them and wash the walls of the vessel.
17 1
18 Dogs
19 Buying a dog
20 Where can I buy a dog? Most of all I would like a little puppy who could make friends
  with my pet cat
21 1
22 Dogs
23 Puppies
24 How many puppies are there in his house?
25 1
26 Dogs
27 Cute animals
28 What are the most beautiful breeds of dogs?
29 2
30 Dogs
31 Lost my dog
32 Ive lost my dog.
33 How can I find it now?
34 1
35 Cats
36 A name for a cat
37 What are the best names for cats?
38 1
39 Cats
40 Mother cat with her kittens
41 My cat has many kittens. How should I take care of them?
42 1
```

43 Cats  
44 Different cats  
45 How many different kinds of cats exist?  
46 1  
47 Cats  
48 Black cat  
49 Is it true that a black cat causes problems?  
50 1  
51 Cats, dogs  
52 Friendship between cats and dogs  
53 Can cats and dogs live together?  
54 2  
55 Cats  
56 Benefit  
57 Did you know that cats can improve your health? Cats help calm down and cope with stress.  
58 In a house where cats live, people quarrel and get annoyed less.  
59 1  
60 Cats  
61 Ancient animal  
62 Cats are very ancient animals. They existed on earth almost fifty million years ago  
63 2  
64 Cats  
65 Age  
66 If the age of a cat is translated to a person, fifteen years for a cat is about seventy human years.  
67 A threeyear-old cat is quite an adult animal, like people in their twenties.  
68 1  
69 Cats  
70 big dormice  
71 Cats spend two thirds of the day sleeping. The rest of the time is devoted to self-care.  
72 1  
73 Dogs  
74 Friendly tail  
75 The tail of the helicopter, which the dog wags from side to side, is a sign of friendliness.  
76 1  
77 Dogs  
78 Lifespan  
79 Dogs live an average of 10 to 14 years.  
80 1  
81 Dogs  
82 understanding  
83 When we pet a dog and look into its eyes, the feel-good hormone is released in both humans and dogs.  
84 1  
85 Dogs  
86 unique nose

87 Just as with human fingerprints, no dog nose prints are the same.  
88 1  
89 Dogs  
90 color discrimination  
91 Dogs are not colorblind, but their eyes dont have red receptors. They see shades of  
black and white, as well as blue and yellow.  
92 1  
93 Dogs  
94 Taste  
95 Dogs have about 1,700 taste buds compared to 9,000 in humans.  
96 1  
97 Dogs  
98 loneliness  
99 Dogs are herd animals, they dont like being alone.  
100 13  
101 Dogs  
102 Text about dog  
103 There are moments in our lives when all the people close to us turn away from us. Or  
when you come home and theres no one there. The best solution to this problem is a  
dog.  
104 No one can say how easy it will be **for** you with this decision, but here you have to  
think it over.  
105 Its hard with a dog, especially the first days. A dog is the same child that is sick  
or sad, sometimes you really want to play.  
106 It seems that you took it in vain, did something stupid and hurried. But hurry up **if**  
you immediately give it away or throw it away. The love of a dog also needs to be  
earned.  
107 As soon as you come home, you mentally ask the question, what did he **do**. If the dog  
wags its tail and its happy eyes are the first thing you see, then everything that  
the dog did is no longer important.  
108 A wet nose, a **long** tongue that tries to kiss you, thats what matters.  
109 With each success of your dog, you become happier. Here the dog understood the command  
to sit, lie down, brought a toy, does not pull the leash, asks to go to the  
toilet, and you are happier than he is.  
110 The dog is protection and support. As soon as you see how your dog does not let  
someone near you or barks **if** something is wrong, then you have nothing to fear.  
111 Even the smallest puppy will not leave its owner in trouble.  
112 You need to communicate with the dog, you need to love it and endure its antics many  
times.  
113 But when you suddenly get sick, and your dog does not leave you a single step, then  
you will understand what it was all **for**, and that it was all worth it.  
114 Dogs often rescue drowning people, find children, and help in rescue operations. The  
dog will undoubtedly be your family member and friend.  
115 And how **do** dogs sit **for** many days near the place where they were thrown out? They can  
neither eat nor drink nor move. They are the most loyal and devoted creatures.  
116 14  
117 Cats  
118 domestic cat cat cat

119 These cute animals belong to the genus of mammals, the cat family, which has existed  
 on earth for 10 thousand years. Man domesticated the cat over 6,000 years ago.  
 120 Everyone knows the domestic cat, but there are also wild cats in nature.  
 121 The life span of a domestic cat is generally 12-14 years. On average, the animal  
 weighs about four kilograms. These gentle, fluffy pets are very capricious.  
 122 Cats purr loudly in a calm, peaceful state, hiss and growl when they are angry.  
 123 If the pet is laid on its back, then it feels protected. And if he starts to twitch or  
 wag his tail, then he is angry. If a cat looks into your eyes and meows  
 plaintively, she is begging for something.  
 124 Sensing danger, the animal arches its back in order to appear larger and frighten the  
 enemy  
 125 This cute pet is very independent and independent. The cat walks where she wants, but  
 in the end she always returns to people.  
 126 There are more than forty types of cats, each breed has its own peculiarity. Some  
 representatives have too short legs and no tail. Most domestic cats are a mixture  
 of several breeds.  
 127 A cats eyesight is six times sharper than a humans. Her eyes glow in the dark. These  
 fluffy pets capture scents that humans cannot smell. They recognize other felines  
 by smell.  
 128 They constantly mark their territory to scare off strangers. Cats have excellent  
 hearing. They are very picky eaters. They love to jump and climb. Good hunters.  
 129 Animals are very clean. They like to wash themselves with their paws, licking them.  
 Cats sleep for most of their lives.  
 130 These pets have fur that keeps them warm in cold weather. In summer, the wool  
 partially falls out.  
 131 Cats are very prolific animals. They often have kittens. At what several at once,  
 about three five kittens at a time. One cat can have kittens of different colors.  
 132 Cats are very intelligent creatures, without which it is difficult for a person to  
 imagine his life.  
 133 10  
 134 Fish  
 135 Aquarium fish  
 136 Aquarium fish are diverse in body shape and color. There are those that can easily fit  
 in a teaspoon, some the size of a childs palm.  
 137 Nature endowed them with different colors - red, yellow, blue, green, blue, white,  
 black, some completely multi-colored: striped, spotted, with neon color.  
 138 The same goes for their body shape. Some fish are flat like a leaf, others with round,  
 thick barrels. Some inhabitants of the aquarium have a lush, large caudal fin or  
 a narrow and long one.  
 139 Mostly common aquarium fish feed on food of animal origin, although some species eat  
 plants and cannot develop normally without them. However, they also eat animal  
 food from time to time.  
 140 The most common food for aquarium fish is dry food, which we buy at a regular pet  
 store.  
 141 However, those who wish to diversify the diet of the inhabitants of aquariums should  
 pay attention to small crustaceans - daphnia, cyclops.  
 142 They can also be caught in ponds, or you can buy them in specialized stores. If  
 possible, diversify the food base of your pets. For small aquarium fish,  
 bloodworms, enchitrei, coretra are suitable.

143 They can be easily bred at home.  
 144 All aquarium fish have well developed sense organs. Most of the tactile cells are  
 located in the upper layer of the body.  
 145 In sturgeon, carp and catfish, they are concentrated in additional organs, such as  
 antennae.  
 146 7  
 147 Fish  
 148 My little fish  
 149 Aquarium fish live in my house: goldfish, neons, swordtails, several snails and a  
 small catfish. My parents gave me an aquarium with fish for my birthday.  
 150 In the pet store, we buy fish not only special dry food, but also live food - these  
 are tiny crustaceans and worms. I feed them a little three times a day.  
 151 The fish live in a large rectangular aquarium. At the bottom of the aquarium is a  
 special soil. And at the bottom there are multi-colored glass pebbles and  
 beautiful shells.  
 152 The fish have a large underwater castle and various labyrinths. Real live algae grow  
 in the aquarium, which create the appearance that the fish are in the ocean. By  
 the way, fish also eat live algae, it is good for their health.  
 153 For a comfortable life of the fish, the parents bought a special lighting for the  
 aquarium. And an air filter is attached to the back wall of the aquarium.  
 154 It simultaneously enriches the water with oxygen and filters it so that it is clean  
 and free of debris. The filter needs to be washed periodically.  
 155 About once every few weeks, my dad and I clean the aquarium. This is necessary so that  
 the water does not turn green and is not cloudy.

Файл test: (тестовые запросы)

1 1  
 2 Feeding  
 3 How to feed my little friends who lives with me?  
 4 1  
 5 Many animals  
 6 Can two black cats live together in peace with a small dog?  
 7 1  
 8 Keeping  
 9 How should I care to keep pet fish? do I need a aquarium?  
 10 1  
 11 aquarium  
 12 Who is small lives in an aquarium and swims in the water?  
 13 1  
 14 friend  
 15 Who is man best fluffy friend?  
 16 1  
 17 relation  
 18 Who needs to be treated with kindness and care?  
 19 1  
 20 valerian  
 21 who is fluffy and loves valerian?  
 22 1  
 23 who is it

24 | a small animal, furry and with four legs?  
25 | 1  
26 | who is it  
27 | very small animal with fins, cant walk, lives in water?  
28 | 1  
29 | qqq  
30 | asdq?  
31 | 1  
32 | together  
33 | What animals can live in an apartment?  
34 | 1  
35 | food  
36 | what should you feed your pets?  
37 | 1  
38 | health  
39 | Do cats help with various diseases?  
40 | 1  
41 | saving  
42 | can dogs save people from dangerous situations?  
43 | 1  
44 | like  
45 | what animals do you like the most of all?  
46 | 1  
47 | present  
48 | what kind of fish did they give you yesterday and do you have an aquarium with a lump  
    for them?  
49 | 1  
50 | help  
51 | my master does not feed me, what should I do?  
52 | 1  
53 | thinks  
54 | I am thinking of getting a cat, but I already have a dog, is it dangerous?  
55 | 1  
56 | eating  
57 | Do dogs eat house fish?  
58 | 1  
59 | legal  
60 | Is it legal to have 100 cats at home?  
61 | 1  
62 | allergic  
63 | I am allergic to wool, can I get a pet?  
64 | 1  
65 | Money  
66 | What is the cheapest pet to keep?  
67 | 1  
68 | charity  
69 | today I have a salary, and I will donate a share of it to a cat and dog shelter!  
70 | 1  
71 | choose

```

72 | my daughter asks for a hamster, but I think that a dog or a cat is better, what should
    | I do?
73 | 1
74 | strange behavior
75 | the cat constantly jumps into the aquarium and swims there, what should I do?
76 | 1
77 | mess
78 | when I come home, all things are scattered around the apartment. Was it the fish or
    | the sleeping cat?
79 | 1
80 | no limit
81 | my dog drinks all the water in a second and eats all the food off the table, how to
    | limit it?
82 | 1
83 | no wool
84 | I want an animal without wool, who should I take?
85 | 1
86 | crossing
87 | If you cross a dog and an aquarium, you get a fish with paws?
88 | 1
89 | bites
90 | do dogs and cats bite their owners, how to protect themselves?
91 | 1
92 | health of animals
93 | What are the diseases of pets and how dangerous is this disease for people?
94 | 1
95 | age
96 | how to calculate the age of my fish in human years?
97 | 1
98 | fear
99 | how to overcome fear of dogs on the street?
100 | 1
101 | emotions
102 | Do animals have emotions, can they smile or cry?

```

console input:

```

./prog learn --input dataset/study --output dataset/stats
./prog classify --stats dataset/stats --input dataset/test --output dataset/testOUT

```

Файл testOUT:

```

1 | fish, dogs
2 | dogs, cats
3 | fish
4 | fish
5 | dogs, cats
6 | fish, dogs
7 | cats
8 | fish

```



9	fish
10	fish, dogs
11	cats
12	fish, dogs
13	cats
14	dogs
15	dogs
16	fish
17	dogs
18	dogs
19	fish
20	cats
21	dogs
22	dogs
23	dogs
24	dogs
25	fish, cats
26	dogs
27	dogs
28	cats
29	fish, dogs
30	dogs, cats
31	cats
32	cats
33	dogs
34	dogs

В текущих данных на обучение дано 1586 слов, среди которых 265 относятся к тегу "cats 254 к тегу "dogs 245 - "fish".

Тестовых запросов подано 34. На каждый из запросов программа предложила набор подходящих к нему тегов. Путем анализа запросов и предполагаемых тегов подсчитано: среди предложенных программой тегов 18 из них точно соответствуют действительности, 10 частично удовлетворяют запросу и 5 не удовлетворяют совсем. Таким образом, округляя, можно сказать, что половина тегов предсказана верно, а оставшая половина либо удовлетворяет запросу частично, либо не удовлетворяет совсем.

## 4 Оценка сложности

Произведем некоторую оценку подсчета вероятности для тегов тестовых наборов.

Пусть  $tests$  — количество данных тестовых наборов;

$types$  — общее число тегов;

$words$  — число слов в конкретном тестовом наборе.

Тогда сложностная оценка подсчета вероятностей тегов для всех тестовых наборов -  $O(tests * types * words)$ .

## 5 Выводы

Подводя итог по выполненному курсовому проекту, наивный Байесовский классификатор является одним из самых популярных и простых алгоритмов в машинном обучении. Он имеет ряд своих плюсов и минусов.

### Положительные стороны:

- Классификация, в том числе многоклассовая, выполняется легко и быстро.
- Когда допущение о независимости выполняется, НБА превосходит другие алгоритмы, такие как логистическая регрессия (logistic regression), и при этом требует меньший объем обучающих данных.
- НБА лучше работает с категориальными признаками, чем с непрерывными. Для непрерывных признаков предполагается нормальное распределение, что является достаточно сильным допущением.

### Отрицательные стороны:

- Если в тестовом наборе данных присутствует некоторое значение категориального признака, которое не встречалось в обучающем наборе данных, тогда модель присвоит нулевую вероятность этому значению и не сможет сделать прогноз. Это явление известно под названием «нулевая частота» (zero frequency). Данную проблему можно решить с помощью сглаживания. Одним из самых простых методов является сглаживание по Лапласу (Laplace smoothing).
- Хотя НБА является хорошим классификатором, значения спрогнозированных вероятностей не всегда являются достаточно точными. Поэтому не следует слишком полагаться на результаты, возвращенные методом *predict\_proba*.
- Еще одним ограничением НБА является допущение о независимости признаков. В реальности наборы полностью независимых признаков встречаются крайне редко.

## Список литературы

- [1] Томас Х. Кормен, Чарльз И. Лейзерсон, Рональд Л. Ривест, Клиффорд Штайн. *Алгоритмы: построение и анализ, 2-е издание*. — Издательский дом «Вильямс», 2007. Перевод с английского: И. В. Красиков, Н. А. Орехова, В. Н. Романов. — 1296 с. (ISBN 5-8459-0857-4 (рус.))