Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

«Санкт-Петербургский политехнический университет Петра Великого»

Институт компьютерных наук и технологий

Высшая школа интеллектуальных систем и суперкомпьютерных технологий

**Расчётное задание 1**

по дисциплине «Теория вероятностей и Математическая статистика»

**Угадывание слов**

Выполнил

Студент группы 3530901/80004 Иванов К. А.

\_\_\_\_\_\_\_\_\_\_\_

Преподаватель Никитин К.В.

\_\_\_\_\_\_\_\_\_\_\_

Санкт-Петербург 2020

**Описание**

Ведущий загадывает слово. Затем в начале каждого хода ведущий сообщает некоторую информацию о слове:

- он выбирает наугад одну из букв в слове и сообщает ее без уточнения ее позиции в слове;

- он выбирает наугад одну из букв в слове и сообщает ее позицию и определенную информацию о ней (гласная/согласная/соединительная, звонкая/глухая, заглавная/строчная);

Игрок на основании информации, полученной от ведущего, строит гипотезы о том, какое это слово. Количество букв в слове в общем случае неизвестно.

1. Попытаться вначале угадать слово самостоятельно, используя как можно меньше подсказок.

2. Используя предоставленную базу слов, вычислить для каждого из слов последовательно вероятности того, что загадано именно это слово. Определить количество подсказок, начиная с которого можно с определенной уверенностью утверждать о том, какое слово загадано.

Построить распределения числа слов-кандидатов как функцию от номера подсказок. По логике количество слов-кандидатов с каждым шагом должно уменьшаться. Загаданное слово принадлежит одному из классов:

Город России (cities\_rus.txt)

Страна мира (countries.txt)

Столица (capitals.txt)

Русское имя (names\_all.txt).

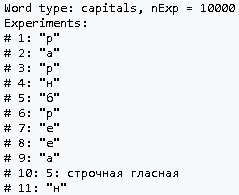
Река (rivers.txt)

Русское существительное (имя нарицательное) (russian\_nouns.txt)

В скобках выше приведены названия текстовых файлов, в которых приведены все возможные варианты существительных для каждого из классов.

**Исходные данные**

Ниже приведены первые несколько строк файла с исходными данными.



Используется набор слов – столицы мира, количество опытов равно 10000. Далее приводятся результаты каждого из опытов. С 1 по 9 опыт сообщаются буквы р, а…В 10-м опыте сообщается, что 5-я буква – строчная гласная и т.д.

**Задание**

1а. После каждого опыта необходимо вычислить ряд распределения апостериорных вероятностей гипотез – какое слово загадано. Представить соответствующие результаты визуально на графике в форме изменения с течением опытов диаграмм распределений вероятностей гипотез. Сузить количество гипотез можно по аналогии с п. 3a. 9 Зам. В данной задаче количество гипотез равно количеству слов в соответствующем файле со словами.

1b. Определять после каждого опыта, какое слово имеет наибольшую вероятность. Визуализировать эволюцию изменения наиболее вероятного слова.

1с. Построить зависимость числа превалирующих гипотез от числа проведенных опытов.

2a. Рассмотреть каждую букву в слове по отдельности и в качестве гипотез выдвигать то, что это за буква (сам символ, заглавная/строчная, гласная/согласная и т.д.). Для гипотезы о том, что за символ общее количество гипотез равно общему количеству возможных символов, для гипотез о классах буквы количество гипотез равно количеству классов. Вычислить для каждой из букв распределения вероятностей гипотез после каждого опыта. Представить результаты визуально по аналогии с п. 1а.

2b. Определить для каждой буквы наиболее вероятную гипотезу на каждом шаге и визуализировать эволюцию этой гипотезы.

2с. Объединить результаты для всех букв, получить наиболее вероятное слово и сравнить его с полученным п.1. Провести анализ сравнения. Зам. Возможной проблемой в данном пункте будет то, что общее количество букв не задано. Возможным способом ее решения является отбрасывание всех букв, у которых номер в слове больше, чем номер последней буквы, сообщенной ведущим.

3a. Определить приближенно частоту тех или иных сообщаемых символов. Исходя из этого можно оценить примерно, сколько каждого из символов в слове. Затем сопоставьте эту информацию с тем, что ведущий говорит о позициях букв и их типе. В результате можно сузить количество возможных слов, используемых в качестве гипотез.

3b. Составьте для каждого слова из базы теоретические профили, содержащие в себе некоторую статистику – количество символов каждого типа, позиция с учетом типа символа. Далее из п. 3а определите аналогичный экспериментальный профиль.

3c. Сопоставить теоретические профили с экспериментальным и найдите их наиболее правдоподобное соответствие в форме слова (слов). Сравнить с полученными результатами в п. 1 и 2. Провести анализ сравнения.

3d. Привести графики изменения экспериментальных профилей для различного количества опытов.

**Теория связанная с заданием**

Формула полной вероятности:

Формула Байеса:

1.Полученная информация на каждом шаге о загаданном шаге сужает количество гипотез. В связи с этим условная вероятность будет варьироваться в зависимости от информации, сказанной ведущим. Вероятность будет равна 1, если назван номер буквы и ее классы.

Следующая вероятность, имеет место, если сказана просто буква:

2. Далее необходимо рассмотреть каждую букву в слове по отдельности, в качестве гипотез были взяты все возможные вариации буквы, которая может находится на какой-либо позиции. Общее количество гипотез равно **n \* m**, где n – количество букв в слове, m – мощность алфавита.

**Р(Hi/A)= 1/𝑛** , если названная ведущим буква совпадает с гипотезой;

**Р(Hi/A)= 1/(𝑚∗𝑛),** если названная ведущим буква не совпадает с гипотезой;

**Р(Hi/A) = 1**, если названа позиция буквы и ее класс.

**Описание логики**

1. Полученная информация на каждом шаге о загаданном слове сужает количество гипотез. В связи с этим условная вероятность, варьироваться в зависимости от информации, сказанной ведущим:

a. если сказана просто буква, **Р(Hi/A)** будет равна количеству названных букв в слове, разделенных на количество букв в слове(гипотезе).

b. если назван номер буквы и ее классы, то если указанная буква в гипотезе того же класса, то вероятность такого условия будет равна 1

2. Далее необходимо рассмотреть каждую букву в слове по отдельности, в качестве гипотез были взяты все возможные вариации буквы, которая может находиться на какой-либо позиции: итого их получилось **n \* m**, где n – количество букв в слове, m – мощность алфавита. В таком случае условная вероятность будет иметь вид:

a. если названная буква ведущим совпадает с гипотезой, то **Р(Hi/A)** будет равна 1/n

b. если названная буква ведущим не совпадает с гипотезой, то **Р(Hi/A)** будет равна 1/ (n \* m), т.к. вероятность того, что не смотря на букву, названную ведущим на данной позиции все равно будет стоять буква гипотезы будет равна вероятности того, что ведущий имел ввиду другой номер буквы, т.е. 1/n, и вероятности того, что на данном месте стоит именно эта буква, т.е. 1/m

c. в случае, если названа позиция буквы и ее класс, то **Р(Hi/A)** будет равно 1

**Решение**

Начиная решать задачу, сразу естественно проситься интуитивное разделение слов на две категории:

1.Гласные буквы – {'о', 'ы', 'я', 'у', 'э', 'ё', 'е', 'а', 'и', 'ю'}

2.Согласные буквы – {'г', 'н', 'б', 'д', 'й', 'ж', 'в', 'р', 'л', 'м', 'з'}

Определяем удобные для работы функции, которые могут нам убрать лишние параметры из работы, например такие как, номер которым идет значение и кавычки. Также функции помогающие взять все подсказки, и соответственно наши главные данные буквы и их частоту, по которым и будем аккуратно искать и перебирать для угадывания правильного слова.

В главной части программы, читаем информацию из файла, и переопределяем все данные в удобный формат таблиц. Которую определим и будем заполнять по ходу продвижения по значениям в файле:

1.Origin – строка из файла

2.reg – заглавная или строчная

3.type – гласная, согласная

4.sound – звонкая или нет

5.idx – номер буквы в подсказке

6.char – наша буква или возможные буквы из подсказки

Также две переменные words, hypothesizes – список слов и гипотезы (предположения, на основании которых и будем делать выводы)

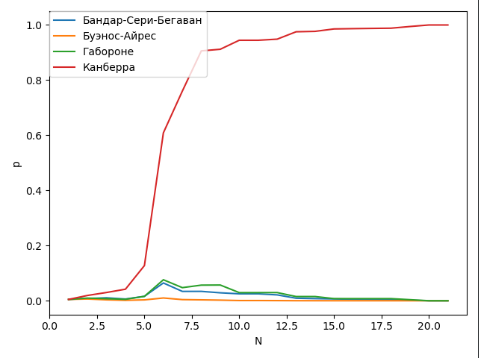
Решения проделаны с помощью Я.П. Python 3, и нескольких библиотек:



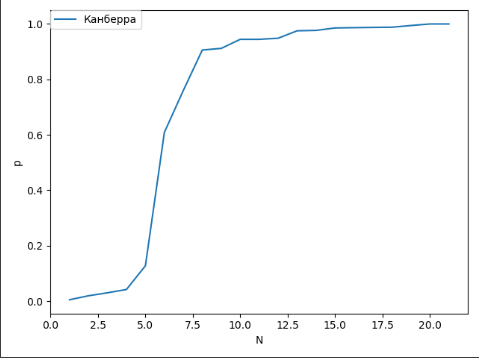
без каждой из которых было бы просто невозможно выполнить множество задач, в том числе и для данной задачи нам понадобятся следующие библиотеки, которые предоставляют базу важных для математики функций; сбором и использовании статистических данных; графическом представлении, соответственно.

По ходу выполнения программы будем смотреть на распределение букв, и тем самым делать некоторые гипотезы, накапливая статистические данные, гипотезы подскажут нам какое слово было загадано.

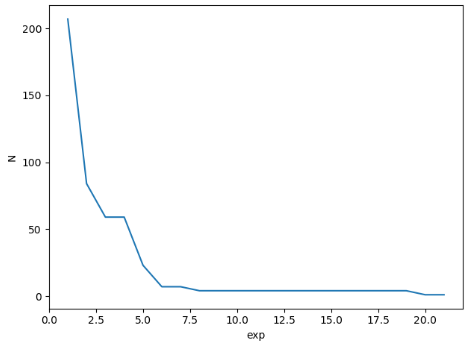
**Графические данные**



**Рис. 1** Предположения о слове с все более глубоким погружением в данные

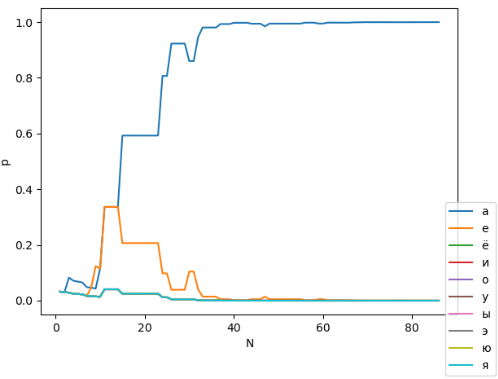


**Рис. 2** Изменения наиболее подходящего слова, в процессе работы



**Рис. 3** Зависимость гипотез от числа проведенных опытов

**Рис. 4** Частота распределения для буквы К

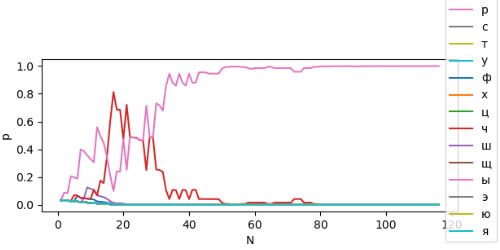


**Рис. 5** Частота распределения для буквы А

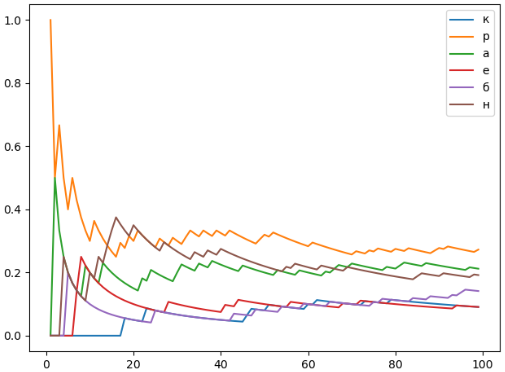
**Рис. 6** Частота распределения для буквы Н

**Рис. 7** Частота распределения для буквы Б

**Рис. 8** Частота распределения для буквы Е



**Рис. 9** Изменение вероятности буквы Р



**Рис. 10** Изменение вероятностей для всех букв