Случайные процессы. Практическое задание 1

- Дедлайн **25 сентября 23:59**.
- Внимательно прочтите правила.
- Обратите внимание на рекомендации по выполнению в конце этого файла.

В Британской империи в Викторианскую эпоху (1837—1901) было обращено внимание на вымирание аристократических фамилий. В связи с этим в своей статье в The Educational Times в 1873 году Гальтон поставил вопрос о вероятности вымирания фамилии. Решение этого вопроса нашел Ватсон и вместе в 1874 году они написали статью "On the probability of the extinction of families". На сайте wikitree.com в свободно распространяемом формате собрано большое количество данных о родословных различных людей. В коллекции есть как люди, жившие во времена поздней античности, так и наши современники. На основе некоторой части этих данных вам предстоит провести исследование о вымирании фамилий.

Вам предоставляются несколько файлов, в которых содержатся данные о некоторых родословных. Вам предстоит проводить исследование на нескольких из этих файлов (каких именно, см. в таблице). Формат файлов следующий:

```
generation \t name \t gender \t birthday \t deathdate \t parents
\t siblings \t spouses \t children
```

Эти данные означают номер поколения, фамилию, пол, дату рождения, дату смерти, родителей, братьев и сестер, супруг, детей соответственно. Если какая-то характеристика неизвестна (кроме номера поколения и фамилии), вместо нее ставится пустая подстрока. Если каких-то характеристик несколько, то они разделены через “;”. Все люди представлены некоторым идентификатором `<id>`, который соответствует адресу `http://www.wikitree.com/wiki/<id>`. Например, идентификатор `Romanov-52` соответствует адресу `http://www.wikitree.com/wiki/Romanov-52`. В файле родословные отделяются друг от друга пустой строкой. Для облегчения вашей работы мы предоставляем вам код, который считывает данные из этого файла и преобразует их в список ветвящихся процессов. Каждый ветвящийся процесс содержит список списков, в каждом из которых содержатся все люди из соответствующего поколения. Чтобы лучше понять требуемое, посмотрите пример работы с кодом, который находится в том же файле. Обратите внимание, что одни и те же родословные могут попасть в разные файлы. В таком случае их следует считать разными.

В предоставленных данных в каждой родословной для каждого мужчины на следующем поколении содержатся все его дети, которые были указаны на сайте. Для женщин дети в данной родословной не указаны. Это связано с тем, что женщины обычно меняют свою фамилию, когда выходят замуж, тем самым, они переходят в другую ветку. С точки зрения ветвящихся процессов, нужно иметь в виду, что если у мужчины родилось 3 мальчика и 4 девочки, то у него 3 потомка как продолжателя фамилии.

Ваша задача — исследовать процесс вымирания фамилий на основе предложенных данных.

Задания: (Первые два задания являются техническими, однако их необходимо выполнить для лучшего понимания задачи)

1. Считайте данные с помощью предложенного кода. Посчитайте количество родословных.

   В имеющихся данных очень много людей, про которых известно лишь то, что они когда-то существовали. Обычно их фамилия неизвестна (вместо фамилии у них может стоять, к примеру, B-290), а у некоторых из них неизвестен даже пол, не говоря уже о родителях и детях. Такие данные стоит удалить.

   Удалите все процессы, состоящие только из одного поколения (в котором, естественно, будет только один человек). Сколько осталось процессов?

2. **(0.5 балла)** Для лучшего понимания задачи и предложенных данных посчитайте следующие характеристики: минимальное, максимальное и среднее число поколений в роду, год рождения самого старого и самого молодого человека, среднюю продолжительность жизни. Постройте гистограмму зависимости количества поколений в родословной от количества родословных (по оси x — количество поколений в роду, по оси y - количество таких родов). На следующем графике отложите на временной оси года рождения всех людей. Для этого наносите точки с нулевой y-координатой, прозрачность точки выставляйте не более 0.2.

3. **(2 балла)** Определите закон размножения в двух случаях: а) считая все процессы одним большим процессом, у которого неизвестно несколько первых поколений; б) считая все процессы разными. Во втором случае должен получиться набор законов размножения. Конкретный способ поиска вероятностного закона вы должны выбрать сами и обосновать его. В качестве базового варианта рекомендуется приближать неизвестный закон размножения пуассоновским распределением и дискретным распределением в общем виде (лучше — оба варианта) на основе оценки максимального правдоподобия (в качестве критерия проверки близости теоретического и эмпирического законов распределения можно использовать критерий хи-квадрат). Гарантируется, что отец указан у всех людей, кроме тех, кто принадлежит нулевому поколению, причем отец стоит первым из родителей.

   Постройте график зависимости вероятности по выбранному закону размножения (в пункте а) от количества потомков.

   Как оценивать количество мужчин, у которых нет детей-мальчиков? Из посчитанного общего количества таких мужчин в предложенных данных нужно вычесть число таких мужчин в последнем поколении каждого рода (возможно, данные неполные), а так же исключить тех, кто родился не раньше 1950 года. Если вы используете свой способ, необходимо четко его описать.

   Автоматическая проверка (доп. 2 балла). Посчитайте параметр пуассоновского распределения (в пункте а), используя предложенную выше схему оценки количества мужчин, у которых нет детей.

   Заполните форму https://goo.gl/forms/jt3koMS03XpKisDg2

4. **(2 балла)** В данном пункте используйте только закон размножения из пункта 3а. Некоторым образом промоделируйте процесс "назад", то есть до единого предка, основываясь на найденном в предыдущем пункте законе размножения. Это

можно сделать, например, простой генерацией случайных величин из закона размножения. Опишите этот процесс подробно. Более сложный вариант предполагает использование данных о годах жизни членов объединенной родословной. Получилось ли дойти до общего предка, или процесс стал “раздуваться”? С чем это может быть связано (с точки зрения случайных процессов)?

5. **(3 балла)** Некоторым образом промоделируйте процесс ”вперед”, основываясь на найденном в пункте 3 закона размножения. Сколько мужчин и фамилий будет через 100, 200, 300 лет? Базовым вариантом выполнения данного пункта будет использование оценки количества поколений за какой-либо большой ($\leqslant 100$) промежуток времени. Более сложный вариант предполагает учёт продолжительности жизни и использование регрессионного анализа. Сделайте моделирование двумя способами — на основе общего закона размножения и считая, что у каждого рода свой закон размножения. В каждом из случаев посчитайте зависимость количества мужчин и фамилий от поколения, начиная с последнего известного поколения в каждом роду. Постройте графики для этих зависимостей. Если вы используете две (или более) модели закона размножения (например, пуассоновскую и общую дискретную), то соответствующие зависимости для разных моделей изобразите на одном графике. Разные зависимости (общий закон или для каждого свой, люди или фамилии) следует изобразить на разных графиках. В каком случае процесс вырождается?

   В данном пункте рекомендуется написать функцию, которая на заданных процессах сгенерирует еще несколько поколений в соответствии с заданным законом (или законами) размножения и возвратит количество человек и фамилий в каждом сгенерированном поколении.

6. **(2 балла)** Посчитать вероятность вырождения для общего закона размножения и вероятность вырождения для каждого процесса в отдельности. Искать корень уравнения можно приближенно. Какая доля процессов имеет вероятность вырождения меньше 0.5? Изобразите все полученные вероятности вырождения на графике. Для этого рисуйте точки с нулевой y-координатой, прозрачность точки выставляйте не более 0.2.

7. **(3 балла)** По результатам каждого пункта исследования сделайте подробный вывод.

Рекомендации по выполнению задания:

1. Не нужно копировать выданный код в ноутбук. Вместо этого можно импортировать из него все необходимое:

   `from BranchingProcess import Person, BranchingProcess, read_from_files`.

2. В третьем пункте удобно воспользоваться `collection.Counter`, который для данного списка считает, сколько раз в нем встречается каждое значение.

3. В пятом пункте для полного копирования списка процессов (не по ссылке) стоит воспользоваться `copy.deepcopy`.

4. В пятом пункте стоит следить за оперативной памятью. Если ее не хватает, то считайте только до 100 лет.

5. На небыстром ноутбуке весь код в базовом варианте выполняется за 6 минут.