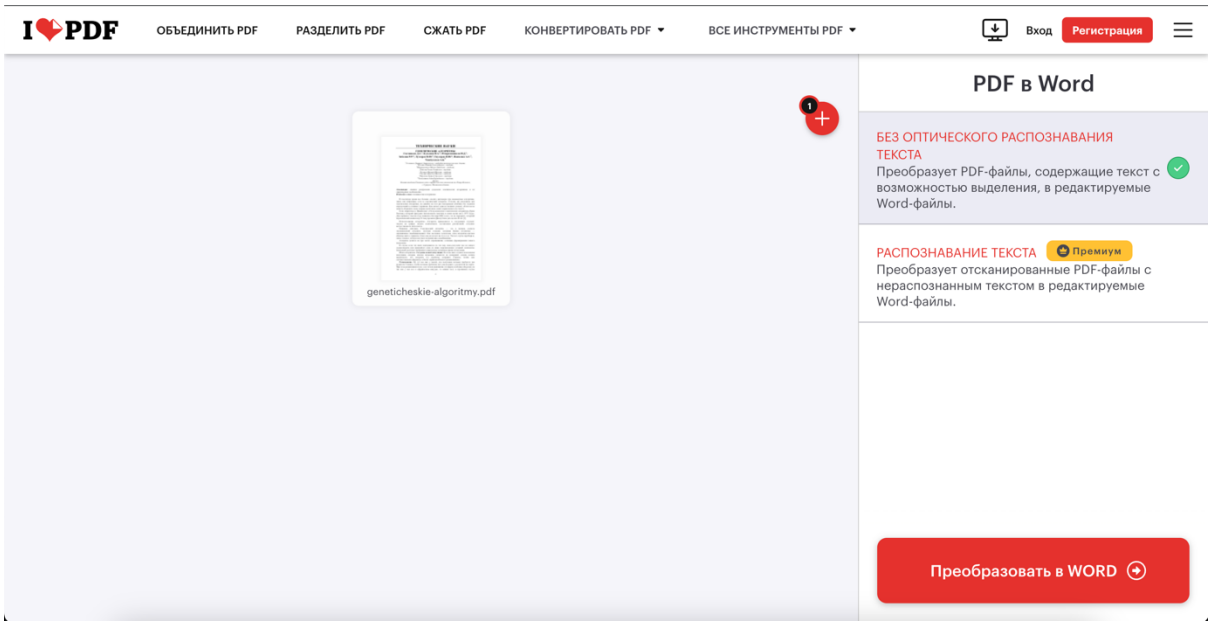
Задание 2.

Создание обучающего датасета состоит из обработки текстовых данных, с целью подведения их к формату, пригодному для использования в программных средствах. Для этого необходимо произвести следующие действия:

– извлечь текстовую информацию из найденных работ;
– совместить тексты в один файл .DOC;
– конвертировать полученный файл в формат .tsv;

Итоговый файл прикладывается к отчету в виде приложения.

Для создания датасета необходимо привести файлы статей к текстовому формату. Если статьи сохранены в формате .doc/docx, переходите сразу к следующему шагу. Иначе воспользуйтесь сервисами конвертации pdf в word. Например, https://www.ilovepdf.com/ru/pdf_to_word.



Далее необходимо очистить файл от лишних разделов: шапки, аннотации, ключевых слов и списка источников.

ТЕХНИЧЕСКИЕ НАУКИ

ГЕНЕТИЧЕСКИЕ АЛГОРИТМЫ

Свечников Д.А.[1], Кузьмин Н.А.[2], Мтиралишвили М.Д.[3], Забелин Р.Р.[4], Лузгарев В.Ю.[5], Гнусарев Я.Ю.[6], Панченко А.С.[7], Чембулатов А.Б.[8]

[1]Свечников Дмитрий Анатольевич - кандидат технических наук, доцент;
[2]Кузьмин Никита Александрович - студент;
[3]Мтиралишвили Михаил Давидович - студент;
[4]Забелин Роман Романович - студент;
[5]Лузгарев Валерий Юрьевич - студент;
[6]Гнусарев Ярослав Юрьевич - студент;
[7]Панченко Андрей Сергеевич - студент;
[8]Чембулатов Артур Борисович - студент,
филиал
Военная академия Ракетных войск стратегического назначения им. Петра Великого,
г. Серпухов, Московская область

Аннотация: *статья раскрывает сущность генетических алгоритмов и их структурные особенности.*
Ключевые слова: *генетические алгоритмы.*

В последнее время все больше «ходят» разговоры про новомодные алгоритмы, такие как нейронные сети и генетический алгоритм. Сегодня мы расскажем про генетические алгоритмы, но давайте на этот раз постараемся обойтись без заумных определений и сложных терминов. Как сказал один из великих ученых: «Если вы не можете объяснить свою теорию своей жене, ваша теория ничего не стоит!»

Если обратиться к Википедии: «Отец-основатель генетических алгоритмов Джон Холланд, который придумал использовать генетику в своих целях аж в 1975 году». Для справки в этом же году появился Альтаир 8800, и нет, это не террорист, а первый персональный компьютер. К тому времени Джону было уже целых 46 лет [1].

Использование алгоритма. Алгоритм применяется в следующих случаях: задачи на графы; задачи компоновки; составление расписаний; создание искусственного интеллекта.

Принцип действия. Генетический алгоритм — это в первую очередь эволюционный алгоритм, другими словами, основная фишка алгоритма — скрещивание (комбинирование). Как несложно догадаться, идея алгоритма наглым образом взята у природы, благо она не подаст на это в суд. Так вот, путем перебора и, самое главное, отбора получается правильная «комбинация».

Алгоритм делится на три части: скрещивание; селекция; формирование нового поколения.

В случае если эти шаги повторяются до тех пор, пока результат нас не начнет удовлетворять или произойдет одно из наших перечисленных условий: количество поколений достигает выбранного максимума; исчерпано время на мутацию.

Шаги алгоритмов. **Создание новой популяции**. На этом шаге создается начальная популяция, которая, вполне возможно, окажется не кошерной, однако велика вероятность, что алгоритм эту проблему исправит. Главное, чтобы они соответствовали «формату» и были «приспособлены к размножению».

Размножение. Ну тут все как у людей, для получения потомка требуется два родителя. Главное, чтобы потомок (ребенок) мог унаследовать у родителей их черты. При этом размножаются все, а не только выжившие (эта фраза особенно абсурдна, но так как у нас все в сферическом вакууме, то можно все), в противном случае

выделится один альфа-самец, гены которого перекроют всех остальных, а нам это принципиально неприемлемо.

Мутации. Мутации схожи с размножением, из мутантов выбирают некое количество особей и изменяют их в соответствии с заранее определенными операциями.

Отбор. Тут начинается самое сладкое, мы начинаем выбирать из популяции долю тех, кто «пойдет дальше». При этом долю «выживших» после нашего отбора мы определяем заранее руками, указывая в виде параметра. Как ни печально, остальные особи должны погибнуть.

Практическая часть

Наше уравнение: a+2b+3c+4d=30

Вы наверно уже подозреваете, что корни данного уравнения лежат на отрезке [1;30], поэтому мы берем 5 случайных значений a, b, c, d (Ограничение в 30 взято специально для упрощения задачи).

Итак, у нас есть первое поколение:

1. (1,28,15,3);
2. (14,9,2,4);
3. (13,5,7,3);
4. (23,8,16,19);
5. (9,13,5,2).

Для того чтобы вычислить коэффициенты выживаемости, подставим каждое решение в выражение. Расстояние от полученного значения до 30 и будет нужным значением.

1. |114-30|=84;
2. |54-30|=24;
3. |56-30|=26;
4. |163-30|=133;
5. |58-30|=28.

Меньшие значения ближе к 30, соответственно они более желанны. Получается, что большие значения будут иметь меньший коэффициент выживаемости. Для создания системы вычислим вероятность выбора каждой (хромосомы). Но решение заключается в том, чтобы взять сумму обратных значений коэффициентов, и исходя из этого вычислять проценты. (P.S. 0.135266 — сумма обратных коэффициентов)

1. (1/84)/0.135266 = 8.80%;
2. (1/24)/0.135266 = 30.8%;
3. (1/26)/0.135266 = 28.4%;
4. (1/133)/0.135266 = 5.56%;
5. (1/28)/0.135266 = 26.4%.

Список литературы

1. Википедия. Джон Холланд. [Электронный ресурс]. Режим доступа:

В последнее время все больше «ходят» разговоры про новомодные алгоритмы, такие как нейронные сети и генетический алгоритм. Сегодня мы расскажем про генетические алгоритмы, но давайте на этот раз постараемся обойтись без заумных определений и сложных терминов. Как сказал один из великих ученых: «Если вы не можете объяснить свою теорию своей жене, ваша теория ничего не стоит!»

Если обратиться к Википедии: «Отец-основатель генетических алгоритмов Джон Холланд, который придумал использовать генетику в своих целях аж в 1975 году». Для справки в этом же году появился Альтаир 8800, и нет, это не террорист, а первый персональный компьютер. К тому времени Джону было уже целых 46 лет [1].

Использование алгоритма. Алгоритм применяется в следующих случаях: задачи на графы; задачи компоновки; составление расписаний; создание искусственного интеллекта.

Принцип действия. Генетический алгоритм — это в первую очередь эволюционный алгоритм, другими словами, основная фишка алгоритма — скрещивание (комбинирование). Как несложно догадаться, идея алгоритма наглым образом взята у природы, благо она не подаст на это в суд. Так вот, путем перебора и, самое главное, отбора получается правильная «комбинация».

Алгоритм делится на три части: скрещивание; селекция; формирование нового поколения.

В случае если эти шаги повторяются до тех пор, пока результат нас не начнет удовлетворять или произойдет одно из ниже перечисленных условий: количество поколений достигает выбранного максимума; исчерпано время на мутацию.

Шаги алгоритмов. **Создание новой популяции**. На этом шаге создается начальная популяция, которая, вполне возможно, окажется не кошерной, однако велика вероятность, что алгоритм эту проблему исправит. Главное, чтобы они соответствовали «формату» и были «приспособлены к размножению».

Размножение. Ну тут все как у людей, для получения потомка требуется два родителя. Главное, чтобы потомок (ребенок) мог унаследовать у родителей их черты. При этом размножаются все, а не только выжившие (эта фраза особенно абсурдна, но так как у нас все в сферическом вакууме, то можно все), в противном случае

выделится один альфа-самец, гены которого перекроют всех остальных, а нам это принципиально неприемлемо.

Мутации. Мутации схожи с размножением, из мутантов выбирают некое количество особей и изменяют их в соответствии с заранее определенными операциями.

Отбор. Тут начинается самое сладкое, мы начинаем выбирать из популяции долю тех, кто «пойдет дальше». При этом долю «выживших» после нашего отбора мы определяем заранее руками, указывая в виде параметра. Как ни печально, остальные особи должны погибнуть.

Практическая часть

Наше уравнение: a+2b+3c+4d=30

Вы наверно уже подозреваете, что корни данного уравнения лежат на отрезке [1;30], поэтому мы берем 5 случайных значений a, b, c, d (Ограничение в 30 взято специально для упрощения задачи).

Итак, у нас есть первое поколение:

1. (1,28,15,3);
2. (14,9,2,4);
3. (13,5,7,3);
4. (23,8,16,19);
5. (9,13,5,2).

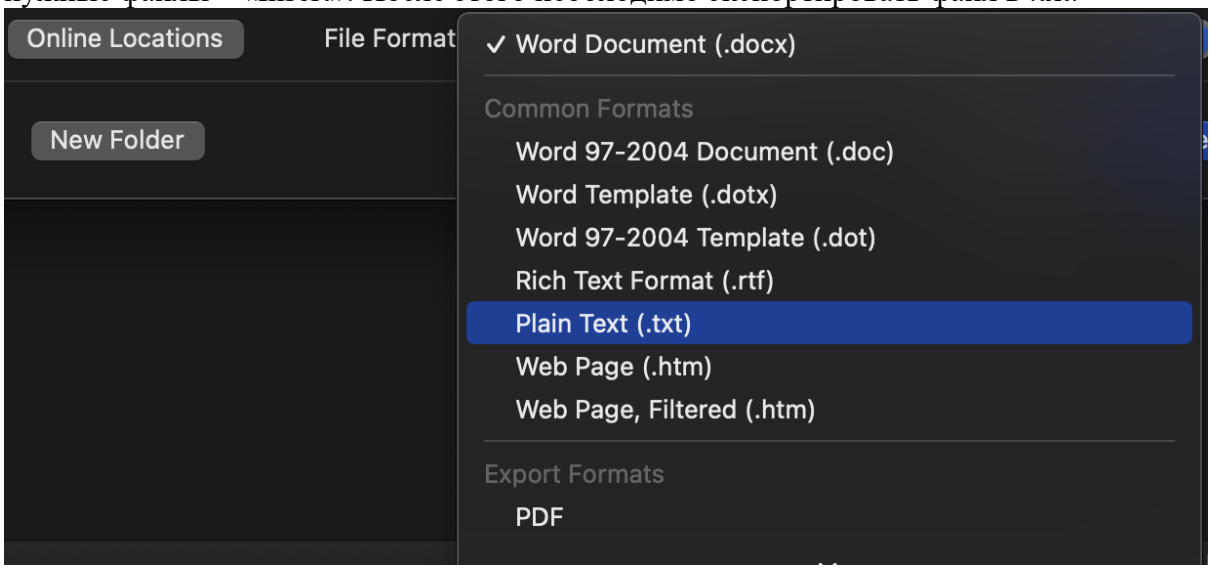
Для того чтобы вычислить коэффициенты выживаемости, подставим каждое решение в выражение. Расстояние от полученного значения до 30 и будет нужным значением.

1. |114-30|=84;
2. |54-30|=24;
3. |56-30|=26;
4. |163-30|=133;
5. |58-30|=28.

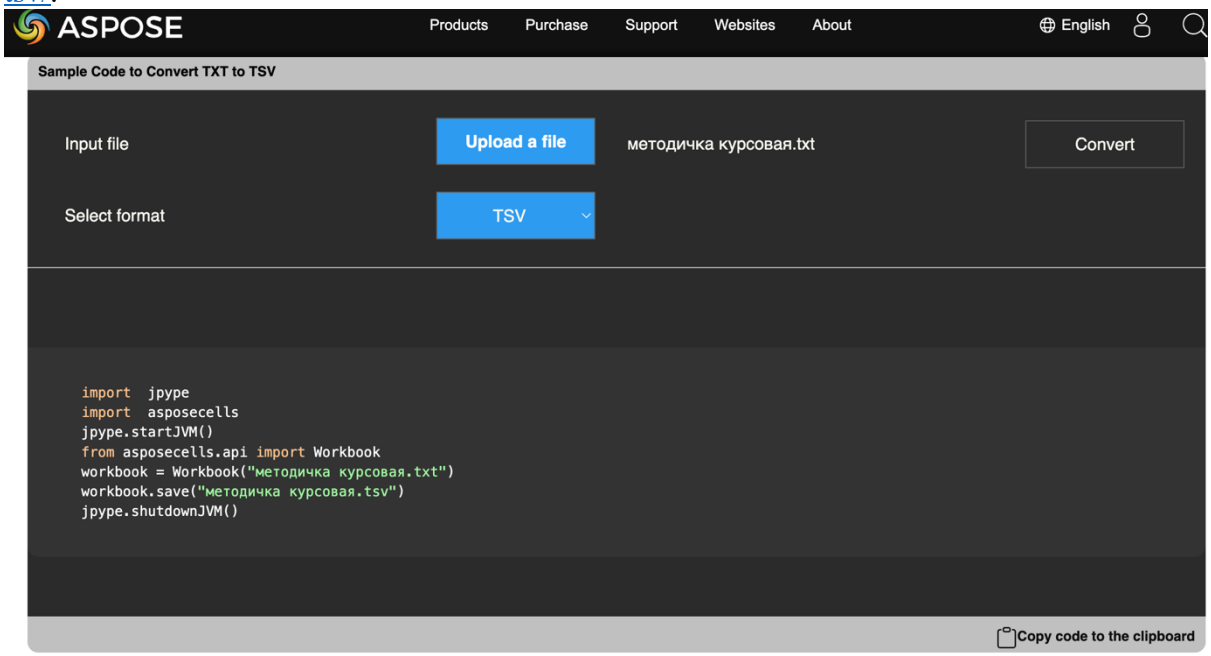
Меньшие значения ближе к 30, соответственно они более желанны. Получается, что большие значения будут иметь меньший коэффициент выживаемости. Для создания системы вычислим вероятность выбора каждой (хромосомы). Но решение заключается в том, чтобы взять сумму обратных значений коэффициентов, и исходя из этого вычислять проценты. (P.S. 0.135266 — сумма обратных коэффициентов)

1. (1/84)/0.135266 = 8.80%;
2. (1/24)/0.135266 = 30.8%;
3. (1/26)/0.135266 = 28.4%;
4. (1/133)/0.135266 = 5.56%;
5. (1/28)/0.135266 = 26.4%.

После того, как данная процедура будет произведена со всеми файлами статей, необходимо объединить их в один файл. «Вставка» - «Текст из файла» - выделить нужные файлы – «Insert». После этого необходимо экспортировать файл в .txt.



Теперь нужно преобразовать получившийся word файл в формат .tsv. Для этого можно воспользоваться сервисом https://products.aspose.com/cells/python-java/conversion/txt-to-tsv/.



Скачать получившийся файл и сохранить. Он будет представлять обучающий датасет для модели.