Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Отчет

по заданию №2
«Работа с регулярными выражениями»

по дисциплине «Компьютерная лингвистика»

Автор: Лакиза Александр Николаевич

Факультет: ИКТ

Группа: К3242

Преподаватель: Чернышева Анастасия Владимировна

Санкт-Петербург 2021

Цель работы: ознакомиться с регулярными выражениями в языке программирования Python. Решить 10 мини-заданий и очистить корпус, собранный в предыдущем задании

Ссылка на исходный код и json файл с корпусом: https://github.com/alexanderlakiza/cs224
Вся работа содержится в файле **task2.ipynb**, очищенный корпус хранится в файле **corpus.json**. Я очень подробно расписывал ход выполнения работы в юпитерской тетрадке (показывая все примеры и все операторы, которые использовал в каждом задании), так что в отчёте не стал приводить примеры по каждому мини-заданию

Ход работы:

Задание 2.1

Необходимо выполнить следующие 10 мини-заданий:

1. Напишите регулярное выражение, которое возвращает список первых двух букв каждого слова строки. Обратите внимание на работу с дефисом.
2. Напишите регулярное выражение, которое выбирает из строки все слова, в которых строго больше 3 символов.
3. Напишите регулярное выражение, которое заменит все подстроки, обозначающие время (только время, не даты), в строке на TBD.
4. Напишите регулярное выражение, которое заменяет произвольное количество пробельных символов внутри строки на один пробел.
5. Напишите регулярное выражение, которое удаляет идущие подряд повторы. Одно слово из группы должно остаться.
6. Напишите регулярное выражение, которое определяет, что подстрока является адресом электронной почты.
7. Напишите регулярное выражение, которое возвращает список аббревиатур в строке.
8. Напишите регулярное выражение, которое разделяет текст на предложения.
9. Напишите регулярное выражение, которое определяет, что строка является номером российского мобильного телефона любого оператора.
10. Напишите регулярное выражение, которое проверяет, что все предложения в строке начинаются с заглавной буквы.

2.1.1. Напишите регулярное выражение, которое возвращает список первых двух букв каждого слова строки. Обратите внимание на работу с дефисом

```
1. re.findall(r'\b\w\w', text)
```

Если мы хотим показать первые две буквы лишь тех слов, которые состоят из двух и более букв

```
1. re.findall(r'\b\w{1,2}', text)
```

Если мы хотим показать еще и слова, состоящие из одной буквы

2.1.2. Напишите регулярное выражение, которое выбирает из строки все слова, в которых строго больше 3 символов.

```
1. re.findall(r'\S{4,}', text)
```

2.1.3. Напишите регулярное выражение, которое заменит все подстроки, обозначающие время (только время, не даты), в строке на TBD.

```
pattern = r'((?:[01]\d|2[0-3])\:(?:[0-5]\d)(?:\:[0-5]\d)?)'
newsubtext1 = re.sub(pattern, "TBD", subtext1)
```

2.1.4. Напишите регулярное выражение, которое заменяет произвольное количество пробельных символов внутри строки на один пробел.

```
pattern = r'\s{1,}'
newsubtext2 = re.sub(pattern, ' ', subtext2)
```

2.1.5. Напишите регулярное выражение, которое удаляет идущие подряд повторы. Одно слово из группы должно остаться.

```
re.sub(r'\b([^\W\d_]+)(\s+\1)+\b', r'\1', subtext3)
```

2.1.6. Напишите регулярное выражение, которое определяет, что подстрока является адресом электронной почты.

```
re.findall(r'\S+@\w+\.\w+', subtext4)
```

2.1.7. Напишите регулярное выражение, которое возвращает список аббревиатур в строке.

```
re.findall(r'[А-ЯA-Z]+\b', subtext5)
```

2.1.8. Напишите регулярное выражение, которое разделяет текст на предложения.

```
re.split(r'(?<=\w[.?!]\s)', text)
```

2.1.9. Напишите регулярное выражение, которое определяет, что строка является номером российского мобильного телефона любого оператора.

```
re.findall(r'(?:\+7|8)\d{10}', subtext6)
```

2.1.10. Напишите регулярное выражение, которое проверяет, что все предложения в строке начинаются с заглавной буквы.

```
re.findall(r'(?:[.!?]\s[A-ZA-Я])|(?:^[A-ZA-Я])', text)
```

Задание 2.2

Необходимо очистить собранный в предыдущем задании корпус текстов от ненужных символов, приведя все данные к одному формату (например, даты).

Сначала я записал весь корпус в одну строку, чтобы легче было работать с регулярными выражениями. В первую очередь, я решил избавиться от символов переноса строки “\n”

```
corpus = re.sub(r'\n', ' ', corpus)
```

Далее я решил удалить все скобки их содержимое, так как именно в них содержится главный мусор такой, как перевод на греческий, латынь и т.д.

```
corpus = re.sub(r'\([^()]*\)', '', corpus)
```

После удаления скобок у меня появились пробелы перед знаками препинания. Их я и решил удалить следующими

```
corpus = re.sub(r'\s+(?=(?:[,.?!:;…]))', '', corpus)
```

Поменял буквы Ё и ё на Е и е соответственно

```
corpus = re.sub(r'Ё', 'E', corpus)
corpus = re.sub(r'ё', 'e', corpus)
```

Текст в википедии всегда сопровождается словами с ударениями, поэтому необходимо удалить диакритические знаки

```
corpus = re.sub(r'aˊ', 'a', corpus)
corpus = re.sub(r'Aˊ', 'A', corpus)
```

Заменять диакритические знаки я решил «в лоб», просто меняя каждую гласную с ударением на обычную гласную. Здесь приведен код только для замены букв «а» и «А»

Было: «Абисса́ль, абисса́льная зона»
Стало: «Абиссаль, абиссальная зона»

И также я заметил, что появились в некоторых местах двойные пробелы, их я тоже исправил с помощью методов из **2.1.4.**

```
corpus = re.sub(r'\s{1,}', ' ', corpus)
```

Корпус выглядел следующим образом:

```
'Абиссаль, абиссальная зона (греч. ἄβυσσος — «бездонный») — зона наибольши
х морских глубин (глубже 3000 м), населённая сообществами бентоса океаниче
ского дна. Рельеф зоны представлен глубоководными котловинами, подводными
хребтами и плато. Абиссаль характеризуется отсутствием дневного света (пос
тоянно находится в вечной темноте), слабой подвижностью вод. Живые организ
мы, такие как живоглоты (лат. Chiasmodon niger), батиптеры (Bathypterois g
rallator), рыбы отряда удильщикообразных (Lophiiformes), населяющие абисса
льные зоны, способны выдерживать значительное глубинное океаническое давле
ние, характеризующее эти зоны (до 775 кг на см²). Животные в основном слеп
ы, отличаются древним происхождением видов. На дне обитают различные виды
иглокожих, губок, анемон, червей и ракообразных. У некоторых видов животны
х образовываются люминесцентные органы. В 1978 г. учёные обнаружили оазисы
жизни абиссали, там около выходов термальных вод и газов появляются уникал
ьные группы организмов. Основой их жизни служит тепловая и химическая энер
гия термальных вод. Воды отличаются слабой подвижностью и очень низкой тем
пературой (+2 °C) с низким содержанием биогенных веществ. Около 90 % дна М
ирового океана покрыто абиссальными отложениями.\nНиже абиссальной зоны на
ходится ультраабиссальная зона.',
 'Автобус (сокращение от автомобиль-омнибус) — безрельсовое механическое м
оторное транспортное средство, технически предназначенное для перевозки де
```

вяти и более пассажиров и способное маневрировать на дороге, приводимое в движение источником энергии, запасённым или производимым из топлива, храняющегося на борту (бывают аккумуляторные, бензиновые, газотопливные, дизельные, суперкондесаторные, а также автобусы на прочих топливных элементах).\nАвтобусы длиной менее 5,5 м называются микроавтобусами (по российской классификации — автобусы особо малого класса), к микроавтобусам также относят минивэны вместимостью от 9 до 16 пассажиров. К ним относят микроавтобусы Fiat Ducato, Ford Transit, IVECO Daily, ГАЗель, РАФ-977, РАФ-2203, а также прочие микроавтобусы различных марок.\nАвтобусы длиной от 5,5 до 7,0 м называются также микроавтобусами малого класса.\nАвтобусы длиной от 7,0 до 10,0 м называются автобусами среднего класса. К ним также относят гибрид микроавтобуса и автобуса среднего класса IVECO VSN700 и прочих автобусов с подобной комплектацией.\nАвтобусы длиной от 10,0 до 16,0 м (до 2016 года до 12 метров) называются автобусами большого класса.\nАвтобусы длиной свыше 16,0 м называются автобусами особо большого класса. К ним относят трехосники (Волжанин-6270, Волжанин СитиРитм 15, МАЗ-107), и сочлененные автобусы.'

Сейчас он выглядит так:

'Абиссаль, абиссальная зона — зона наибольших морских глубин, населенная сообществами бентоса океанического дна. Рельеф зоны представлен глубоководными котловинами, подводными хребтами и плато. Абиссаль характеризуется отсутствием дневного света, слабой подвижностью вод. Живые организмы, такие как живоглоты, батиптеры, рыбы отряда удильщикообразных, населяющие абиссальные зоны, способны выдерживать значительное глубинное океаническое давление, характеризующее эти зоны. Животные в основном слепы, отличаются древним происхождением видов. На дне обитают различные виды иглокожих, губок, анемон, червей и ракообразных. У некоторых видов животных образовываются люминесцентные органы. В 1978 г. ученые обнаружили оазисы жизни абиссали, там около выходов термальных вод и газов появляются уникальные группы организмов. Основой их жизни служит тепловая и химическая энергия термальных вод. Воды отличаются слабой подвижностью и очень низкой температурой с низким содержанием биогенных веществ. Около 90 % дна Мирового океана покрыто абиссальными отложениями. Ниже абиссальной зоны находится ультраабиссальная зона. Автобус — безрельсовое механическое моторное транспортное средство, технически предназначенное для перевозки девяти и более пассажиров и способное маневрировать на дороге, приводимое в движение источником энергии, запасенным или производимым из топлива, хранящегося на борту. Автобусы длиной менее 5,5 м называются микроавтобусами, к микроавтобусам также относят минивэны вместимостью от'

(Взята лишь часть корпуса)

Я более чем уверен, что в корпусе еще места, которые надо будет исправить, так как сам корпус содержит более 30 тысяч слов, и найти все проблемные места сразу было бы просто невозможно, но самую основную работу по очистке я провёл, и если в будущем я столкнусь с «кривыми местами» в корпусе, мне не составит труда их быстро исправить.