Домашнее задание №12

Асланов А.Б., ИУ9-21М

Реализовать простой метод автоматического аннотирования.
Для статьи:

1. Посчитать частотный список лемм;
2. Исключить служебные слова;
3. Отобрать четыре предложения, в которых сумма частот слов максимальна.

И другой вариант делим сумму частот на длину предложения (т.е. усредняем) – выбираем предложение с максимальным значением.

Применяем для трех статей не менее страницы (можно взять статью Коммерсанта, РБК).
Отчет: взятые статьи, полученные аннотации.
Ваша оценка: хорошо ли получилось.

In [1]:
```
corpus_name = 'rbk_article.txt'
with open(corpus_name) as fin:
    article = fin.read()
```

In [2]:
```
import re
from pymystem3 import Mystem
from operator import itemgetter
mystem = Mystem()

# Убираем знаки препинания (все, кроме точки, запятой, точки с запятой, восклицательного и вопросительного знаков:
# они пригодятся при разделении текста на предложения после лемматизации)
article = re.split("[ \n«»%$,()—'`–:']", article)

# Убираем пустые строки
article = list(filter(None, article)) #None - потому что фильтрует пустые строк
article = ' '.join(article)

# Заменяем знаки конца предложения на точку (чтобы делить потом на предложения было удобнее)
article = article.replace(';', '.')
article = article.replace('?', '.')
article = article.replace('!', '.')
article = article.replace('\n', '.')

# Обработка для получения лемм
lemmas = ''.join(mystem.lemmatize(article))
lemmas = lemmas[:-2]

# Частотные списки лемм
freq_lemmas = {}
for word in lemmas.strip().lower().split():
    freq_lemmas[word] = freq_lemmas.get(word, 0) + 1
del freq_lemmas['.']

# Первые 20 наиболее частотных лемм с учетом стоп-слов
sorted(freq_lemmas.items(), key=itemgetter(1), reverse=True)[:20]
```

Out[2]:
```
[('в', 30),
 ('на', 21),
 ('и', 19),
 ('год', 14),
 ('по', 13),
 ('экономика', 10),
 ('рост', 10),
 ('быть', 10),
 ('уровень', 10),
 ('цель', 8),
 ('до', 8),
 ('с', 8),
 ('для', 8),
 ('к', 8),
 ('план', 7),
 ('национальный', 7),
 ('россия', 7),
 ('2024', 7),
 ('год.', 7),
 ('мера', 6)]
```

In [3]:
```
from nltk.corpus import stopwords
stopwords_nltk = stopwords.words('russian')


lemmas = lemmas.split()
lemmas = list(filter(lambda word: word not in stopwords_nltk, lemmas))
```

In [4]:
```
lemmas = ' '.join(lemmas)
lemsents = lemmas.split('.')
```

In [5]:
```
def choose_sents_with_max_freqs(lemsents, freq_lemmas):
    """СПОСОБ 1: выбираем 4 предложения, в которых сумма частот слов максимальна"""

    # Сами предложения
    lemsentwords = [sent.split() for sent in lemsents]
    # Список списков с частотами (каждый список в списке - это одно предложение)
    freqs_list = [[freq_lemmas.get(word, 1) for word in sent] for sent in lemsentwords]
    # Сумма частот слов на каждом предложении
    freqs_sums = [sum(freq) for freq in freqs_list]

    # ранжируем предложения по частотности
    lemsentwords_joined = [' '.join(sent) for sent in lemsentwords]
    d = {}
    for sent, freq_sum in zip(lemsentwords_joined, freqs_sums):
        d[sent] = freq_sum
    for sent in sorted(d, key=d.get, reverse=True)[:4]:
        print(sent, '|||СУММА ЧАСТОТ =',d[sent])
        print('-'*50)


choose_sents_with_max_freqs(lemsents, freq_lemmas)
```

```
согласно приложение единый план реальный располагать денежный доход население 2019 год вырастать незначительно 0 5 однако 2021 год темп рост превышать 2 2024 год достигать 2 4 |||СУММА ЧАСТОТ = 124
--------------------------------------------------
входить топ-5 экономика мир правительство отмечать темп рост российский экономика последний год стабилизироваться уровень 1 5 2 5 входить пятерка крупный экономика мир понадобиться выходить показатель высоко 3 |||СУММА ЧАСТОТ = 119
--------------------------------------------------
перелом тренд падение реальный доход россиянин обеспечение рост пенсия темп выше инфляция правительство определять следующий мера федеральный мрот ежегодно устанавливаться уровень прожиточный минимум трудоспособный население второй квартал предыдущий год |||СУММА ЧАСТОТ = 102
--------------------------------------------------
единый план представлять набор основной мера действие ключевой инструмент достижение национальный цель развитие определенный президент майский указ 2024 год |||СУММА ЧАСТОТ = 78
--------------------------------------------------
```

In [6]:
```
import numpy as np

def choose_sents_with_max_average_freq(lemsents, freq_lemmas):
    """СПОСОБ 2: делим сумму частот на длину предложения"""

    # Сами предложения
    lemsentwords = [sent.split() for sent in lemsents]
    # Список списков с частотами (каждый список в списке - это одно предложение)
    freqs_list = [[freq_lemmas.get(word, 1) for word in sent] for sent in lemsentwords]
    # Сумма частот слов на каждом предложении
    freqs_sums = np.array([sum(freq) for freq in freqs_list])
    # Длины предложений
    sents_lens = np.array([len(sent) for sent in lemsentwords])
    average_freqs = freqs_sums / sents_lens
    average_freqs = average_freqs.tolist()

    # ранжируем предложения по частотности
    lemsentwords_joined = [' '.join(sent) for sent in lemsentwords]
    d = {}
    for sent, avg_freq in zip(lemsentwords_joined, average_freqs):
        d[sent] = avg_freq
    for sent in sorted(d, key=d.get, reverse=True)[:4]:
        print(sent, '|||СРЕДНЕЕ ПО ЧАСТОТАМ =', d[sent])
        print('-'*50)


choose_sents_with_max_average_freq(lemsents, freq_lemmas)
```

```
премьер-министр подписывать единый план достижение национальный цель развитие россия 2024 год |||СРЕДНЕЕ ПО ЧАСТОТАМ = 6.0
--------------------------------------------------
рост производительность труд |||СРЕДНЕЕ ПО ЧАСТОТАМ = 5.333333333333333
--------------------------------------------------
повышение ожидать продолжительность жизнь 78 год 2030 год 80 год |||СРЕДНЕЕ ПО ЧАСТОТАМ = 5.2
--------------------------------------------------
достижение устойчивый рост реальный доход также рост пенсия выше инфляция |||СРЕДНЕЕ ПО ЧАСТОТАМ = 5.2
--------------------------------------------------
```

Выводы

В целом лучше себя показывает аннотирование по максимальным средним частотам, так как содержания аннотаций охватывают статью в целом, в то время как аннотирование по максимальной сумме частот лучше достаёт факты из статьи, но они фрагементарны и содержат много лишнего: аннотация получается слишком длинной.
Также оба метода имеют некоторые ошибочные предложения в аннотациях, но их количество незначительно.

Аннотирование предложений по максимальной сумме частот

Статья 1
https://www.rbc.ru/economics/08/05/2019/5cd2f77c9a794768881ddad1?from=center

```
согласно приложение единый план реальный располагать денежный доход населе
ние 2019 год вырастать незначительно 0 5 однако 2021 год темп рост превыша
ть 2 2024 год достигать 2 4 |||СУММА ЧАСТОТ = 124
--------------------------------------------------
входить топ-5 экономика мир правительство отмечать темп рост российский эк
ономика последний год стабилизироваться уровень 1 5 2 5 входить пятерка кр
упный экономика мир понадобиться выходить показатель высоко 3 |||СУММА ЧАС
ТОТ = 119
--------------------------------------------------
перелом тренд падение реальный доход россиянин обеспечение рост пенсия тем
п выше инфляция правительство определять следующий мера федеральный мрот е
жегодно устанавливаться уровень прожиточный минимум трудоспособный населен
ие второй квартал предыдущий год |||СУММА ЧАСТОТ = 102
--------------------------------------------------
единый план представлять набор основной мера действие ключевой инструмент
достижение национальный цель развитие определенный президент майский указ
2024 год |||СУММА ЧАСТОТ = 78
--------------------------------------------------
```

Статья 2
https://www.rbc.ru/newspaper/2019/04/30/5cc6d6cb9a79478856ec409a

```
сша турция китай наращивать оборонный расход первый место рейтинг оборонны
й трата SIPRI традиционно занимать соединять штат 649 млрд 3 2 ввп второй
оказываться китай 250 млрд 1 9 ввп третий позиция саудовский аравия 67 6 м
лрд 8 8 ввп четвертый индия 66 5 млрд 2 4 ввп замыкать пятерка франция 63
8 млрд 2 3 ввп |||СУММА ЧАСТОТ = 244
--------------------------------------------------
SIPRI стараться включать свой оценка весь расход действовать вооруженный с
ила военный деятельность число расход военизированный структура росгвардия
гражданский персонал оборонный ведомство социальный пособие военный семья
оборонный исследование разработка военный строительство военный помощь стр
ана |||СУММА ЧАСТОТ = 190
--------------------------------------------------
данные SIPRI объем военный расход история независимый россия достигать мак
симум 2016 год 69 2 млрд 5 3 ввп показатель начинать снижаться |||СУММА ЧА
СТОТ = 119
--------------------------------------------------
эксперт также указывать великобритания который рейтинг военный расход SIPR
I идти вслед россия показатель 50 млрд это британский сила общий назначени
е мало российский большинство составлять примерно десять добавлять кашин |
||СУММА ЧАСТОТ = 117
--------------------------------------------------
```

Статья 3
https://www.rbc.ru/newspaper/2019/04/30/5cc48a549a79475b870c850e

```
импортозамещение справляться Oracle американский производитель программный
обеспечение поставлять рекорд продажа россия несмотря программа импортозам
ещение объем закупка программный обеспечение американский Oracle госструкт
ура госкомпания 2018 год достигать 13 3 млрд руб |||СУММА ЧАСТОТ = 99
--------------------------------------------------
```

это максимум последний пять год итог 2018 год госорган госкомпания потрати
ть закупка лицензия услуга техподдержка продукт американский корпорация Or
acle 13 3 млрд руб |||СУММА ЧАСТОТ = 82

счетный палата выявлять рост нарушение госзакупки 5 2 год экономика правил
о крупный заказчик являться клиент Oracle закупать новый лицензия использо
вание продукт компания редко необходимо увеличивать количество функция име
ться клиент система связь рост количество пользователь |||СУММА ЧАСТОТ = 7
8

отношение госкомпания конец 2018 год действовать директива первый зампред
правительство глава минфин антон силуанов который должный доводить 50 доля
отечественный софт 2021 год |||СУММА ЧАСТОТ = 68

Аннотирование предложений по максимальным средним частотам

Статья 1
https://www.rbc.ru/economics/08/05/2019/5cd2f77c9a794768881ddad1?from=center

премьер-министр подписывать единый план достижение национальный цель разви
тие россия 2024 год |||СРЕДНЕЕ ПО ЧАСТОТАМ = 6.0

рост производительность труд |||СРЕДНЕЕ ПО ЧАСТОТАМ = 5.333333333333333

повышение ожидать продолжительность жизнь 78 год 2030 год 80 год |||СРЕДНЕ
Е ПО ЧАСТОТАМ = 5.2

достижение устойчивый рост реальный доход также рост пенсия выше инфляция
|||СРЕДНЕЕ ПО ЧАСТОТАМ = 5.2

Статья 2
https://www.rbc.ru/newspaper/2019/04/30/5cc6d6cb9a79478856ec409a

---
целое 2018 год мировой военный расход составлять 1822 млрд |||СРЕДНЕЕ ПО Ч
АСТОТАМ = 10.0

разный страна военный расход считать по-разному |||СРЕДНЕЕ ПО ЧАСТОТАМ = 8
.833333333333334

согласно оценка 2018 год военный расход россия составлять 61 4 млрд 3 9 вв
п |||СРЕДНЕЕ ПО ЧАСТОТАМ = 8.285714285714286

рост военный расход китай продолжаться 24-й год подряд 1994 год увеличиват
ься десять раз |||СРЕДНЕЕ ПО ЧАСТОТАМ = 7.230769230769231

Статья 3
https://www.rbc.ru/newspaper/2019/04/30/5cc48a549a79475b870c850e

19 2017 год рекорд последний пять год |||СРЕДНЕЕ ПО ЧАСТОТАМ = 4.285714285
714286

данные TAdviser количество закупка продукция Oracle 2018 год сокращаться 2
50 штука 380 2017 год |||СРЕДНЕЕ ПО ЧАСТОТАМ = 3.5714285714285716

это максимум последний пять год итог 2018 год госорган госкомпания потратить закупка лицензия услуга техподдержка продукт американский корпорация Oracle 13 3 млрд руб |||СРЕДНЕЕ ПО ЧАСТОТАМ = 3.5652173913043477