Московский авиационный институт
Факультет прикладной математики и физики

Лабораторная работа №8

по курсу:
«Информационный поиск»
по теме:
«TF-IDF ранжирование»
2 семестр

Студент: Ахмед С. Х.
Преподаватель: Калинин А. Л.
Группа: 8О-106М

Москва, 2019 г

Постановка задачи

Необходимо сделать ранжированный поиск на основании схемы ранжирования TF-IDF. Теперь, если запрос содержит в себе только термины через пробелы, то его надо трактовать как нечёткий запрос, т.е. допускать неполное соответствие документа терминам запроса и т.п. Примеры запросов:

- [роза цветок]
- [московский авиационный институт]

Если запрос содержит в себе операторы булева поиска, то запрос надо трактовать как булев, т.е. соответствие должно быть строгим, но порядок выдачи должен быть определён ранжированием TF-IDF. Например:

- [роза && цветок]
- [московский && авиационный && институт]

В отчёте нужно привести несколько примеров выполнения запросов, как удачных, так и не удачных.

Ход работы

Оперируемые формулы

TF(term) = 1 + log10(frequency)

IDF вычисляется классическим способом

IDF(document) = log10(NUM_DOCUMENTS/ num_documents_with_term)

TF-IDF (term,document) = TF(term)*IDF(document)

TF-IDF оценка для запроса = сумма TF-IDF слов в запросе

В качестве нечеткого запроса я решил воспользоваться заменой пробелов операцией OR (раньше заменялось операцией AND)

Код ранжирования:

Считывания словаря из диска(считывание файла)

```
def ranger(self,requests,answer):
    dictionary = self.load_dictionary_tf()[0]
```

```
tf_idf_dicts = []
for term in requests:
    tf_idf_dicts.append(self.load_postings_tf(term,dictionary[term][0],dictionary[term][1]))
tf = []
for idx in range(len(answer)):
    res = 0
    for jdx in range(len(tf_idf_dicts)):
        term_ =list(tf_idf_dicts[jdx].keys())[0]
        if answer[idx] not in tf_idf_dicts[jdx][term_]:
            print('Here')
            continue
        else:
            res += tf_idf_dicts[jdx][term_][answer[idx]]
    print(res)
    tf.append((answer[idx],res))
return sorted(tf,key = lambda tup: tup[1],reverse = True)
```

Хранение TF-IDF оценки

Для подсчета TF-IDF запроса необходимо подправить структуру обратного индекса, чтобы без проблем считывать и ранжировать данные. В качестве решения на этапе индексации я стал считать для каждого термина TF-IDF значение для каждого документа, сведя задачу к простому считыванию нужных TF-IDF оценок документов, с последующей сортировкой по значениям: при этом в структуре обратного индекса произойдут следующие изменения при сжатии индекса

M байт – количество документов (Y) в	N байт – doc_id в сжатом виде	8 байт – TF-IDF для блока координат	P байт – длина блока координат	Z байт – координаты coord_n	K байт – смещение для перехода вперед по списку координат (прыжок на

сжатом виде		в несжатом виде	в байтах с прыжками в сжатом виде	в сжатом виде	координату coord_n+jump) Блок повторяется 1 раз в jump координат
	Повторение блока Y раз				

При несжатом формате:

4 байт – количество документо в D	N байт – doc_id	8 байт – TF-IDF для блока координат в несжатом виде	4 байт – длина блока координат в байтах с прыжками	Z байт – координат ы coord_n	K байт – смещение для перехода вперед по списку координат (прыжок на координату coord_n+jump) Блок повторяется 1 раз в jump координат
	Повторение блока D раз				

Примеры Запросов

Запрос	**P@1**	**P@10**	**NDCG**
сибирская платформа (сибирская and платформа P@1 = 1 и P@10 = 1)	**1**	0.5(учитывалась значимость слова и наличие слов запроса)	DCG@10 = 17.56 IDCG@10 = 18.28 NDCG@10 = 0.96
лесной кот (лесной AND кот имеет P@1 = 1 и P@10 = 1)*	**0**	0.4(учитывалась значимость слова и наличие слов запроса)	DCG@10 = 18.89 IDCG@10 = 21.15 NDCG@10 = 0.89

советского государства (советского AND государства)(топ 10 элементов совпало)	1	1	DCG@10 = 25.05 IDCG@10= 27.05 NDCG@10 = 0.91

- **Д**ля булевых запросов метрика качества не считалась

Как можно заметить, метрика Precision очень хромает для нечетких запросов, однако даже в этом случае какое-то понятие является превалирующим в документе, что сказывается на информативности и полезности документа.

Примеры:

(сибирская платформа, лесной кот (в нем перевес в сторону кота)). То есть нечеткий запрос будет страдать в смысле метрики Precision и в смысле этой метрики они являются плохими примерами.

Также плохим примером может быть такой запрос, чей термин входит в маленькую статью. В таком случае без учета дополнительной информации данному документу присваивается больший ранг, нежели более релевантным, но более длинным документам.

Качество поиска

Качество поиска оценивалось запросами из лабораторной работы по Булевому поиску. Сравнивались поисковики Google, Wiki и мой поисковик. Вычислялись метрики P, DCG, NDCG. Оценки SERP выдачи

служить всю жизнь государству	0	4	4	2	2	2	2	5	2	2
соглашение о свободной торговле	1	1	4	1	1	4	4	5	1	1
зимний дворец	5	3	3	4	4	2	1	2	3	4
российская империя	1	2	5	2	2	1	1	3	2	2
англо русской войне	5	3	2	2	2	2	2	2	2	2
всероссийский съезд советов	4	4	4	4	3	2	3	4	2	2
транссибирская магистраль	5	3	2	2	3	2	2	2	2	2
свидетели иеговы	5	4	3	3	2	2	3	3	2	2
владимир ильич ульянов	5	1	1	1	1	1	1	1	1	1
крымский мост	5	2	2	2	2	2	2	2	2	3
ионообменные смолы	5	4	2	2	0	0	0	0	0	0
красный террор	5	5	2	2	2	2	3	2	2	2
аутсайдер моя жизнь как интрига	5	1	1	1	1	1	1	1	1	1
день защитника отечества	5	3	2	2	2	2	3	2	2	2
день пограничника	4	2	3	3	4	2	1	1	1	1
ударный музыкальный инструмент	5	5	1	5	5	5	5	2	5	5
генеральный штаб вооружённых сил российской федерации	5	5	2	5	5	2	2	3	2	2
чирлидинг	5	3	3	4	4	2	1	2	3	4
голштинских заявлений	5	0	0	0	0	0	0	0	0	0
быстрая сортировка	5	4	4	2	2	2	2	2	1	1
умберто боччони	5	2	2	2	2	2	2	2	2	2
охотники за привидениями	5	4	4	3	3	5	2	4	4	4
божественный пёс	5	2	2	2	2	3	2	2	2	2
эдди мерфи	5	4	3	3	3	3	3	3	1	3
бегущий по льду	5	5	5	5	5	5	4	2	4	4
человек кусает собаку	5	2	2	2	2	3	2	2	2	2
военный коммунизм	5	2	4	2	4	3	3	3	2	2
петропавловская крепость	2	2	3	4	2	2	2	3	3	2
эрмитаж	3	5	2	4	2	2	2	3	3	2
ссср	5	2	2	2	3	2	3	2	3	2

:

Оценки Точности:

служить всю жизнь государству	False	True	True	False	False	False	False	True	False	False
соглашение о свободной торговле	False	False	True	False	False	True	True	True	False	False
зимний дворец	True	True	True	True	True	False	False	False	True	True
российская империя	False	False	True	False	False	False	False	True	False	False
англо русской войне	True	True	False	False	False	False	False	False	False	False
всероссийский съезд советов	True	True	True	True	True	False	True	True	False	False
транссибирская магистраль	True	True	False	False	True	False	False	False	False	False
свидетели иеговы	True	True	True	True	False	False	True	True	False	False
владимир ильич ульянов	True	False	False	False	False	False	False	False	False	False
крымский мост	True	False	False	False	False	False	False	False	False	True
ионообменные смолы	True	True	False	False	False	False	False	False	False	False
красный террор	True	True	False	False	False	False	True	False	False	False
аутсайдер моя жизнь как интрига	True	False	False	False	False	False	False	False	False	False
день защитника отечества	True	True	False	False	False	False	True	False	False	False
день пограничника	True	False	True	True	True	False	False	False	False	False
ударный музыкальный инструмент	True	True	False	True	True	True	True	False	True	True
генеральный штаб вооружённых сил российской федерации	True	True	False	True	True	False	False	True	False	False
чирлидинг	True	True	True	True	True	False	False	False	True	True
голштинских заявлений	True	False	False	False	False	False	False	False	False	False
быстрая сортировка	True	True	True	False	False	False	False	False	False	False
умберто боччони	True	False	False	False	False	False	False	False	False	False
охотники за привидениями	True	True	True	True	True	True	False	True	True	True
божественный пёс	True	False	False	False	False	True	False	False	False	False
эдди мерфи	True	True	True	True	True	True	True	True	False	True
бегущий по льду	True	True	True	True	True	True	True	False	True	True
человек кусает собаку	True	False	False	False	False	True	False	False	False	False
военный коммунизм	True	False	True	False	True	True	True	True	False	False
петропавловская крепость	False	False	True	True	False	False	False	True	True	False
эрмитаж	True	True	False	True	False	False	False	True	True	False
ссср	True	False	False	False	True	False	True	False	True	False

P1

```
[0,
 0,
 1,
 0,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 0,
 1,
 1]
```

P3

```
[0.6666666666666666,
 0.3333333333333333,
 1.0,
 0.3333333333333333,
 0.6666666666666666,
 1.0,
 0.6666666666666666,
 1.0,
 0.3333333333333333,
 0.3333333333333333,
 0.6666666666666666,
 0.6666666666666666,
 0.3333333333333333,
 0.6666666666666666,
 0.6666666666666666,
 0.6666666666666666,
 0.6666666666666666,
 1.0,
 0.3333333333333333,
 1.0,
 0.3333333333333333,
 1.0,
 0.3333333333333333,
 1.0,
 1.0,
 0.3333333333333333,
 0.6666666666666666,
 0.3333333333333333,
 0.6666666666666666,
 0.3333333333333333]
```

P10

```
[0.3,
 0.4,
 0.7,
 0.2,
 0.2,
 0.7,
 0.3,
 0.6,
 0.1,
 0.2,
 0.2,
 0.3,
 0.1,
 0.3,
 0.4,
 0.8,
 0.5,
 0.7,
 0.1,
 0.3,
 0.1,
 0.9,
 0.2,
 0.9,
 0.9,
 0.2,
 0.6,
 0.4,
 0.5,
 0.4]
```

Стоит заметить, что в среднем значения Precision заметно улучшилось по сравнению с Булевым поиском, что есть хорошо. Это обусловлено способом ранжирования поисковой выдачи (вес термина в документе).

CG оценки

Google

CG@3

Out[53]:

служить всю жизнь государству	7
соглашение о свободной торговле	14
зимний дворец	15
российская империя	10
англо русской войне	11
всероссийский съезд советов	15
транссибирская магистраль	10
свидетели иеговы	15
владимир ильич ульянов	13
крымский мост	12
ионообменные смолы	11
красный террор	10
аутсайдер моя жизнь как интрига	9
день защитника отечества	11
день пограничника	14
ударный музыкальный инструмент	12
генеральный штаб вооружённых сил российской федерации	12
чирлидинг	12
голштинских заявлений	10
быстрая сортировка	11
умберто боччони	9
охотники за привидениями	15
божественный пёс	11
эдди мерфи	12
бегущий по льду	15
человек кусает собаку	11
военный коммунизм	14
петропавловская крепость	15
эрмитаж	15
ссср	10

dtype: int64

CG@10

Out[58]:

служить всю жизнь государству	25
соглашение о свободной торговле	39
зимний дворец	40
российская империя	30
англо русской войне	22
всероссийский съезд советов	46
транссибирская магистраль	28
свидетели иеговы	41
владимир ильич ульянов	31
крымский мост	32
ионообменные смолы	33
красный террор	32
аутсайдер моя жизнь как интрига	23
день защитника отечества	35
день пограничника	35
ударный музыкальный инструмент	47
генеральный штаб вооружённых сил российской федерации	40
чирлидинг	40
голштинских заявлений	24
быстрая сортировка	28
умберто боччони	24
охотники за привидениями	44
божественный пёс	25
эдди мерфи	31
бегущий по льду	44
человек кусает собаку	25
военный коммунизм	33
петропавловская крепость	40
эрмитаж	40
ссср	30

dtype: int64

Wiki

CG@3

Out[62]:

служить всю жизнь государству	8
соглашение о свободной торговле	11
зимний дворец	13
российская империя	9
англо русской войне	10
всероссийский съезд советов	15
транссибирская магистраль	10
свидетели иеговы	15
владимир ильич ульянов	9
крымский мост	14
ионообменные смолы	13
красный террор	10
аутсайдер моя жизнь как интрига	9
день защитника отечества	13
день пограничника	11
ударный музыкальный инструмент	12
генеральный штаб вооружённых сил российской федерации	12
чирлидинг	12
голштинских заявлений	7
быстрая сортировка	13
умберто боччони	9
охотники за привидениями	15
божественный пёс	9
эдди мерфи	12
бегущий по льду	15
человек кусает собаку	9
военный коммунизм	14
петропавловская крепость	13
эрмитаж	13
ссср	9

dtype: int64

CG@10

Out[63]:

служить всю жизнь государству	25
соглашение о свободной торговле	36
зимний дворец	31
российская империя	26
англо русской войне	24
всероссийский съезд советов	46
транссибирская магистраль	28
свидетели иеговы	37
владимир ильич ульянов	28
крымский мост	27
ионообменные смолы	28
красный террор	30
аутсайдер моя жизнь как интрига	23
день защитника отечества	30
день пограничника	14
ударный музыкальный инструмент	47
генеральный штаб вооружённых сил российской федерации	40
чирлидинг	40
голштинских заявлений	19
быстрая сортировка	28
умберто боччони	23
охотники за привидениями	44
божественный пёс	24
эдди мерфи	31
бегущий по льду	44
человек кусает собаку	24
военный коммунизм	33
петропавловская крепость	31
эрмитаж	31
ссср	26

dtype: int64

Мой поиск

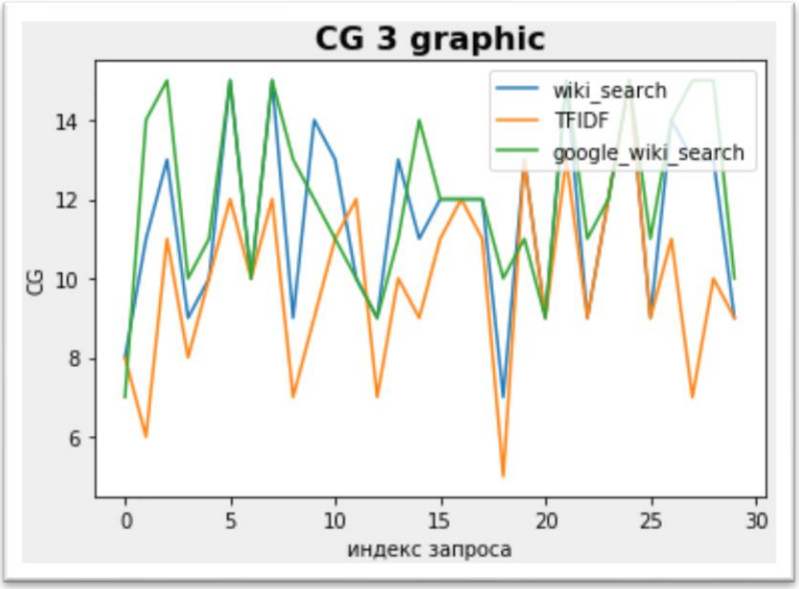
CG@3

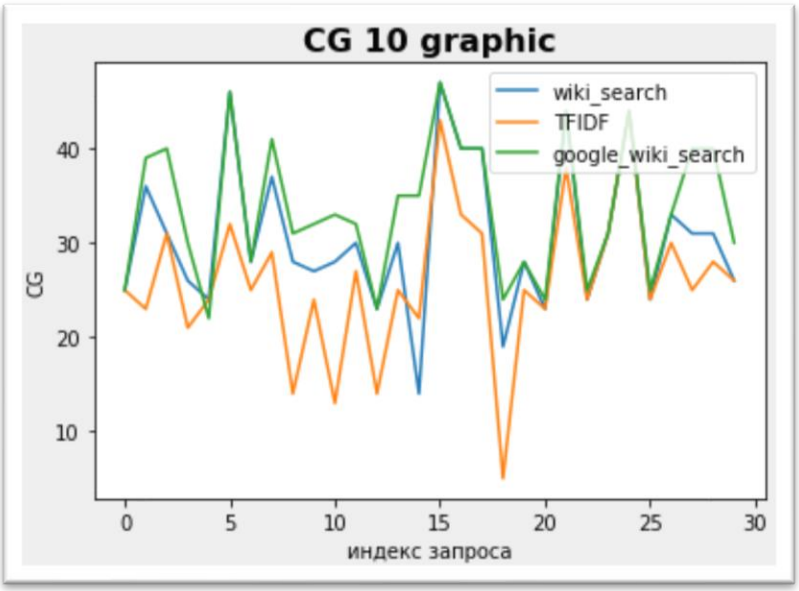
Out[66]:
```
служить всю жизнь государству                                 8
соглашение о свободной торговле                               6
зимний дворец                                                11
российская империя                                            8
англо русской войне                                          10
всероссийский съезд советов                                  12
транссибирская магистраль                                    10
свидетели иеговы                                             12
владимир ильич ульянов                                        7
крымский мост                                                 9
ионообменные смолы                                           11
красный террор                                               12
аутсайдер моя жизнь как интрига                               7
день защитника отечества                                     10
день пограничника                                             9
ударный музыкальный инструмент                               11
генеральный штаб вооружённых сил российской федерации       12
чирлидинг                                                    11
голштинских заявлений                                         5
быстрая сортировка                                           13
умберто боччони                                               9
охотники за привидениями                                     13
божественный пёс                                              9
эдди мерфи                                                   12
бегущий по льду                                              15
человек кусает собаку                                         9
военный коммунизм                                            11
петропавловская крепость                                      7
эрмитаж                                                      10
ссср                                                          9
dtype: int64
```

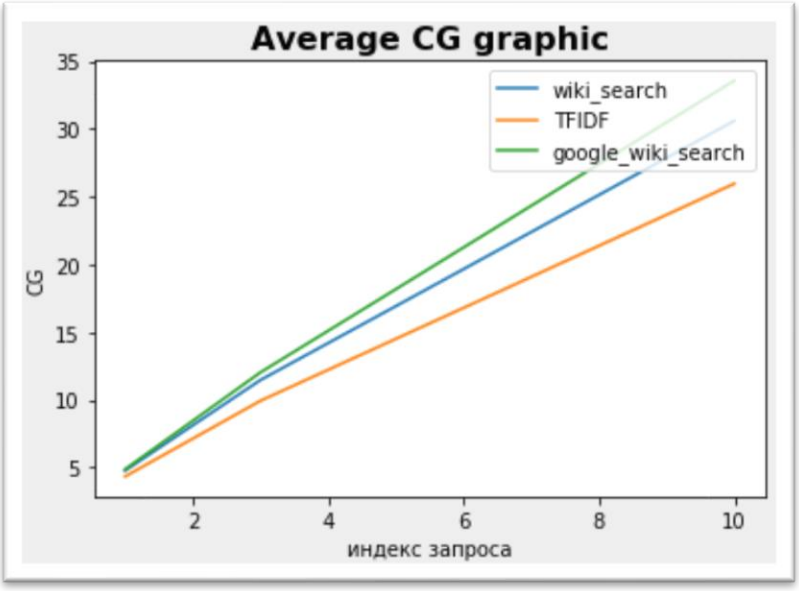
CG@10

Out[67]:
```
служить всю жизнь государству                                25
соглашение о свободной торговле                              23
зимний дворец                                                31
российская империя                                           21
англо русской войне                                          24
всероссийский съезд советов                                  32
транссибирская магистраль                                    25
свидетели иеговы                                             29
владимир ильич ульянов                                       14
крымский мост                                                24
ионообменные смолы                                           13
красный террор                                               27
аутсайдер моя жизнь как интрига                              14
день защитника отечества                                     25
день пограничника                                            22
ударный музыкальный инструмент                               43
генеральный штаб вооружённых сил российской федерации       33
чирлидинг                                                    31
голштинских заявлений                                         5
быстрая сортировка                                           25
умберто боччони                                              23
охотники за привидениями                                     38
божественный пёс                                             24
эдди мерфи                                                   31
бегущий по льду                                              44
человек кусает собаку                                        24
военный коммунизм                                            30
петропавловская крепость                                     25
эрмитаж                                                      28
ссср                                                         26
dtype: int64
```

Поведение оценок

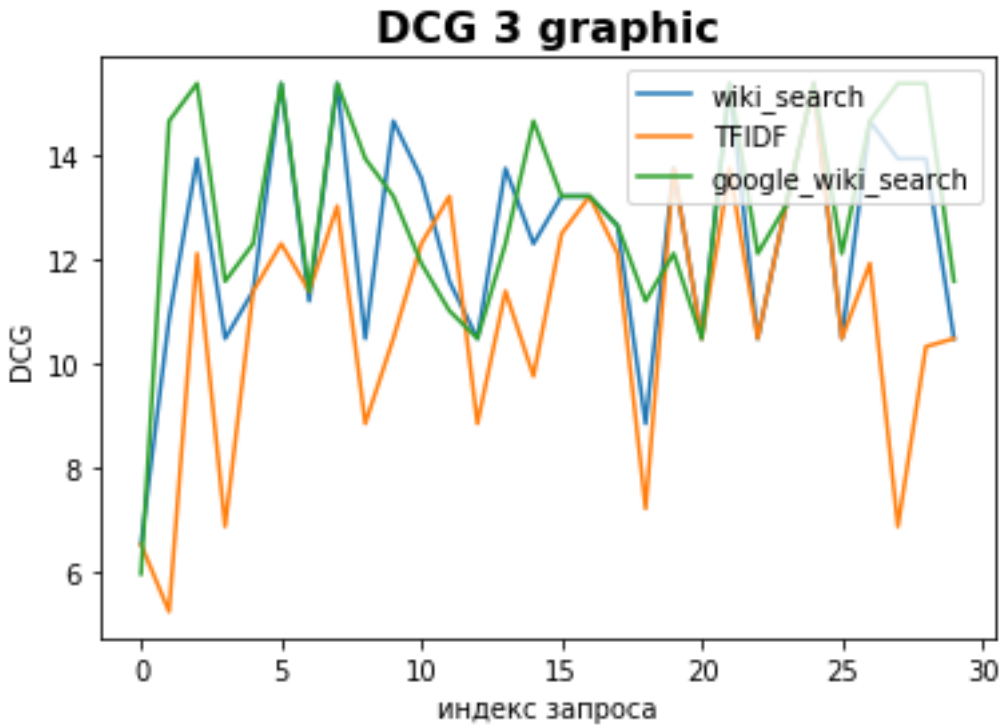


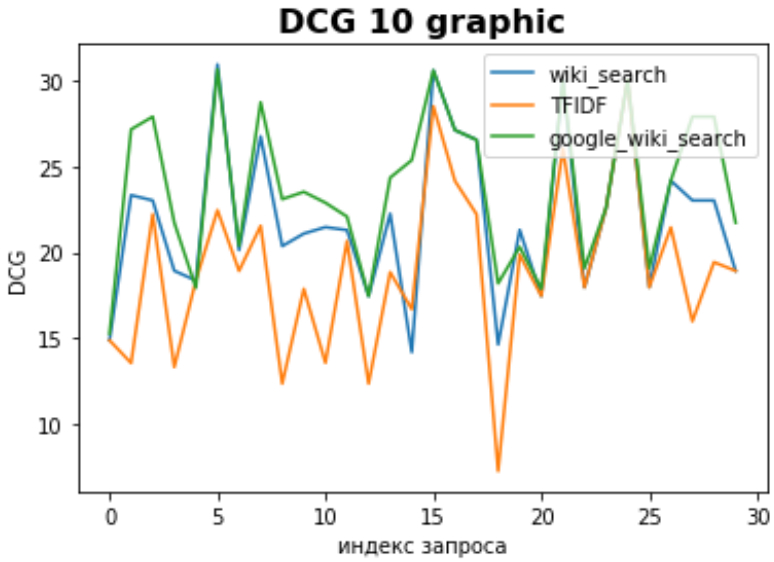


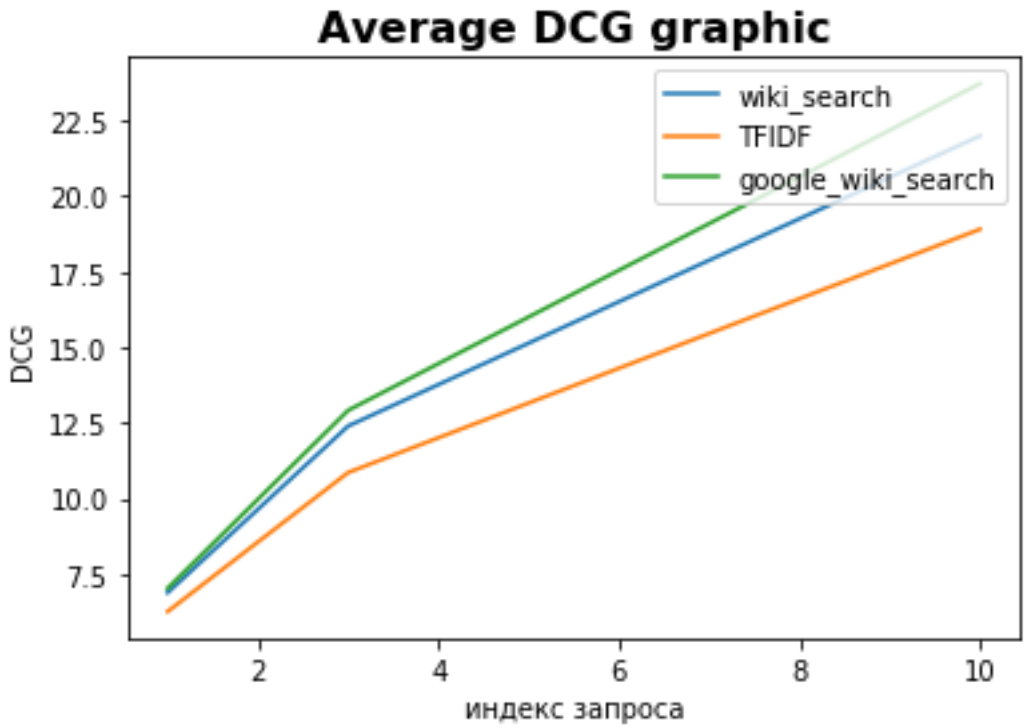


Ожидаемо Поиск от Google ведет себя намного лучше и стабильнее и показывает более существенный рост полезности выдачи. Стоит заметить, что и выдача отранжированная с помощью tf-idf в некоторых запросах, ведет себя хорош, достигая уровня google и Wiki, но в некоторых сильно проседает, однако несмотря на колеблющийся характер изменения оценок, они крутятся вокруг одного среднего числа. Также, данный поиск показывает стабильный рост полезности выдачи

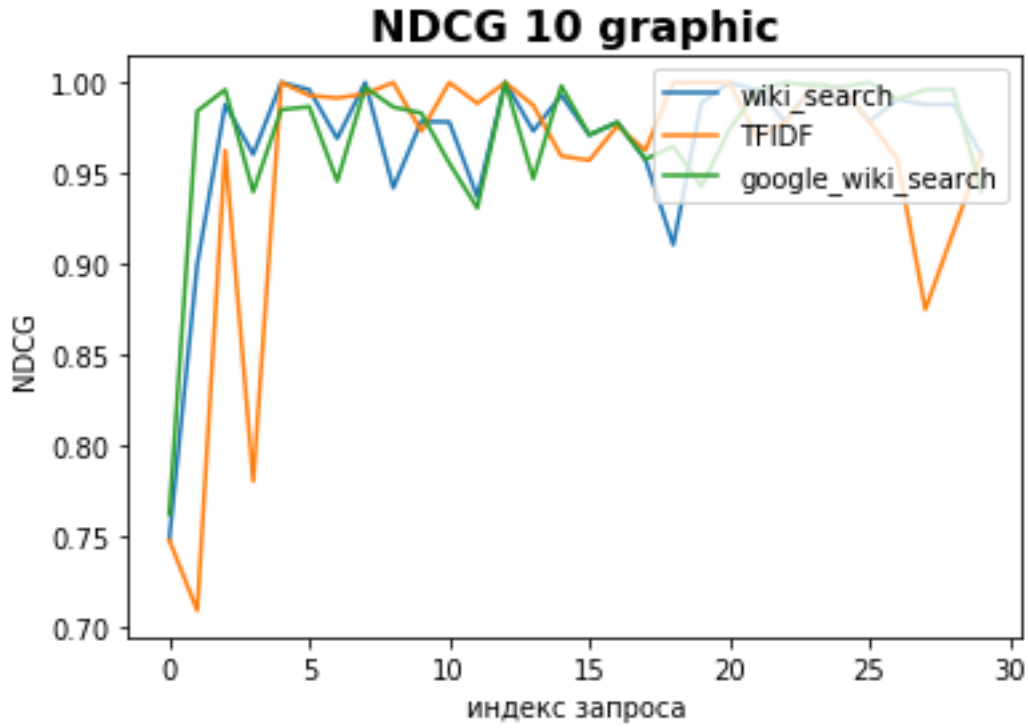
DCG(графики поведения)

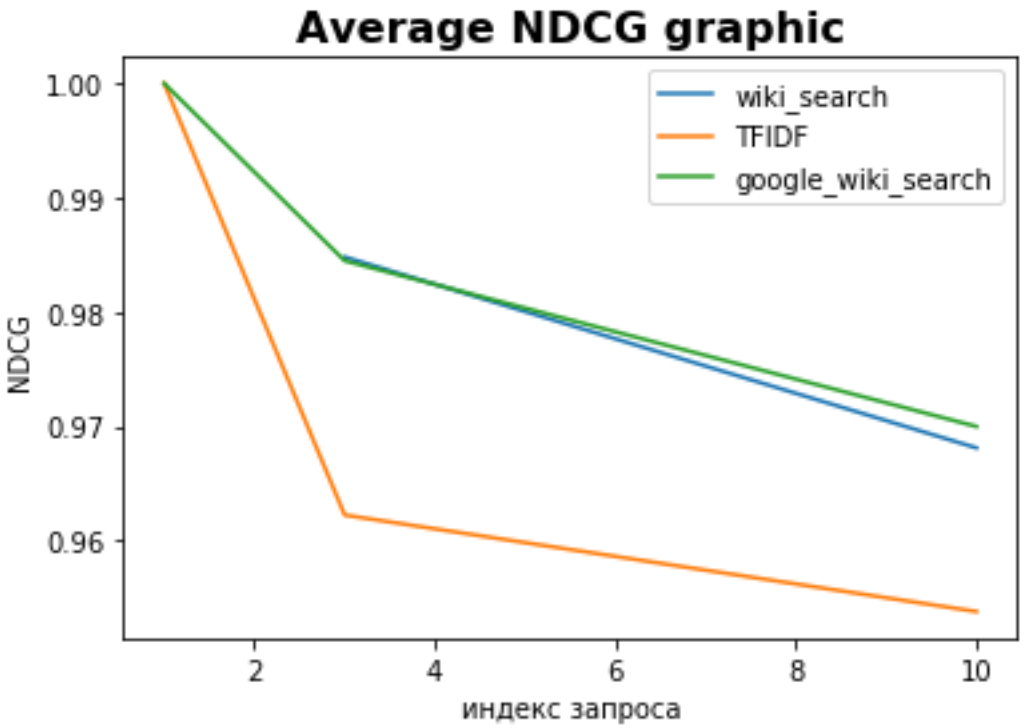






NDCG(графики поведения)





Заключение

По результатам работы можно сделать вывод, что ранжирование по TF-IDF без учета какой-либо еще информации может иметь негативные последствия: короткие статьи с одним вхождением слова N будут иметь более большой ранг, чем большая статья, в котором слово N входит также один раз, хотя именно большая статья может быть тем, что должно быть расположено в самом верху поисковой выдачи. Также стоит отметить, что tf-idf ранжирование не очень хорошо взаимодействует с нечетким, а также с булевским или поиском, однако хорош в строгих запросах на вхождение. Сравнение с другими поисковиками показало, что мой поисковик проигрывает основным поисковикам, в полезности и стабильности выдачи, также показывает общее падение нормированного коэффицента. Однако, он существенно лучше булева поиска, что показало сравнение и текущие результаты.