Московский авиационный институт
Факультет прикладной математики и физики

Лабораторная работа №4

по курсу:
«Информационный поиск»
по теме:
«Булев поиск»
2 семестр

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

Москва, 2019 г

Постановка задачи

Нужно реализовать ввод поисковых запросов и их выполнение над индексом, получение поисковой выдачи. Синтаксис поисковых запросов:

- Пробел или два амперсанда, «&&», соответствуют логической операции «И».
- Две вертикальных «палочки», «||» – логическая операция «ИЛИ»
- Восклицательный знак, «!» – логическая операция «НЕТ»
- Могут использоваться скобки. Парсер поисковых запросов должен быть устойчив к переменному числу пробелов, максимально толерантен к введённому поисковому запросу. Примеры запросов:
- [московский авиационный институт]
- [(красный || желтый) автомобиль]
- [руки !ноги]

Так же должна быть реализована утилита командной строки, загружающая индекс и выполняющая поиск по нему для каждого запроса на отдельной строчке входного файла.

В отчёте должно быть отмечено:

- Скорость выполнения поисковых запросов.
- Примеры сложных поисковых запросов, вызывающих длительную работу.
- Каким образом тестировалась корректность поисковой выдачи

Ход решения

Итак, я решил воспользоваться встроенным типом данных в Python: множества, для которых определены операции пересечения и объединения. Также я решил воспользоваться деревом выражений в узлах которых определены операции AND и OR соответсвующие && и || соответсвенно. Но предварительно решил совершить препроцессинг данных, обработать случай отрицания целого выражения, замена пробелов && там где нужно и обработка особых случаев

Сам алгоритм поиска тривиален. Первое мы нормализуем слова, дальше зависит от представления словаря и блоков. В случае дерева вычисляю хэш строки, затем вытаскиваю значения по ключу и смотрю на операцию, если вижу операцию || объединяю блоки операндов, если && пересекаю, если вижу отрицание выполняю операцию вычитания из списка всех блоков.

Одной из причин реализации дерева выражения является наличия приоритета вычисления (наличие скобок), а также скорость прохода по дереву.

В качестве одной из базовой структуры данных я решил взять Множества, встроенные в Python, которые обладают эффективными реализациями пересечения и объединения. Основным опасением вызванным появлением множества, является замедление скорости, однако на практике этого не происходит. Также, здесь стоит понимать, что в питоне, в отличии от компилируемых языков программирования, переход, осуществляемый со списка на множество дешевый (смена типа в указателе PyObject). В остальном же, ничего не меняется. Еще одно говорящее в пользу множеств, операции над ними быстрее, чем над списком, а также список и множество относятся к одному классу объектов.

Несколько элементов координатного блока(10 элементов(байты))

```
[[b'\xb5+'], [b'\xb1\x04'], [b'\x8eD'], [b'\xb5+'], [b'\xb5+'], [b'\xb5+'], [
b', '], [b'\xc8_'], [b'\x00\x87', b'\x02\x02', b'\x03Y', b'\nD', b'\x10\x8a',
b'\x11\x14', b'\x11\x1a', b'\x11k', b'\x11\x8a', b'\x12\x8a', b'\x15\x05', b'
\x17\x02', b'\x18\x1a', b'\x19,', b'\x19C', b'\x1bC', b'\x1db', b'\x1e7', b'
\x04', b'${', b'$\x88', b'+\x15', b'-I', b'.{', b'/\x02', b'5h', b'6l', b'7#'
, b'7$', b'7\x85', b'88', b'9A', b'<,', b'>\x12', b'>c', b'>}', b'@\x1f', b'D
u', b'E\x1f', b'F\x80', b'GH', b'H\x1b', b"O'", b'Q\x1f', b'Zt', b'a9', b'bz'
, b'c"', b'fy', b'f\x83', b'g\x86', b'l\x86', b'm\x00', b'm\x85', b'n\x1e', b
'n-', b'n\x85', b'pl', b'q\x86', b'r\x86', b's|', b'ut', b'u\x85', b'x\x15',
b'x;', b'x\x85', b'{\x1e', b'\x7fC', b'\x8a1', b'\x8c/', b'\x8c\x8b', b'\x8d\
x01', b'\x8e\x12', b'\x90\x12', b'\x90%', b'\x90q', b'\x93T', b'\x96\x8b', b'
\x97l', b'\x9b$', b'\x9bl', b'\xa50', b'\xa5n', b'\xad@', b'\xb3\x1f', b'\xb5
\x11', b'\xbd}', b'\xc1\x15', b'\xc4a', b'\xc5?', b'\xc5g', b'\xc5\x84', b'\x
c6O', b'\xc6z', b'\xc7\x06', b'\xc9g', b'\xcb~', b'\xcc\x86', b'\xce\x0f', b'
\xce.', b'\xce}', b'\xd0$', b'\xd2\x1b', b'\xd5y', b'\xd8\x1d', b'\xe0/', b'\
xe2.', b'\xe8\x84', b'\xedt', b'\xeeN', b'\xee\x84', b'\xf1\x82', b'\xf1\x86'
, b'\xf2\x86', b'\xf3\x86', b'\xf67', b'\xf7\x1a', b'\xf7d', b'\xf9y', b'\xfc
9', b'\xfdE', b'\xff3'], [b'\x0e\x01', b'\x15w', b'\x1bC', b'\x1db', b' \x12'
, b'+\x15', b'?\x8b', b'`\x03', b'u4', b'u9', b'\x8d\x01', b'\x8e\x00', b'\x9
4k', b'\xa1.', b'\xb6\x04', b'\xcb{', b'\xcc\x89', b'\xcf\x85', b'\xf8\x1f']]
```

В ходе работы выяснилось, что скорость выполнения запроса оказалось связанным с длинной поискового запроса (как мне показалось странным, по сути я предполагал, что это будет зависеть от скорости выполнения операции пересечения (для сложных поисковых запросов)) Однако, операции пересечения и объединения являются встроенными в питон, то они реализованы крайне эффективно

Примеры поисковых запросов находятся в гит репозитории, вместе с замерами по времени.

Проверка результатов осуществлялась с помощь утилит командной строки grep и т.п для поиска вхождения слов. Также заранее был произведен отбор статей, как статьи для теста. Сама суть проверки заключалась в пословной проверке вхождения слов из запроса в документ. В случае с булевым поиском МЕТРИКА точности равна 1. Можно попробовать оценить полноту: то есть посчитать, кол-во документов, которых не вошло в результат, однако он оценивает наш словарь и булев поиск имеет такую же оценку и там.

Итоговая точность на пуле запросов:

r.SearchRequest('(всю && (жизнь служить)) государству')

r.SearchRequest('правительством && (феврале || (года (вступил || силу)))')

r.SearchRequest('(заодно || или) && правительством && (феврале || (года (вступил || силу)))')

r.SearchRequest('!зимний дворец санкт петербурге гг дворец искусств прошлом главный императорский дворец россии расположенный по адресу дворцовая площадь дворцовая набережная нынешнее здание дворца пятое построено годах русским архитектором итальянского происхождения бартоломео франческо растрелли стиле пышного елизаветинского барокко элементами французского рококо интерьерах начиная советского времени стенах дворца размещена основная экспозиция эрмитажа момента окончания строительства году по год использовался качестве официальной зимней резиденции российских императоров году николай ii перенёс постоянную резиденцию александровский дворец царском селе октября года до ноября года во дворце работал госпиталь имени царевича алексея николаевича июля по ноябрь года во дворце размещалось временное правительство январе года во дворце открыт государственный музей революции разделявший здание государственным эрмитажем вплоть до года зимний дворец дворцовая площадь образуют красивейший архитектурный ансамбль современного города являются одним из главных объектов международного туризма история первый дворец свадебные палаты файл по рисунку махаева вид зимнего дворца jpg thumb right px второй дворец зимний дворец петра третий дворец дворец анны иоанновны всего за период годов городе возводилось пять зимних дворцов первоначально пётр поселился построенном на скорую руку году недалеко от петропавловской крепости одноэтажном доме первый зимний дворец свадебные палаты петра пётр великий владел участком между невой миллионной улицей на месте нынешнего эрмитажного театра году здесь глубине участка строится деревянный зимний дом небольшой двухэтажный дом высоким крыльцом черепичной крышей году были выстроены каменные свадебные палаты петра этот дворец стал подарком губернатора санкт петербурга александра даниловича меншикова свадьбе петра екатерины алексеевны второй зимний дворец дворец петра зимней канавки году архитектор георг маттарнови по приказу царя приступил постройке нового зимнего дворца на углу невы зимней канавки которую тогда называли зимнедомным каналом году пётр со всем своим')

%time r.SearchRequest('!зимний дворец санкт петербурге гг дворец искусств прошлом главный императорский дворец россии расположенный по адресу дворцовая площадь дворцовая набережная нынешнее здание дворца пятое построено годах русским архитектором итальянского происхождения бартоломео франческо растрелли стиле пышного елизаветинского барокко элементами французского рококо интерьерах начиная советского времени стенах дворца размещена основная экспозиция эрмитажа момента окончания строительства году по год использовался качестве официальной зимней резиденции российских императоров году николай ii перенёс постоянную резиденцию александровский дворец царском селе октября года до ноября года во дворце работал госпиталь имени царевича алексея николаевича июля по ноябрь года во дворце размещалось временное правительство январе года во дворце открыт государственный музей революции разделявший здание государственным эрмитажем вплоть до года зимний дворец дворцовая площадь образуют красивейший архитектурный ансамбль современного города являются одним из главных объектов международного туризма история первый дворец свадебные палаты файл по рисунку махаева вид зимнего дворца jpg thumb right px второй дворец зимний дворец петра третий дворец дворец анны иоанновны всего за период годов городе возводилось пять зимних дворцов первоначально пётр поселился построенном на скорую руку году недалеко от петропавловской крепости одноэтажном доме первый зимний дворец свадебные палаты петра пётр великий владел участком между невой миллионной улицей на месте нынешнего эрмитажного театра году здесь глубине участка строится деревянный зимний дом небольшой двухэтажный дом высоким крыльцом черепичной крышей году были выстроены

каменные свадебные палаты петра этот дворец стал подарком губернатора санкт петербурга александра даниловича меншикова свадьбе петра екатерины алексеевны второй зимний дворец дворец петра зимней канавки году архитектор георг маттарнови по приказу царя приступил ((постройке || нового) && зимнего) дворца на углу невы зимней канавки которую тогда называли зимнедомным каналом году пётр со всем своим')

r.SearchRequest('руины храма аполлона дельфах файл santuario delfos jpg thumb px модель античного святилища аполлона дельфах де льфы древнегреческий город юго восточной фокиде греция общегреческий религиозный центр храмом оракулом аполлона по легенде был назван по имени сына аполлона дельфа начала vi века до вплоть до конца iv века здесь проходили общегреческие пифийские игры археологический заповедник дельфы включён список всемирного наследия раскопками начавшимися году были открыты храм аполлона пифийского vi iv века до сокровищницы сифносцев около до афинян начало до стоя галерея портик афинян до театр ii век до стадион vi век до другие сооружения честь дельф была названа среда разработки delphi одноимённый язык также извилина дельфы на спутнике юпитера европе география руины древних дельф расположены километрах от побережья коринфского залива город итея на юго западном склоне горы парнаса на высоте метров над уровнем моря современный малый город дельфы находится неподалёку западнее руин община дельфы входит периферийную единицу фокиду община включает себя приморский малый город галаксидион этимология греческое слово восходит корню матка лоно утроба отсюда происходят слова брат или букв единоутробный дельфин новорожденный младенец утробный возможно из за внешнего сходства младенцем или из за того что крик дельфина похож на крик ребёнка причина такого названия видимо связана тем что представлении древних греков неподалёку от храма аполлона находился пуп земли мифология зевс послал краёв света двух орлов они встретились на пифийской скале либо там встретились лебеди либо вороны эта встреча обозначила что там находился пуп земли который охраняли две горгоны некогда доля дельф принадлежала гее она отдала её фемиде та подарила аполлону по трезенскому рассказу дельфы ранее принадлежали посейдону калаврия аполлону позднее они поменялись местностями югу от храма аполлона дельфах был храм геи статую аполлона дельфах виде колонны упоминает евмел городе было также прорицалище диониса по орфикам там гроб старшего диониса он лежит под треножником или омфале см загрей')

Оценка качества поиска

Оценка качества поиска проводилась на 30 запросах, представленных ниже. Сам поисковик сравнивался с поисковиком Google и встроенным поиском по Wikipedia. Для чистоты эксперимента, поиск Google был ограничен поиском по Wiki:

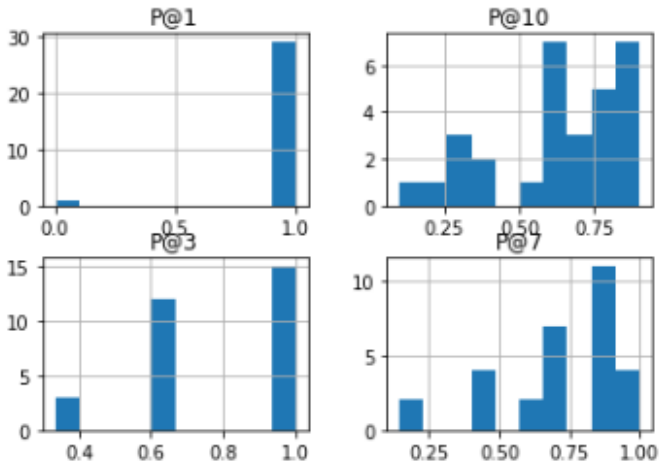
GOOGLE (вместе с распределением оценок):

---

[6]:

	P@1	P@3	P@7	P@10
служить всю жизнь государству	0	0.333333	0.571429	0.4
соглашение о свободной торговле	1	1.000000	0.857143	0.8
зимний дворец	1	1.000000	0.857143	0.8
российская империя	1	0.666667	0.714286	0.6
англо русской войне	1	0.666667	0.428571	0.3
всероссийский съезд советов	1	1.000000	0.857143	0.9
транссибирская магистраль	1	0.666667	0.571429	0.6
свидетели иеговы	1	1.000000	1.000000	0.8
владимир ильич ульянов	1	1.000000	0.714286	0.5
крымский мост	1	0.666667	0.857143	0.6
ионообменные смолы	1	0.666667	0.714286	0.7
красный террор	1	0.666667	0.714286	0.7
аутсайдер моя жизнь как интрига	1	0.333333	0.142857	0.1
день защитника отечества	1	0.666667	0.857143	0.6
день пограничника	1	1.000000	0.714286	0.6
ударный музыкальный инструмент	1	0.666667	0.857143	0.9
генеральный штаб вооружённых сил российской федерации	1	0.666667	0.857143	0.9
чирлидинг	1	0.666667	0.857143	0.9
голштинских заявлений	1	0.666667	0.428571	0.4
быстрая сортировка	1	1.000000	0.714286	0.6
умберто боччони	1	0.333333	0.142857	0.2
охотники за привидениями	1	1.000000	1.000000	0.9
божественный пёс	1	1.000000	0.428571	0.3
эдди мерфи	1	1.000000	1.000000	0.9
бегущий по льду	1	1.000000	1.000000	0.9
человек кусает собаку	1	1.000000	0.428571	0.3
военный коммунизм	1	1.000000	0.857143	0.7
петропавловская крепость	1	1.000000	0.857143	0.8
эрмитаж	1	1.000000	0.857143	0.8
ссср	1	0.666667	0.714286	0.6

Распределение оценок

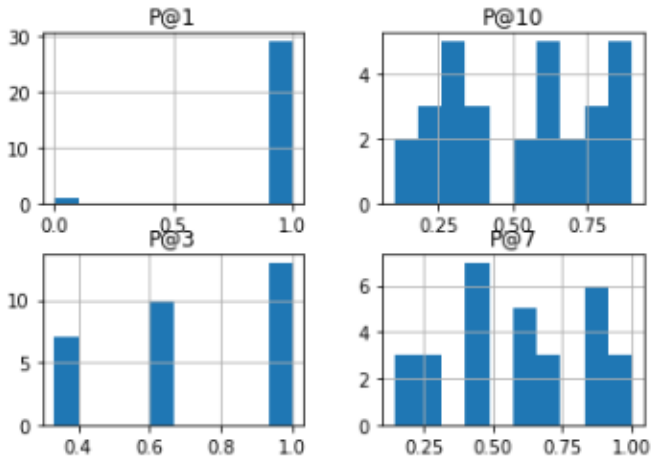


Wikipedia (вместе с распределением оценок):

[10]:

	P@1	P@3	P@7	P@10
служить всю жизнь государству	0	0.666667	0.285714	0.3
соглашение о свободной торговле	1	1.000000	0.857143	0.8
зимний дворец	1	1.000000	0.571429	0.6
российская империя	1	0.333333	0.428571	0.4
англо русской войне	1	0.666667	0.285714	0.2
всероссийский съезд советов	1	1.000000	1.000000	0.9
транссибирская магистраль	1	0.666667	0.571429	0.6
свидетели иеговы	1	1.000000	1.000000	0.8
владимир ильич ульянов	1	1.000000	0.714286	0.5
крымский мост	1	0.333333	0.428571	0.3
ионообменные смолы	1	0.666667	0.714286	0.7
красный террор	1	0.666667	0.428571	0.5
аутсайдер моя жизнь как интрига	1	0.333333	0.142857	0.1
день защитника отечества	1	1.000000	0.571429	0.4
день пограничника	1	0.666667	0.428571	0.3
ударный музыкальный инструмент	1	0.666667	0.857143	0.9
генеральный штаб вооружённых сил российской федерации	1	0.666667	0.857143	0.9
чирлидинг	1	0.666667	0.857143	0.9
голштинских заявлений	1	0.333333	0.142857	0.2
быстрая сортировка	1	1.000000	0.714286	0.6
умберто боччони	1	0.333333	0.142857	0.1
охотники за привидениями	1	1.000000	1.000000	0.9
божественный пёс	1	0.333333	0.285714	0.2
эдди мерфи	1	0.666667	0.857143	0.8
бегущий по льду	1	1.000000	0.428571	0.3
человек кусает собаку	1	1.000000	0.428571	0.3
военный коммунизм	1	1.000000	0.857143	0.7
петропавловская крепость	1	1.000000	0.571429	0.6
эрмитаж	1	1.000000	0.571429	0.6
ссср	1	0.333333	0.428571	0.4

Распределение оценок

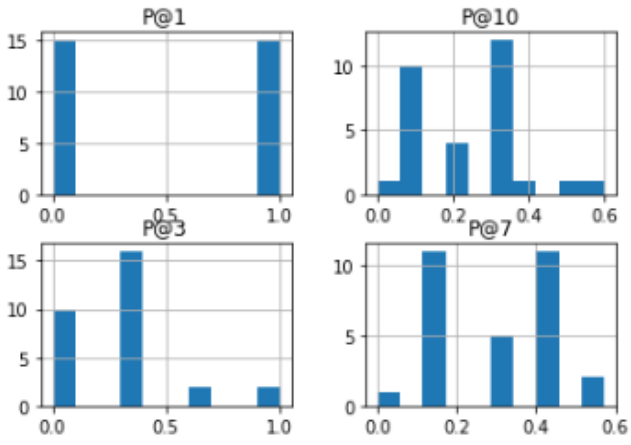


Булев поиск

[31]:

	P@1	P@3	P@7	P@10
служить всю жизнь государству	1	0.333333	0.285714	0.3
соглашение о свободной торговле	1	0.333333	0.142857	0.1
зимний дворец	0	0.000000	0.428571	0.3
российская империя	1	0.333333	0.428571	0.4
англо русской войне	0	0.000000	0.285714	0.2
всероссийский съезд советов	1	0.333333	0.285714	0.2
транссибирская магистраль	0	0.333333	0.142857	0.1
свидетели иеговы	1	0.333333	0.142857	0.1
владимир ильич ульянов	1	0.333333	0.142857	0.1
крымский мост	0	0.000000	0.000000	0.0
ионообменные смолы	1	1.000000	0.428571	0.3
красный террор	1	0.333333	0.142857	0.1
аутсайдер моя жизнь как интрига	1	0.333333	0.142857	0.1
день защитника отечества	0	0.666667	0.285714	0.2
день пограничника	1	0.333333	0.428571	0.3
ударный музыкальный инструмент	0	0.333333	0.571429	0.6
генеральный штаб вооружённых сил российской федерации	0	0.000000	0.428571	0.3
чирлидинг	0	0.000000	0.428571	0.3
голштинских заявлений	1	0.333333	0.142857	0.1
быстрая сортировка	0	0.000000	0.142857	0.3
умберто боччони	1	0.333333	0.142857	0.1
охотники за привидениями	0	0.333333	0.285714	0.2
божественный пёс	0	0.000000	0.142857	0.1
эдди мерфи	0	0.333333	0.571429	0.5
бегущий по льду	1	0.666667	0.428571	0.3
человек кусает собаку	1	1.000000	0.428571	0.3
военный коммунизм	1	0.333333	0.142857	0.1
петропавловская крепость	0	0.000000	0.428571	0.3
эрмитаж	0	0.000000	0.428571	0.3
ссср	0	0.000000	0.428571	0.3

Распределение оценок



Подробнее о результатах булева поиска:

Out[29]:

	P@1	P@3	P@7	P@10
count	30.000000	30.000000	30.000000	30.000000
mean	0.500000	0.288889	0.295238	0.230000
std	0.508548	0.273102	0.154379	0.134293
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.142857	0.100000
50%	0.500000	0.333333	0.285714	0.250000
75%	1.000000	0.333333	0.428571	0.300000
max	1.000000	1.000000	0.571429	0.600000

Сравнение с Wiki и Google

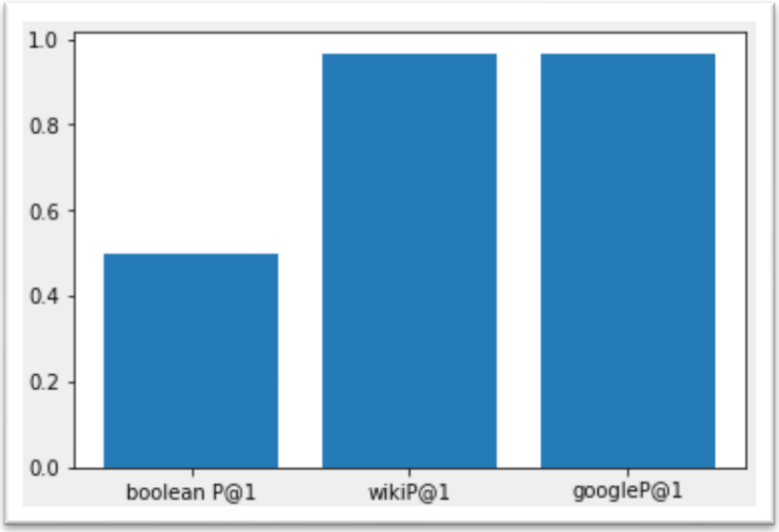


Рис. 1 Сравнение по метрике P@1

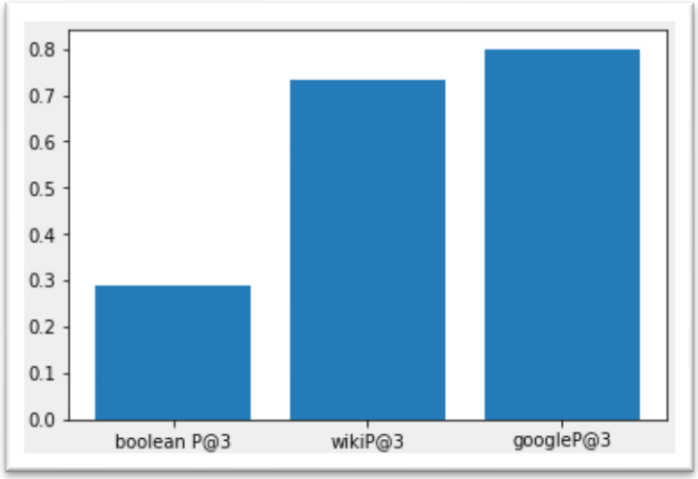


Рис. 2. Сравнение по метрике P@3

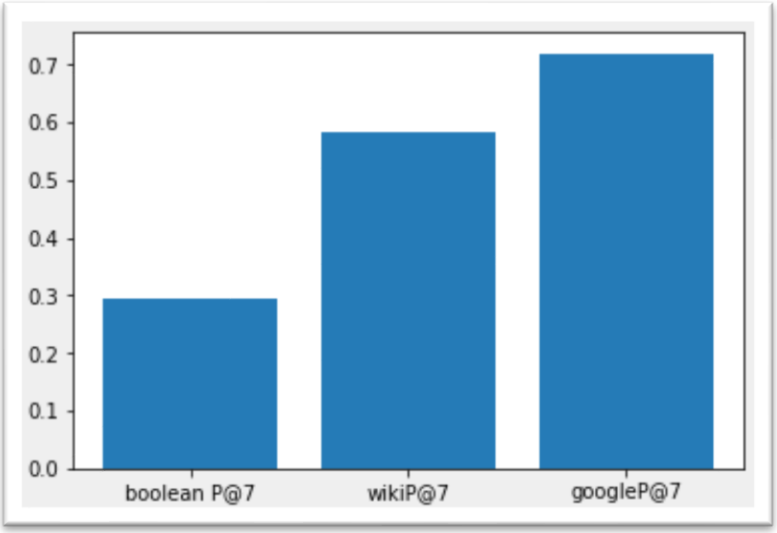


Рис. 3. Сравнение по метрике P@7

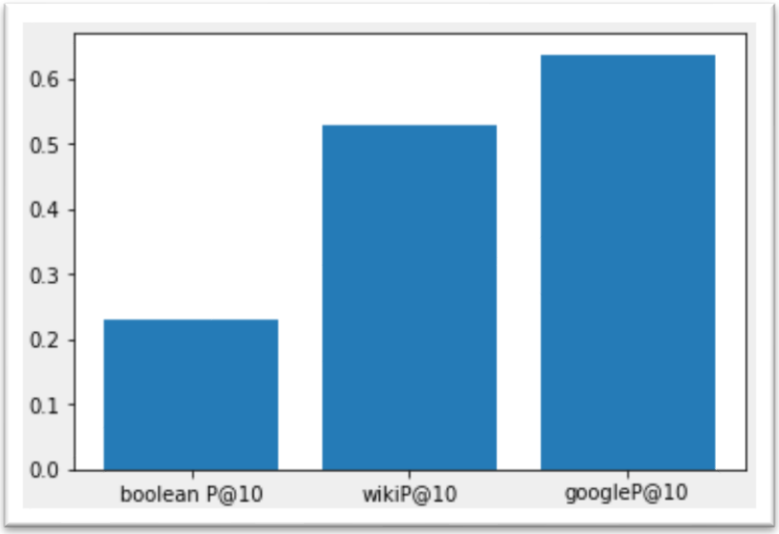


Рис. 4. Сравнение по метрике P@10

Анализируя результаты, можно заметить большое превосходство поиска от Google над булевым поиском. Стоит заметить, что булев поиск просто находит вхождение слов запроса в документ и не ранжирует никак полученные документы, что сильно сказывается на оценке поиска (в данном случае на точности). В теории можно рассмотреть это так: булев поиск ранжирует документы по их номеру, и посчитать оценки DCG и NDCG, однако это будет сильно притянуто. Также стоит заметить превосходство поиска Google над выдачей Википедии, что очень заметно. Однако, если посмотреть на распределения оценок этих поисковых выдач то они относительно похожи. В тоже самое время распределение выдачи Булева поиска очень разрежено