

Московский Авиационный Институт
(Научный Исследовательский Институт)

Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Отчёт по лабораторной работе №2 по курсу «Информационный поиск»

Выполнил: Ефименко Н.А.

Группа: 8О-106М

Преподаватель: Калинин А.Л.

Москва, 2019

ЛР2: Оценка качества поиска

Постановка задачи:

Необходимо оценить качество своего поиска и сравнить их с двумя альтернативами (для Википедии можно собственный поиск по Википедии, поиск Google или Яндекса с ограничением по сайту Википедии). Как минимум, нужно измерить P, DCG, NDCG, ERR уровней @1, @3 и @5, приветствуется использование дополнительных метрик качества.

Для оценки качества необходимо придумать 25 запросов, отражающих интересы пользователей или, если есть доступ к настоящим запросам пользователей, то выбрать репрезентативную подборку.

Привести анализ результатов оценки качества. Какие у какой поисковой системы сильные и слабые стороны? Как можно бороться с недостатками, что можно сделать, чтобы улучшить качество?

Этапы выполнения:

- Придумать 25 поисковых запросов по выбранной тематике учитывая интересы пользователей
- Получить поисковую выдачу с Google и Википедии, учитывая только те статьи, которые содержатся в корпусе документа, полученном на предыдущем этапе
- Представить полученные данные в удобном для оценки виде
- Оценка
- Измерить P, CG, DCG, NDCG, ERR уровней @1, @3 и @5
- Представить измеренные данные в удобном для анализа виде
- Анализ полученных измерений

Выполнение:

Чтобы обеспечить качественную оценку поиска поисковых движков нужно построить поисковый индекс на одном корпусе документов лишь тогда оценка будет качественной. Для сравнения с поисковым движком реализованным в курсе “Информационный поиск” была выбрана технология Elasticsearch. Она предоставляет удобные настройки для полнотекстового поиска. Представляя собой документоориентированную базу данных она обеспечивает простую связь с клиентским приложением по средством протокола http что очень удобно.

В качестве корпуса для сравнения была выбрана подвыборка из корпуса документов из ЛР-1: Документы из категории Википедия ‘Category:Computer programming’

Тестовые запросы:

- python
- very simple graph algorithm with code example
- where to apply map reduce
- emulate ray movement
- what are the game engines
- ray casting
- faster sorting algorithm
- modern optimization methods
- new programming languages
- old dead programming languages
- esoteric programming languages
- django framework
- django & framework & !(python)
- (redux | flux | MVVC) & architecture
- SPA frontend frameworks
- most popular framework
- big data algorithms
- popular search engine

- $n \cdot \log(n)$

Figure 1

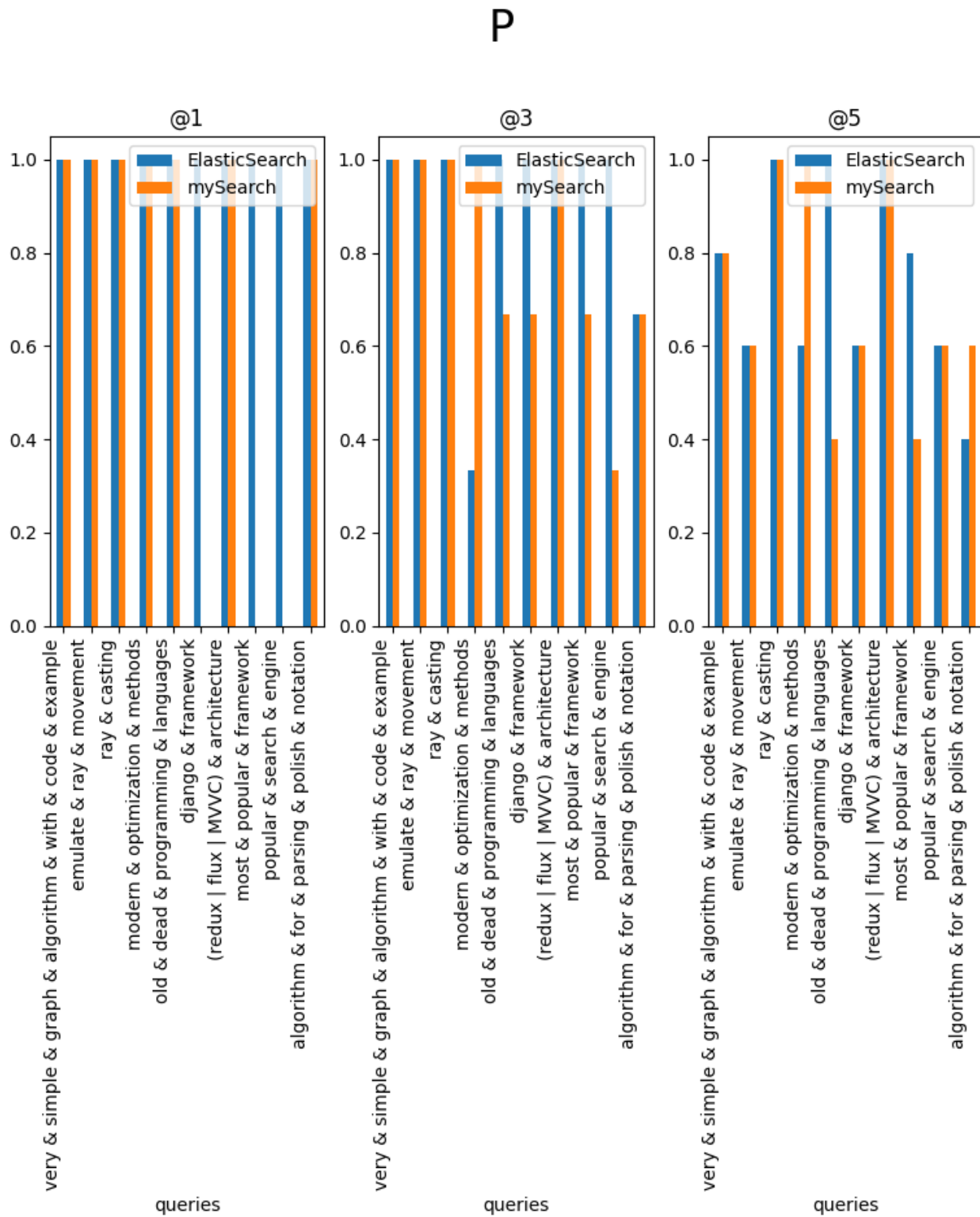


Figure 2

CG

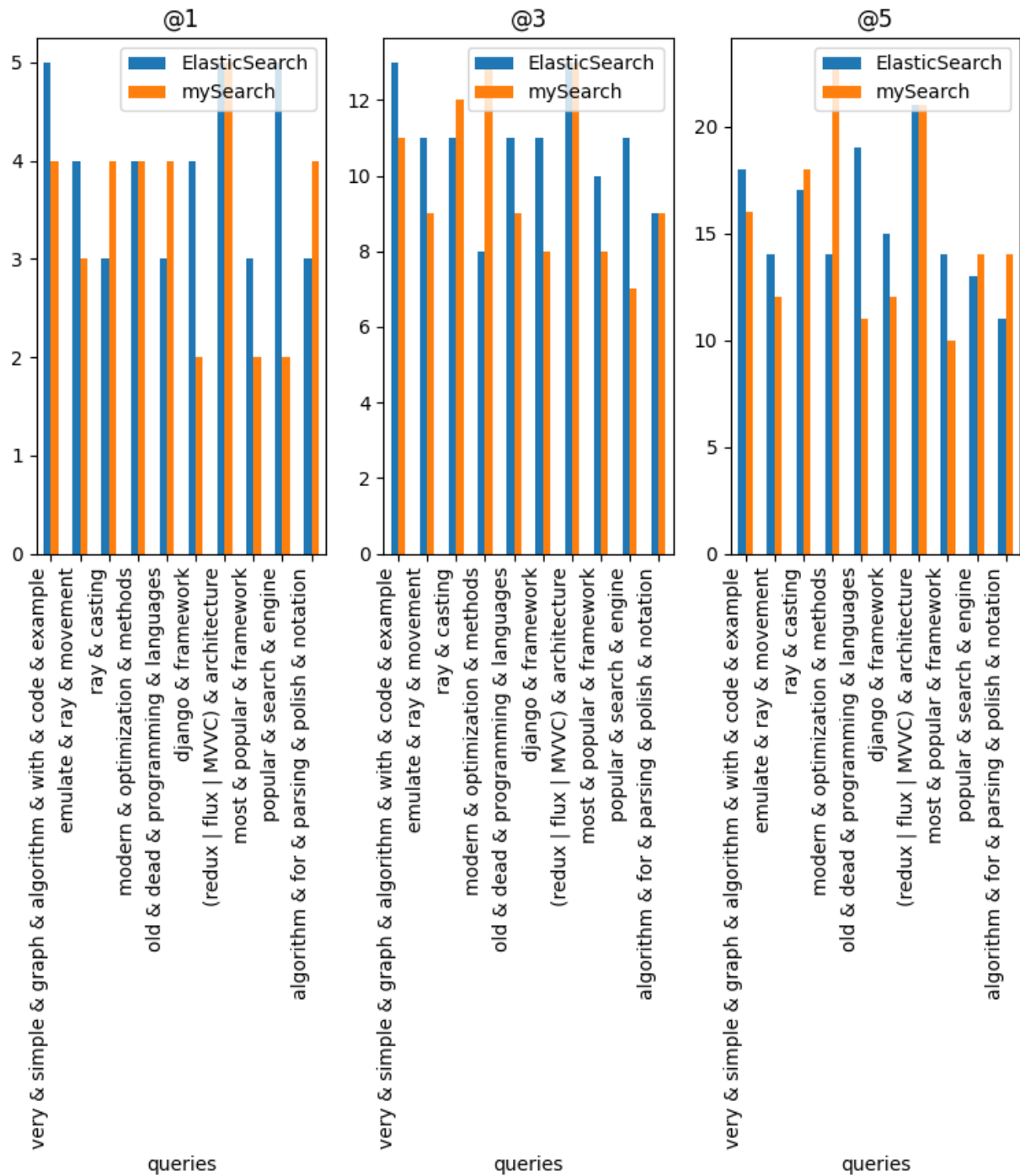


Figure 3

DCG

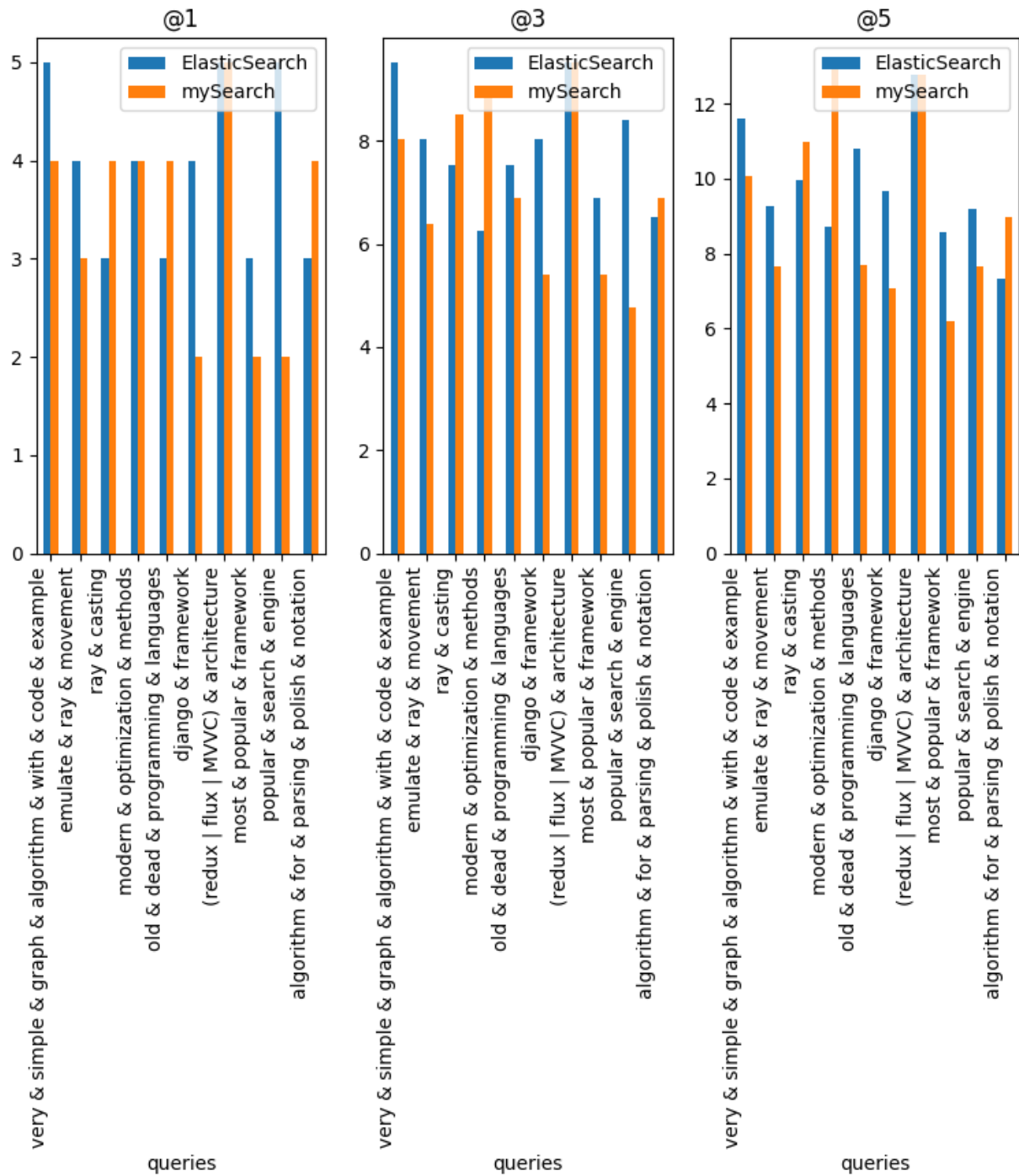


Figure 4

NDCG

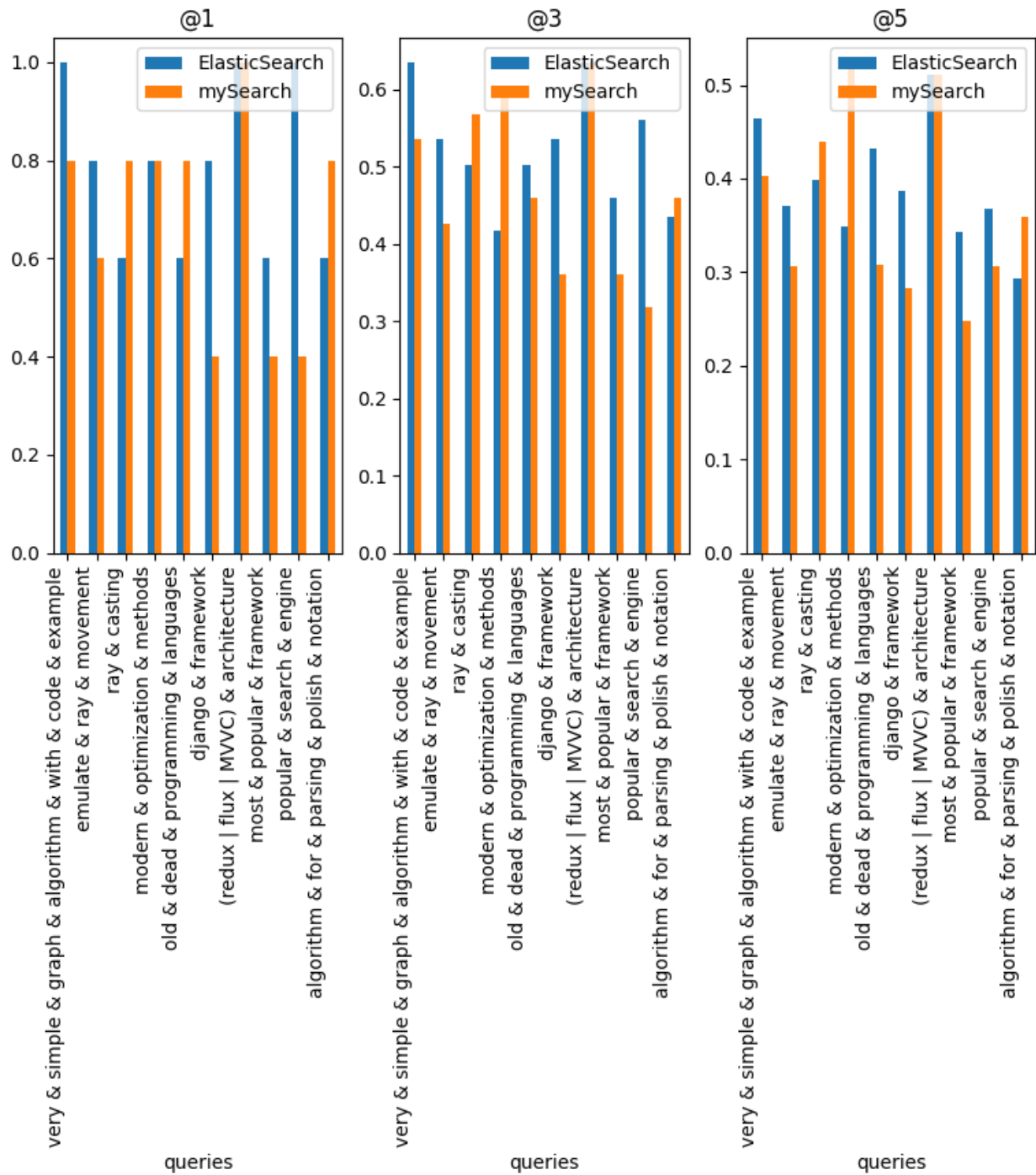
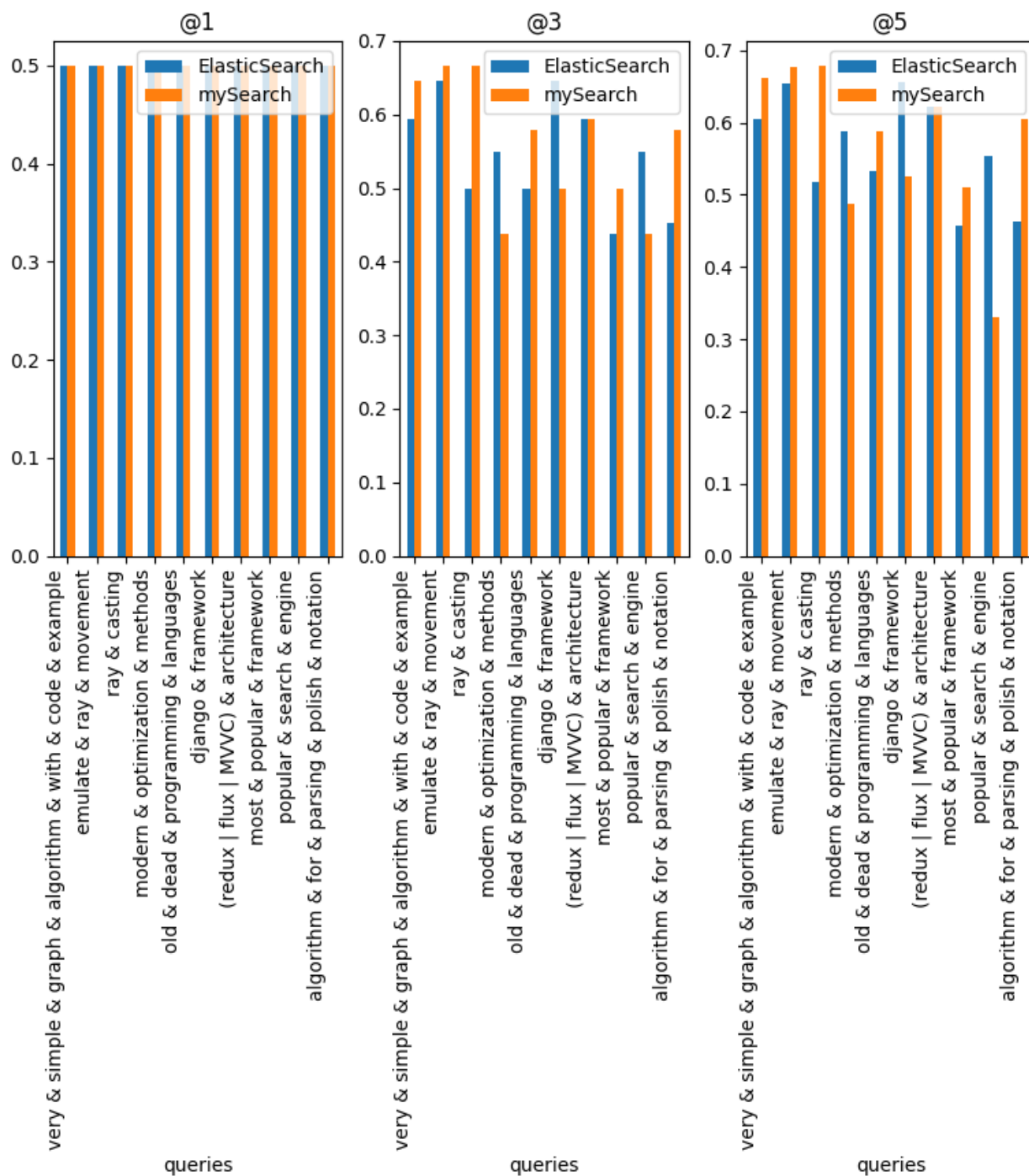


Figure 5

ERR



x= y=0.577168

Из графиков можно заключить, что на заданном наборе документов движки дают разные результаты. Видно что алгоритмы ранжирования в них реализованы по-другому однако неплохо справляются со своей работой. Выводы сделана на основе предоставленных оценок, которые в идеале должны проставляться экспертами (не 1 человеком)

в соответствующей области. Кроме того на запросах мой поиск просто не дает результата из-за простого алгоритма лемматизации и неучета числовых и знаковый запросов ($n \cdot \log(n)$)

Вывод

В ходе выполнения лабораторной работы были придуманы запросы и получены оценки качества поиска на них. Результаты полученные в лабораторной работе не могут давать сравнение двух поисков в полной мере, а лишь служат отправной точкой для дальнейшего сравнения своего поискового решения с другими решениями.