МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика»

Лабораторная работа №*4*
по курсу «Информационный поиск»

«Стемминг»

Выполнил: *Ермаков Ярослав Валерьевич*
Группа: *М8О-407Б-22*
Преподаватели: *А.А. Кухтичев*

Москва, *2025*

Условия

Добавить в созданную поисковую систему лемматизацию. В простейшем случае, это просто поиск без учёта словоформ. В более сложном случае, можно давать бонус большего размера за точное совпадение слов.

Лемматизацию можно добавлять на этапе индексации, можно на этапе выполнения поискового запроса. В отчёте должна быть включена оценка качества поиска, после внедрения лемматизации. Стало ли лучше? Изучите запросы, где качество ухудшилось. Объясните причину ухудшения и как можно было бы улучшить качество поиска по этим запросам, не ухудшая остальные запросы?

Описание

Цель лабораторной работы - добавить в поисковую систему стемминг, чтобы поиск не зависел от словоформ и разные грамматические формы слова сопоставлялись как один терм. Стемминг применяется как нормализация токенов и позволяет повысить полноту поиска: документы находятся даже тогда, когда ключевое слово в тексте встречается в другой форме. Для корректной работы одинаковые правила обработки должны применяться как при индексации, так и при обработке поискового запроса.

Журнал выполнения и решение проблем

Стемминг был внедрён как дополнительный шаг после токенизации: сначала выделяются токены по правилам и выполняется базовая нормализация , после чего при включённом режиме токен преобразуется стеммером. Основная практическая проблема при внедрении заключалась в том, что стемминг может не только улучшать выдачу, но и ухудшать её. Это проявляется в случаях, когда разные слова сводятся к одному стему, из-за чего в результатах появляются нерелевантные документы. Также на “технических” токенах стемминг может давать нестабильные эффекты, поэтому такие токены целесообразно либо не стеммировать, либо обрабатывать отдельным правилом. В ходе проверки также было важно обеспечить согласованность режимов: если индекс построен со стеммингом, то запрос должен обрабатываться тем же режимом, иначе результаты будут неполными.

Исходный код

В системе реализован режим включения стемминга через параметр запуска (например, --stemming 0/1). При построении индекса каждый токен перед добавлением в индекс проходит одинаковую цепочку преобразований, включая стемминг в соответствующем режиме. При выполнении запроса входная строка токенизируется теми же правилами и, при включённом режиме, токены также пропускаются через стеммер, после чего

выполняется поиск по индексу. Такой подход делает поиск устойчивым к словоформам и обеспечивает корректное сопоставление запроса и данных индекса.

Оценка качества поиска и анализ ухудшений

Оценка качества выполнялась сравнением выдачи для одинаковых запросов в двух режимах: без стемминга и со стеммингом. В большинстве случаев стемминг улучшает полноту: увеличивается количество релевантных документов, особенно для запросов с русскими словами, где в текстах часто встречаются разные падежи, числа и производные формы. При этом обнаруживаются запросы, где качество ухудшается. Основные причины ухудшения связаны с “слипанием” разных слов в один стем , из-за чего выдача расширяется за счёт нерелевантных документов. Также ухудшение может происходить для коротких токенов и для терминов с цифрами/дефисами, где преобразование не соответствует ожиданиям пользователя.

Чтобы улучшить качество на таких запросах, не ухудшая остальные, можно использовать комбинированный подход: выполнять поиск по стемам, но дополнительно повышать приоритет документов, где встречается точная форма слова из запроса, либо ограничивать применение стемминга для коротких токенов и технических обозначений. Ещё один практический вариант - хранить в индексе как стем, так и исходную форму и учитывать это на этапе ранжирования.

Выводы

В лабораторной работе в поисковую систему был добавлен стемминг, применяемый согласованно при индексации и при обработке запросов. В результате поиск стал менее зависимым от словоформ, что повысило полноту выдачи. При этом выявлены запросы, где качество может снижаться из-за “слипания” разных слов в один стем и из-за особенностей технических токенов. Для снижения негативного эффекта целесообразно ограничивать стемминг для отдельных классов токенов и учитывать точные совпадения при формировании выдачи.