тов в разных валютах, пополнение счетов, отправка средств с одного на другой счет. Сотрудник банка может просматривать счета клиентов и блокировать их, отправлять сообщения клиентам.

ЛИТЕРАТУРА

1 Интернет-банкинг // Википедия [Электронный ресурс]. – 2020. – Режим доступа: https://ru.wikipedia.org/wiki/Интернет-банкинг – Дата доступа: 18.04.2020.

2 Hibernate // Википедия [Электронный ресурс]. – 2020. – Режим доступа: https://ru.wikipedia.org/wiki/Hibernate_(библиотека) – Дата доступа: 18.04.2020.

3 Электронный ресурс: https://bootstrap-4.ru/– Дата доступа: 18.04.2020.

УДК 004. 934.2

Магистрант А. С. Демещик
Науч. рук. ст. преп. И. Г. Сухорукова
(кафедра программной инженерии, БГТУ)

АНАЛИЗ АЛГОРИТМОВ ПРЕДОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ

При анализе текста очевидно, что не все слова в тексте несут полезную информацию. Кроме того, в силу гибкости естественных языков формально различные слова (синонимы и т. п.) на самом деле означают одинаковые понятия. Таким образом, удаление неинформативных слов, а также приведение близких по смыслу слов к единой форме значительно сокращают время анализа текстов. Устранение описанных проблем выполняется на этапе предварительной обработки текста. Для предварительной обработки текста используют различные подходы [1].

Очистка текстовых данных подразумевает удаление HTML-тегов, мета-тегов, различного «информационного мусора».

Фильтрации стоп-слов. Стоп-словами называются слова, которые являются вспомогательными и несут мало информации о содержании документа. Обычно заранее составляются списки таких слов, и в процессе предварительной обработки они удаляются из текста.

Стемминг заключается в преобразовании каждого слова к его нормальной форме. Нормальная форма исключает склонение слова, множественную форму, особенности устной речи и т. п. Например, слова "сжатие" и "сжатый" должны быть преобразованы в нормальную форму слова "сжимать". Алгоритмы морфологического разбора

учитывают языковые особенности и вследствие этого являются языковозависимыми алгоритмами.

N-граммы– это альтернатива морфологическому разбору и удалению стоп-слов. По сравнению со стеммингом или удалением стоп-слов, N-граммы менее чувствительны к грамматическим и типографическим ошибкам. Кроме того, N-граммы не требуют лингвистического представления слов, что делает данный прием более независимым от языка. Однако N-граммы, позволяя сделать текст более строгим, не решают проблему уменьшения количества неинформативных слов. Знак «_» заменяет пробелы и в будущем позволяет конвертировать последовательности в текст, корректно расставляя границы слов.

Подход *приведение к регистру* заключается в приведении всех слов к одному регистру, чтобы исключить случаи, когда «ТЕКСТ» и «текст» рассматриваются в различном контексте.

Кроме названных алгоритмов в работе использовался алгоритма TF-IDF. Его основная задача не обработка текста, а анализ оценки важности слова в корпусе текста, но его можно применять совместно с другими алгоритмами, например, для фильтрации текстового корпуса. Предварительно «пропустив» текст через TF-IDF, на выходе мы получаем матрицу важности слова в коллекции документов. Можно использовать эту информацию, чтобы убрать малозначительные и не очень важные слова из текста с помощью фильтрации стоп-слов, это поможет сделать текст более чистым от «информационного мусора».

ЛИТЕРАТУРА

1. Барсегян А.А. Анализ данных и процессов. 3-е издание - Санкт-Петербург: БХВ-Петербург, 2009

2. Кузьмина А., Баяндин Н.И. Технологии анализа данных. Москва: МЭСИ, 2011

3. Климов Д.В. Предобработка текстовых сообщений для метрического классификатора. /– МТИ, Москва –2017.

УДК 004.4

Студ. С. О. Гончар
Науч. рук. доц. А.П. Лащенко
(кафедра программной инженерии, БГТУ)

КРАУДФАНДИНГОВАЯ ПЛАТФОРМА

Краудфандинг – коллективное сотрудничество людей (доноров), которые добровольно объединяют свои деньги или другие ресурсы вместе, как правило, через Интернет, чтобы поддержать усилия других людей или организаций (реципиентов).