Титульный лист

Введение

В современном мире обработка текстовых данных является одной из самых востребованных и быстроразвивающихся отраслей. Большинство крупных компаний постоянно ищут способы оптимизировать свои бизнес-процессы, связанные с обработкой документов, писем или иной информации на естественном языке. В современных реалиях сложно представить организацию, для которой внедрение алгоритмов анализа текстовых данных не принесло бы существенной выгоды. Основными задачами данной отрасли являются:

1. Машинный перевод;
2. Классификация текстов;
3. Кластеризация текстов;
4. Извлечение информации;
5. Системы вопрос-ответ;
6. Генерация текстов;
7. Проверка правописания;
8. Оптическое распознавание текстов;
9. Проверка на соответствие образцу.

Машинный перевод – это процесс при котором текст, написанный на одном естественном языке, переводится на другой естественный язык с использованием специального программного обеспечения. Яркими примерами такого программного обеспечения являются сервисы «Google переводчик» и «Яндекс.переводчик». Они представляют собой Web-приложения, в которые можно загрузить текст и перевести его на другой язык, сохраняя с некоторым уровнем точности его смысловую нагрузку и синтаксис.

Классификация текстов – это процесс при котором специальное программное обеспечение может присвоить полученному тексту один из классов, заданных разработчиком. Например, новостной портал «Яндекс Дзен» может автоматически присвоить написанной статье категорию, в зависимости от ее наполнения.

Кластеризация текстов – это процесс, при котором специальное программное обеспечение получившее на вход некоторое количество текстовых документов может разделить их на заданное количество категорий.

Извлечение информации – это процесс, при котором специальное программное обеспечение, получившее на вход текстовый документ, может найти и вывести пользователю обособленные факты или конкретные участки текста, необходимы пользователю.

Системы вопрос-ответ – это системы, способные коммуницировать с пользователем, подбирая текст ответа из заранее заготовленного множества ответов, в зависимости от запроса пользователя.

Генерация текстов – это процесс, при котором из некоторого количества данных, например: темы, часть текста или полнотекстового документа специальное программное обеспечение генерирует текст, заданного разработчиком объема.

Проверка правописания – это итеративный процесс, часто встречающийся в текстовых редакторах или системах контроля документации на предприятии. Позволяющий проверять синтаксис, орфографию и пунктуацию вводимого пользователем текста после каждого введённого символа.

Для решения данных задач существуют разнообразные методы и алгоритмы, применение которых позволяет достигнуть необходимого результата. Используемые практики можно разделить на несколько уровней:

1. Уровень символов;
2. Уровень слов – морфология;
3. Уровень предложений – синтаксис;
4. Уровень смысла – семантика.

Классификация текстов один из самых широко используемых инструментов в современной обработке естественного языка. Он позволяет решать огромное количество задач, подавляющее большинство которых позволяет оптимизировать процессы, связанные с документооборотом на предприятии. Однако, классификация текстов может применяться и для удобства более широкого круга пользователей, например, в новостных агрегаторах или иных сервисах, работающих со статьями пользователей, например: Elsevier Journal Finder

Бесплатный онлайн-сервис, разработанных компанией RELX и используемый издательство Elsevier позволяет пользователь может найти наиболее точно соответствующие тематике статьи журналы издательства. С помощью этого сервиса возможно:

* Выбрать журнал для статьи пользователя, при отсутствии практического опыта в публикации;
* Выбрать журнал для работ, написанных в междисциплинарных областях;
* Найти журналы, позволяющие опубликовать статью бесплатно и в открытом доступе.

Для реализации подбора журнала требуются следующие входные данные:

1. Заголовок статьи;
2. Текст статьи;
3. Релевантные ключевые слова;
4. Предполагаемая пользователем сфера исследований.

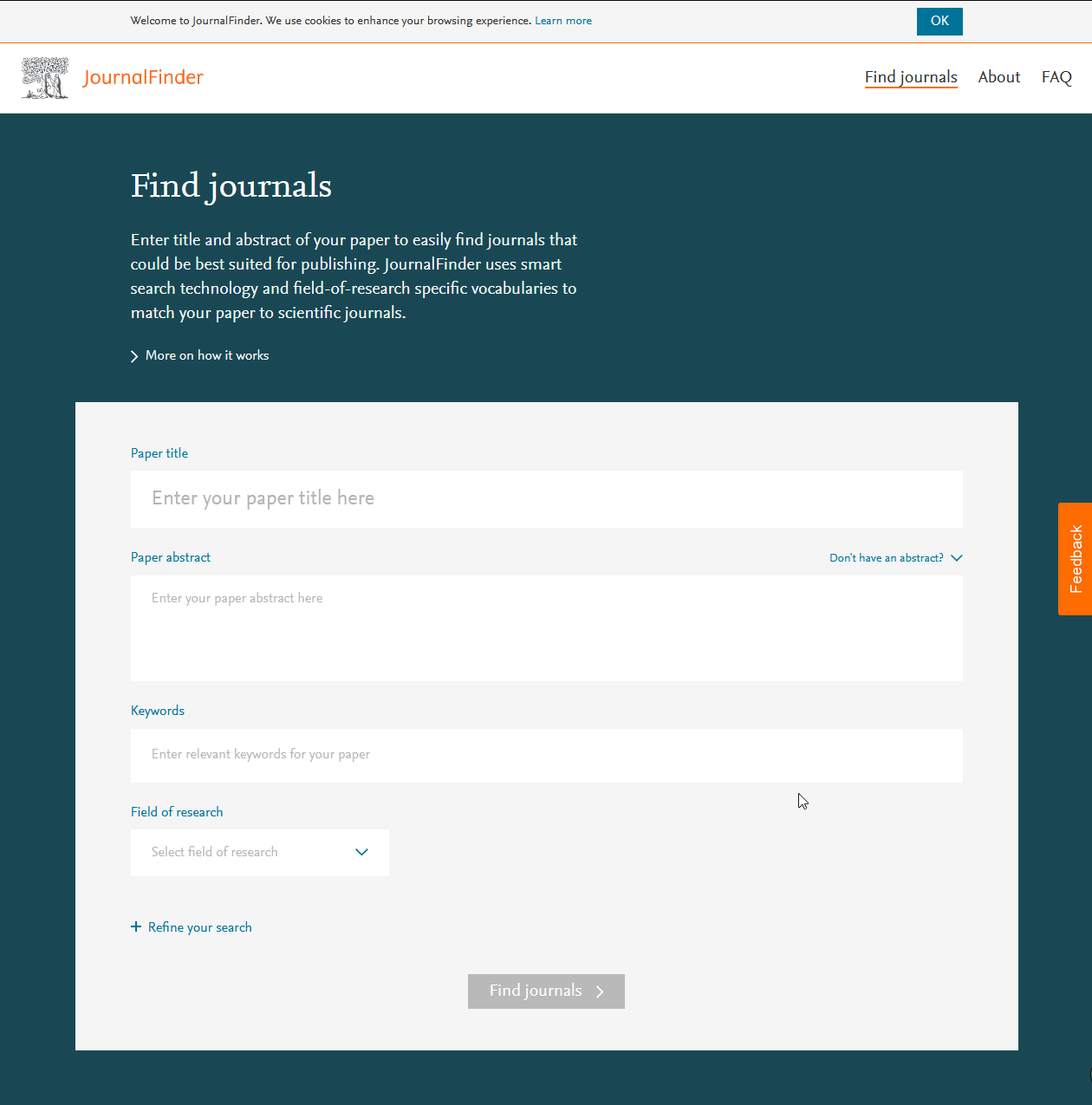


Рисунок 1 Интерфейс сервиса Journal Finder

Аналогичным функционалом обладают сервисы:

1. Springer Journal Suggester;
2. Manuscript matcher;
3. Edanz Journal Selector;
4. Researchers.One;
5. Transpose.

Несмотря на довольно большое количество сервисов, ни один из них не работает со статьями, написанными на русском языке. В связи с этим фактом имеется смысл в разработке аналога этих сервисов, работающего со статьями российских пользователей без необходимости перевода работ на английский язык и публикации в русскоязычных издательствах.

Для того, чтобы разработать сервис, способный предлагать пользователю журналы для публикации необходимо решить ряд задач:

1. Разработать и реализовать схему взаимодействия с пользователем;
2. Выбрать и разметить корпуса текстов, необходимых для обучения алгоритма и последующей проверки на точность.
3. Решить задачу преобразования текста в числовой формат для обработки компьютером.
4. Изучить применяемые алгоритмы классификации текстов и выбрать наиболее подходящую.

Теория используемых методов.

Для взаимодействия с пользователем могут применяться различные методы и виды сервисов, например:

1. Веб-приложение

Приложение, развернутое на сервере, имеющие CMS (Content Management System/Систему управление контентом, базу данных и отображающееся в браузере клиента.

1. Веб-сайт;

Набор файлов HTML/CSS/JS располагающийся на сервера и backend скрипт, позволяющий реализовать взаимодействие между клиентом и сервером.

1. Чат-бот;

Чат, развернутый на сервере третьего лица, позволяющий, информационной системе, выполняющей действия автоматически или по заданному расписанию выполнять какие-либо действия через интерфейсы, предназначенные для пользователей.

1. Почтовая система;

Система, позволяющее взаимодействовать с пользователем по средствам обмена электронной почтой.

1. Приложение типа Клиент/Сервер.

Приложение, имеющее две составляющие. Первое – приложение, устанавливаемое на компьютер пользователя, представляющее собой набор интерфейсов и второе – приложение, развернутое на сервере, занятое обработкой информации, предоставленной клиентским приложением.