Оглавление

[Введение 3](#_Toc515289794)

[Глоссарий 3](#_Toc515289795)

[Описание предметной области 4](#_Toc515289796)

[Неформальная постановка задачи 5](#_Toc515289797)

[Обзор существующих решений 7](#_Toc515289798)

[Постановка задачи 7](#_Toc515289799)

[Требования к окружению 8](#_Toc515289800)

[Требования к аппаратному обеспечению 8](#_Toc515289801)

[Требования к программному обеспечению 8](#_Toc515289802)

[Спецификация данных 8](#_Toc515289803)

[Формат данных 8](#_Toc515289804)

[Функциональные требования 16](#_Toc515289805)

[Требования к интерфейсу 17](#_Toc515289806)

[Интерфейс командной строки 17](#_Toc515289807)

[Графический интерфейс пользователя 17](#_Toc515289808)

[Проект 22](#_Toc515289809)

[Описание метода 24](#_Toc515289810)

[Извлечение именованных сущностей 24](#_Toc515289811)

[Оценка качества модели 28](#_Toc515289812)

[Результаты 29](#_Toc515289813)

[Извлечение библиографических ссылок 30](#_Toc515289814)

[Извлечение локаций с координатной привязкой 31](#_Toc515289815)

[Извлечение ключевых слов и словосочетаний 32](#_Toc515289816)

**Аннотация**

В данной работе решается задача автоматизации подготовки метаинформации из документов с научным уклоном. Для этого реализуются модели, которые извлекают метаданные из текста. Так же реализуется приложение, которое дает возможность дополнить, отредактировать и опубликовать эти метаданные в систему управления пространственной информацией – GeoNetwork.

# **Введение**

## **Глоссарий**

RNN - Recurrent Neural Networks (рекуррентные нейронные сети)

LSTM - Long short-term memory (сети с долгой краткосрочной памятью)

CRF – Conditional random fields (Условные случайные поля) – алгоритм машинного обучения с учителем. Разновидность марковских случайных полей

SVM - Support vector machine (Метод опорных векторов) – алгоритмы [обучения с учителем](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81_%D1%83%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D0%B5%D0%BC). Используется для задач линейной классификации.

## **Описание предметной области**

Тихоокеанский научно-исследовательский рыбохозяйственный центр (ФГБНУ «ТИНРО-Центр») — Федеральное государственное бюджетное научное учреждение и научно-исследовательский центр на Дальнем Востоке России.

Центр является самой крупной рыбохозяйственной научно-исследовательской организацией России. Его история началась в 1925 г., когда был создан Тихоокеанской научно-промысловой станции (ТОНС). Позже, в 1928 г. была преобразована в Тихоокеанский институт рыбного хозяйства (ТИРХ), а в 1934 г. институт был переименован в Тихоокеанский научно-исследовательский институт рыбного хозяйства и океанографии (ТИНРО).

Институт занимается такими направлениями исследований как: морская биология, [биоценология](https://ru.wikipedia.org/wiki/%D0%91%D0%B8%D0%BE%D1%86%D0%B5%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F), [океанография](https://ru.wikipedia.org/wiki/%D0%9E%D0%BA%D0%B5%D0%B0%D0%BD%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F), разработка новых комплексных технологий переработки морского сырья.

Кроме подразделений Владивостока, в состав ТИНРО-Центра входят Хабаровское и Чукотское отделения, а также База научно-исследовательского флота (БИФ).

ТИНРО-Центр входит в структуру [ВНИРО](https://ru.wikipedia.org/wiki/%D0%92%D0%9D%D0%98%D0%A0%D0%9E) (Всероссийский научно-исследовательский институт рыбного хозяйства и океанографии).

ТИНРО-Центр объединяет все дальневосточные рыбохозяйственные научные организации и координирует исследования на дальневосточном рыбопромысловом бассейне. Интеграция рыбохозяйственной науки в последние годы особенно стала актуальной. Следствием этого стало создание в 2000 г. единой дальневосточной рыбохозяйственной научной структуры – Ассоциации «Научно-техническое объединение ТИНРО».

Также ТИНРО-Центр представляет Россию в международной организации PICES.

**The North Pacific Marine Science Organization** ([PICES](http://meetings.pices.int/about/organizationstructure), Северо-тихоокеанская морская научная организация) - межправительственная научная организация, которая была создана в 1992 году. Ее нынешними членами являются: Канада, Китайская Народная Республика, Япония, Республика Корея, Российская Федерация и Соединенные Штаты Америка.

Цели Организации заключаются в следующем:

* Содействовать и координировать морские исследования в северной части Тихого океана и прилегающих морях.
* Продвигать научные знания об окружающей среде океана, погоде и изменении климата, живых организмах и их экосистемах, а также о воздействии человеческой деятельности на эти экосистемы.
* Содействовать сбору и обмену научной информацией по этим вопросам.

Для управления и публикации метаданных по пространственным данным и связанным с ними информации. Используется система GeoNetwork.

GeoNetwork — это система управления ресурсами с пространственной привязкой. Он предоставляет мощные функции редактирования и поиска метаданных, а также интерактивный просмоторщик веб-карт. В настоящее время он используется во многих проектах по всему миру.

GeoNetwork предоставляет простой в использовании веб-интерфейс для поиска информации, для этого используются каталоги с метаданными.

Извлечение метаописаний широко применяется в задачи информационного поиска. Эта процедура помогает более эффективно искать информацию, удовлетворяющую критериям поиска.

Составление метаописаний – это многоэтапный процесс. Данные проходят несколько стадий проверок качества, прежде чем быть опубликованными для публичного просмотра. Сейчас метаописания составляются вручную. Из-за того, что это затратный по времени процесс, возникла необходимость его автоматизации.

Извлечение метаданных – это задача выделения структурированных данных из источников с неструктурированной или слабо структурированной информацией, которая связана с обработкой данных на естественных языках.

## **Неформальная постановка задачи**

В общем случае задача состоит в том, чтобы из предложенного текста, выделить информацию об именованных объектах в тексте (информацию о классах, которые они соответствуют), ключевых словах, авторах, и т. д. Набор классов фиксируется заранее.

Конкретно в моей задаче – нужно извлечь информацию из текста, предоставить интерфейс, где можно её отредактировать и дополнить, а также опубликовать в систему GeoNetwork.

Одной из подзадач в извлечении метаописаний – это выделение именованных сущностей (Named Entity Recognition, NER). Цель этой задачи выделить в тексте объектов определенного типа, это могут быть имена людей, географические объекты, биологические объекты и т.д.

Эта задача начала активно развиваться в середине 90-х, но и сейчас она остается актуальной, так как с каждым днем появляется потребность классификации, данный с новыми классами.

Под термином именованная сущность (named entity) понимается объект, который имеет такие характеристики как имя или идентификатор.

Сейчас существует много разных стратегий и подходов к решению поставленной задачи. В основном это системы, основанные на словарях, правилах или машинном обучении. Наиболее популярны методы машинного обучения с учителем. Такая система изначально тренируется на множестве примеров. Плюс этой системы в том, что её можно использовать на произвольных данных. Она имеет преимущество, по сравнению с моделями, основанными на словарях, так как может распознавать данные, которые раньше не встречала.

Ещё одна подзадача извлечения данных из текста – это выделение ключевых слов и словосочетаний. Ключевые слова – важные слова и словосочетания, которые дают высокоуровневое описание текста. Выделение таких слов, сделает информационный поиск более эффективным и точным.

Сейчас существуют алгоритмы, основанные на правилах и машинном обучении. Задача сводится к фильтрации текста и далее производится извлечение словосочетаний, из которых потом выделяются наиболее важные.

Следующая задача - связывание именованных сущностей с пространственным местоположением.

После извлечения географических объектов, хотелось бы получить информацию о местоположении объекта. Данную задачу можно решить с помощью словаря объектов и их координат, а для более качественного поиска таких объектов использовать алгоритма неточного сравнения строк.

Эта подзадача не совсем связана с извлечением информации, но она так же актуальна, так как в системе GeoNetwork поиск может осуществляться по пространственным координатам.

Все перечисленное делает задачу извлечения метаописаний достаточно полезной и интересной не только с точки зрения исследователей, но и с точки зрения автоматизации процесса.

## **Обзор существующих решений**

На просторах сети интернет существует множество систем, которые так или иначе решают задачу извлечения информации из неструктурированного или слабо структурированного текста.

Приведу некоторые примеры.

**Cermine** – это web сервис с открытым исходным кодом. У него неплохой интерфейс, но на предложенных ему документах он справился не самым лучшим образом.

Этот сервис использует машинное обучение, а именно CRF, SVM и K-means.

У него модульная структура, где для каждой подзадачи используются свои алгоритмы.

**GROBID** – сервис, который предлагается инсталлировать на свою машину и использовать.

Он имеет открытый исходный код и использует SVM для извлечения информации.

**ParseCit** – это бесплатная программа для извлечения информации из текстов.

Он имеет открытый исходный код и использует CRF для извлечения информации.

**Text Razor** сервис для извлечения именованных сущностей из текста.

Несмотря на то, что представленные решения справляются с задачей извлечения информации, в той или иной степени, не одно из них не предоставляет данные в нужном формате, для публикации в GeoNetwork. Также в них нет специфических функция, таких как связь географических объектов с координатами.

Следовательно, возникает необходимость разработать свою систему для извлечения метаописаний.

## **Постановка задачи**

Задача данной работы состоит в разработке программного продукта для извлечения метаописаний из документов с последующим редактированием и публикацией в системе GeoNetwork. Для этого требуется:

1. Исследовать существующие методы извлечения информации из текстов;
2. Разработать методы извлечения информации из текстов;
3. Выполнить программную реализацию разработанных методов;
4. Реализовать клиентское приложение для системы GeoNetwork, в которое будет загружаться текстовый документ и извлекать метаданные из него. Также должна быть возможность отредактировать и дополнить эти данные, а затем опубликовать их.
5. Провести тестирование разработанной системы.

# **Требования к окружению**

## **Требования к аппаратному обеспечению**

**Минимальные требования к аппаратному обеспечению:**

Процессор: 2 ГГц

Оперативная память: 2 ГБ

Жесткий диск: 3 ГБ свободного места

**Рекомендованные требования к аппаратному обеспечению:**

Процессор: 3 ГГц

Оперативная память: 4 ГБ

Жесткий диск: 3 ГБ свободного места

Видеокарта с поддержкой CUDA.

## **Требования к программному обеспечению**

ОС: Windwos 10 / GNU/Linux

Окружение: python 3.5 и выше

Зависимости: tensorflow 1.4.0 и выше

Если tensorflow с поддержкой gpu: CUDA Toolkit 9.0, cuDBB v7.0

GUI: PyQT5

Другие зависимости: numpy, nltk, pdfminer

# **Спецификация данных**

## **Формат данных**

После извлечения информации из pdf документа, её нужно опубликовать в GeoNetwork, но прежде её нужно представить в стандартизированном виде, понятном системе. Для этого используются определённые стандарты для описания метаданных.

Метаданные, обычно определяются как “данные о данных”. Они представляют собой структурированный набор информации, который описывает данные, хранящиеся в административных системах. Метаданные могут содержать краткое содержание, свойства объекта, данные местоположения, а также информацию, связанную с его созданием.

Стандарты метаданных представляют собой правила для описания данных, что позволяет пользователям оценивать полезность данных.

GeoNetwork поддерживает несколько форматов метаданных, из которых были выбраны наиболее популярные: ISO19115 и FGDC

ISO19115 Geographic information Metadata (Географическая информация. Метаданные) – международный стандарт метаданных для описания географической информации.

ISO 19115 обеспечивает универсальное, независимое от кодирования представление метаданных, для описания пространственных, временных и других характеристик географических данных.

Описывается с помощью языка XML.

|  |  |
| --- | --- |
| gdm:dateStamp | Дата создания метаданных |
| gmd:language | Язык создания метаданных |
| gmd:characterSet | Обозначение и наименование стандарта символьной кодировки данных |
| gmd:hierarchyLevel | Наименование уровня иерархии, для которого созданы метаданные |
| gmd:metadataStandardName | Обозначение и наименование используемого стандарта метаданных (включая обозначение и наименование профиля) |
| gmd:metadataStandardVersion | Версия используемого стандарта (профиля) метаданных |
| gmd:dataSetURI | URI данные |
| gmd:topicCategory | Предметная область данных |
| gmd:topicCategory> gmd:MD\_TopicCategoryCode | Описание предметных областей |
| gmd:abstract | Краткое описание содержания набора данных |

*“Табл. 1. Основные данные”*

|  |  |
| --- | --- |
| gmd:contact | Субъект, ответственный за метаданные |
| gmd:contact>CI\_ResponsibleParty | Информация об ответственном субъекте |
| gmd:contact>CI\_ResponsibleParty> individualName | Имя субъекта |
| gmd:contact>CI\_ResponsibleParty> organisationName | Наименование юридического лица |
| gmd:contact>contactInfo | Информация о субъекте |
| gmd:contact>contactInfo> CI\_Contact | Контактная информация для связи с ответственным субъектом |
| gmd:contact>contactInfo> CI\_Contact >phone | Номера контактных телефонов |
| gmd:contact>contactInfo> CI\_Contact >phone >CI\_Telephone | Номера контактных телефонов для связи с ответственным субъектом |
| gmd:contact>contactInfo> CI\_Contact >phone >CI\_Telephone >voice | Номер контактного телефона для связи с ответственным субъектом |
| gmd:contact>contactInfo> CI\_Contact >phone >CI\_Telephone >facsimile | Номер контактного факса для связи с ответственным субъектом |
| gmd:contact>contactInfo> CI\_Contact >address | Информация о местоположении субъекта |
| gmd:contact>contactInfo> CI\_Contact >address >CI\_Address | Адрес |
| gmd:contact>contactInfo> CI\_Contact >address >CI\_Address >deliveryPoint | Адрес |
| gmd:contact>contactInfo> CI\_Contact >address >city | Город |
| gmd:contact>contactInfo> CI\_Contact >address >administrativeArea | Область, край |
| gmd:contact>contactInfo> CI\_Contact >address >postalCode | Почтовый или другой индекс |
| gmd:contact>contactInfo> CI\_Contact >address >country | Страна |
| gmd:contact>contactInfo> CI\_Contact >address >electronicMailAddress | Адрес электронной почты |

*“Табл. 2. Данные об ответственном субъекте”*

|  |  |
| --- | --- |
| gmd:identificationInfo | Общая информация о данных |
| gmd:identificationInfo >MD\_DataIdentification | Информация о данных |
| gmd:identificationInfo >MD\_DataIdentification >citation | Ссылка на описание и информации |
| gmd:identificationInfo >MD\_DataIdentification >citation >CI\_Citation | Ссылка на описание и информация об данных |
| gmd:identificationInfo >MD\_DataIdentification >citation >CI\_Citation >title | Название |
| gmd:identificationInfo >MD\_DataIdentification >citation >CI\_Citation>date | Дата |
| gmd:identificationInfo >MD\_DataIdentification >citation >CI\_Citation>date >CI\_Date | Дата |
| gmd:identificationInfo >MD\_DataIdentification >citation >CI\_Citation>date >CI\_Date >date | Дата |
| gmd:identificationInfo >MD\_DataIdentification >citation >CI\_Citation>date >CI\_Date >date >DateTime | Дата |

*“Табл. 3. Данные о документе”*

|  |  |
| --- | --- |
| gmd:pointOfContact | Субъект, ответственный за создание набора данных |
| gmd:pointOfContact >CI\_ResponsibleParty | Информация об ответственном субъекте |
| gmd:pointOfContact >CI\_ResponsibleParty >individualName | Имя субъекта |
| gmd:pointOfContact >CI\_ResponsibleParty >organisationName | Наименование юридического лица |
| gmd:pointOfContact >CI\_ResponsibleParty >contactInfo | Контактная информация ответственного субъекта |
| gmd:contact>contactInfo> CI\_Contact | Контактная информация для связи с ответственным субъектом |
| gmd:contact>contactInfo> CI\_Contact >phone | Номера контактных телефонов |
| gmd:contact>contactInfo> CI\_Contact >phone >CI\_Telephone | Номера контактных телефонов для связи с ответственным субъектом |
| gmd:contact>contactInfo> CI\_Contact >phone >CI\_Telephone >voice | Номер контактного телефона для связи с ответственным субъектом |
| gmd:contact>contactInfo> CI\_Contact >phone >CI\_Telephone >facsimile | Номер контактного факса для связи с ответственным субъектом |
| gmd:contact>contactInfo> CI\_Contact >address | Информация о местоположении субъекта |
| gmd:contact>contactInfo> CI\_Contact >address >CI\_Address | Адрес |
| gmd:contact>contactInfo> CI\_Contact >address >CI\_Address >deliveryPoint | Адрес |
| gmd:contact>contactInfo> CI\_Contact >address >city | Город |
| gmd:contact>contactInfo> CI\_Contact >address >administrativeArea | Область, край |
| gmd:contact>contactInfo> CI\_Contact >address >postalCode | Почтовый или другой индекс |
| gmd:contact>contactInfo> CI\_Contact >address >country | Страна |
| gmd:contact>contactInfo> CI\_Contact >address >electronicMailAddress | Адрес электронной почты |

*“Табл. 4. Данные об авторах”*

|  |  |
| --- | --- |
| gmd:extent | Состояние местности |
| gmd:extent >EX\_Extent | Пространственно-временные характеристики |
| gmd:extent >EX\_Extent >description | Описание местности |
| gmd:extent >EX\_Extent >geographicElement | Географический компонент протяженности набора данных |
| gmd:extent >EX\_Extent >geographicElement >EX\_GeographicBoundingBox | Географические координаты данных |
| gmd:extent >EX\_Extent >geographicElement >EX\_GeographicBoundingBox >westBoundLongitude | Долгота крайней западной точки |
| gmd:extent >EX\_Extent >geographicElement >EX\_GeographicBoundingBox >eastBoundLongitude | Долгота крайней восточной точки |
| gmd:extent >EX\_Extent >geographicElement >EX\_GeographicBoundingBox >southBoundLatitude | Широта крайней южной точки |
| gmd:extent >EX\_Extent >geographicElement >EX\_GeographicBoundingBox >northBoundLatitude | Широта крайней северной точки |

*“Табл. 5. Данные о пространственно-временных характеристиках”*

|  |  |
| --- | --- |
| gmd:descriptiveKeywords | Набор ключевых слов, характеризующих набор данных и их источник |
| gmd:descriptiveKeywords >MD\_Keywords | Информация о ключевых словах |
| gmd:descriptiveKeywords >MD\_Keywords >keyword | Часто употребляемые, общепринятые слова и фразы |

*“Табл. 6. Данные о ключевых словах”*

|  |  |
| --- | --- |
| gmd:citation | Ссылка на описание данных |
| gmd:citation >CI\_Citation | Ссылка на описание |
| gmd:citation >CI\_Citation >title | Заголовок |
| gmd:citation >CI\_Citation >date | Дата |
| gmd:identificationInfo >MD\_DataIdentification >citation >CI\_Citation>date >CI\_Date | Дата |
| gmd:citation >CI\_Citation >date >CI\_Date >date | Дата |
| gmd:citation >CI\_Citation >date >CI\_Date >date >DateTime | Дата |

*“Табл. 7. Данные о цитировании”*

FGDC (Federal Geographic Data Committee) - стандарт для описания цифровых геопространственных метаданных, получивший широкое распространение и применяющийся в США и во всем мире уже много лет.

Описывается с помощью языка XML.

|  |  |
| --- | --- |
| metadata | Метаданные |
| metadata>idinfo | Информация о методанных |
| metadata>idinfo >descript | Характеристика набора данных, включая его предполагаемое использование и ограничения |
| metadata>idinfo > descript >abstract | Краткое описательное данных |
| metadata>idinfo > descript >purpose | Краткое изложение целей, с которыми был разработан набор данных |
| metadata>idinfo > descript >supplinf | Другая описательная информация о наборе данных |
| metadata>idinfo > timeperd | Период времени для которого соответствует набор данных |
| metadata>idinfo > timeperd >timeinfo | Информация о дате и времени события |
| metadata>idinfo > timeperd >timeinfo >rngdates | Временной диапозон |
| metadata>idinfo > timeperd >timeinfo >rngdates >begdate | Начало события |
| metadata>idinfo > timeperd >timeinfo >rngdates >enddate | Конец события |
| metadata>idinfo > timeperd >current | Дата публикации данных |
| metadata>idinfo > status | Информация о состоянии набора данных. |
| metadata>idinfo > status >progress | Состояние набора данных ("Complete","In work","Planned") |
| metadata>idinfo > status >update | Частота внесения изменений в данные ("Continually","Daily",…) |

*“Табл. 11. Основные данные”*

|  |  |
| --- | --- |
| citation | Информацию, которая будет использоваться для ссылки на набор данных. |
| onlink>citeinfo | Ссылка для набора данных |
| onlink>citeinfo>origin | Название организации или отдельного лица, которые разработали набор данных |
| onlink>citeinfo>pubdate | Дата публикации |
| onlink>citeinfo>title | Заголовок |
| onlink>citeinfo>geoform | Режим, в котором представлены геопространственные данные |
| onlink>citeinfo>pubinfo | Сведения о публикации |
| onlink>citeinfo>pubinfo>pubplace | Название городе, где были опубликованы данные |
| onlink>citeinfo>pubinfo>publish | Имя человека или организации, опубликовавшей набор данных |
| onlink>citeinfo>onlink | Ссылка на данные |

*“Табл. 8. Данные о цитировании”*

|  |  |
| --- | --- |
| ptcontac | Контактная информация об авторах |
| ptcontac >cntinfo | Контактная информация об авторе |
| ptcontac >cntinfo >cntorgp | Организация и член организации, связанные с набором данных |
| ptcontac >cntinfo >cntorgp >cntorg | Название организации |
| ptcontac >cntinfo >cntorgp >cntper | Имя человека |
| ptcontac >cntinfo >cntaddr | Адрес для организации или человека |
| ptcontac >cntinfo >cntaddr >addrtype | Информация, предоставленная об адресе |
| ptcontac >cntinfo >cntaddr >address | Адрес |
| ptcontac >cntinfo >cntaddr >city | Город |
| ptcontac >cntinfo >cntaddr >state | Штат  или провинция |
| ptcontac >cntinfo >cntaddr >postal | Почтовый код |
| ptcontac >cntinfo >cntaddr >country | Страна |

*“Табл. 9. Данные о авторах”*

|  |  |
| --- | --- |
| spdom | Географический ареал набора данных |
| spdom > bounding | Облость |
| spdom > bounding >westbc | Долгота крайней западной точки |
| spdom > bounding >eastbc | Долгота крайней восточной точки |
| spdom > bounding >northbc | Широта крайней северной точки |
| spdom > bounding >southbc | Широта крайней южной точки |

*“Табл. 10. Данные о пространственно-временных характеристиках”*

|  |  |
| --- | --- |
| Keywords | Слова или фразы описывающие данные |
| keywords >theme | Предметы, охватываемые набором данных |
| keywords >theme >themekt | Ссылка на официально зарегистрированный тезаурус или аналогичный авторитетный источник ключевых слов |
| keywords >theme >themekey | Ключевое слово |

*“Табл.11. Данные о ключевых словах”*

# **Функциональные требования**

Разработанная система должна предоставлять пользователю:

* Извлекать метаданные из pdf документа
  + Извлекать заголовок
  + Извлекать авторов
  + Извлекать локации с координатной привязкой
  + Извлекать ключевые слова и словосочетания
  + Извлекать библиографические ссылки
* Давать возможность редактировать компоненты данных
* Давать возможность удалять компоненты данных
* Давать возможность добавлять компоненты данных
* Сохранять данные в форматах ISO19115 и FGDC
* Публиковать ранее сохраненнные данные в форматах ISO19115 и FGDC в систему GeoNetwork
* Публиковать данные в систему GeoNetwork

# **Требования к интерфейсу**

Система предоставляет интерфейс командной строки и графический интерфейс пользователя.

## **Интерфейс командной строки**

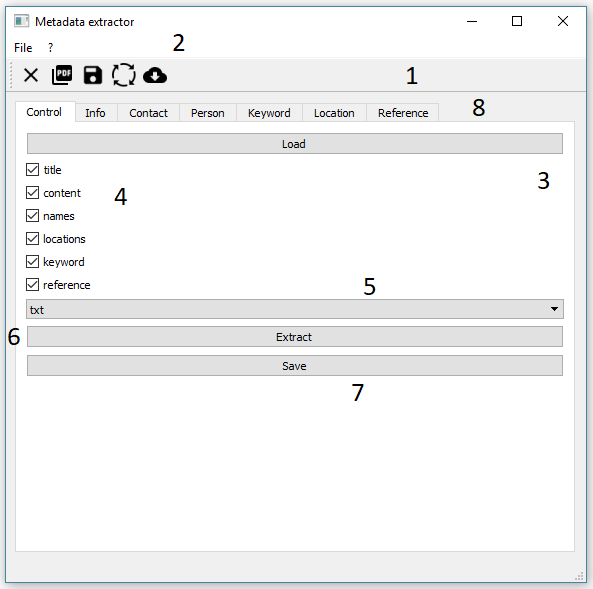
|  |  |
| --- | --- |
| --type -t | Тип выходного файла *txt*, *iso19115* или *fgdc* |
| --input\_file -in | Название входного файла |
| --output\_file -out | Название выходного файла |
| --metaTitle | Извлечение заголовка |
| --metaContent | Извлечение оглавления |
| --metaName | Извлечение авторов |
| --metaLocation | Извлечение локаций |
| --metaKeyWord | Извлечение ключевых слов и словосочетаний |
| --metaRef | Извлечение библиографических ссылок |
| --metaAll | Извлечение всех метаданных |

*“Табл.12. Данные о ключевых словах”*

## **Графический интерфейс пользователя**

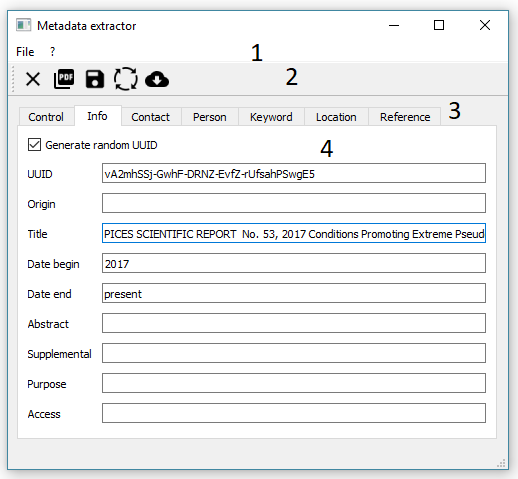
Для реализации графического интерфейса использовалась библиотека PyQT5.

Для более комфортной работы отдельные части метаданных разбиты на вкладки.



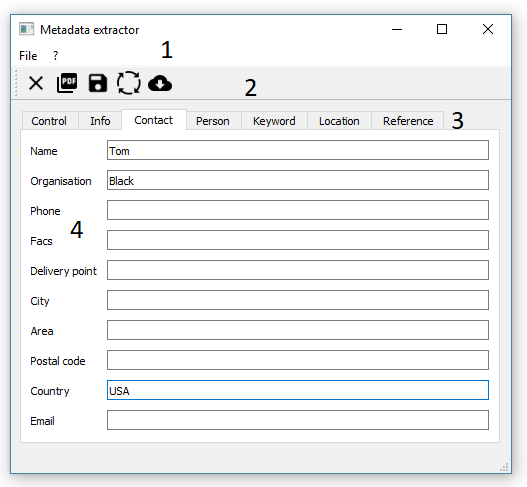
*“Рис. 1. Вкладка управления извлечением метаинформации”*

1. Toolbar
2. Меню
3. Кнопка загрузки PDF файла
4. Чекбоксы для выбора, какие метаданные необходимо извлекать
5. Выбор формата для сохранания метаданных txt / iso19115v2 / FGDC
6. Кнопка для извлечения метаданных
7. Кнопка сохранения метаданных в указанном формате из пункта 5
8. Вкладки с метаинформацией



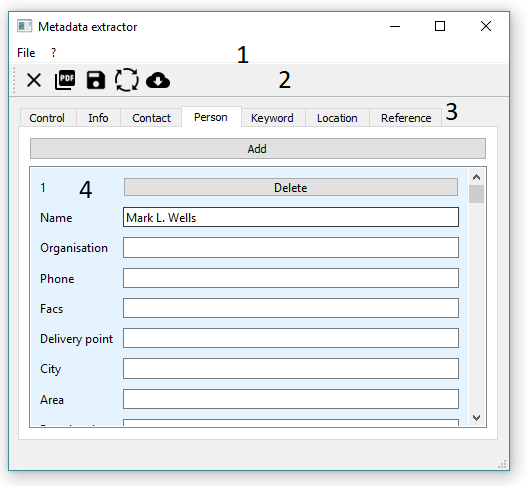
*“Рис. 2. Вкладка управления основной метаинформацией”*

1. Меню
2. Toolbar
3. Вкладки
4. Вкладка с основной информацией



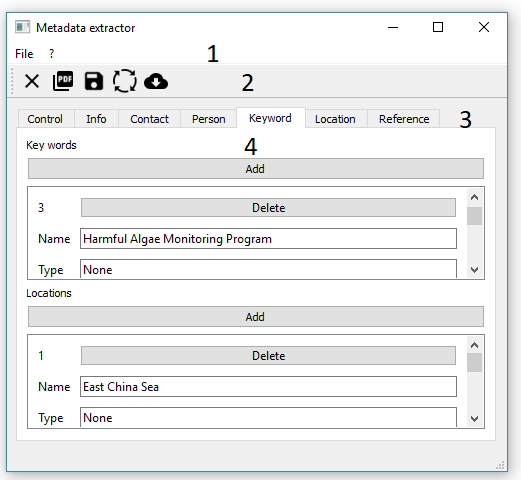
*“Рис. 3. Вкладка управления информацией о субъекте, ответственным за данные”*

1. Меню
2. Toolbar
3. Вкладки
4. Вкладка с информацией о контактном лице, ответственном за данные



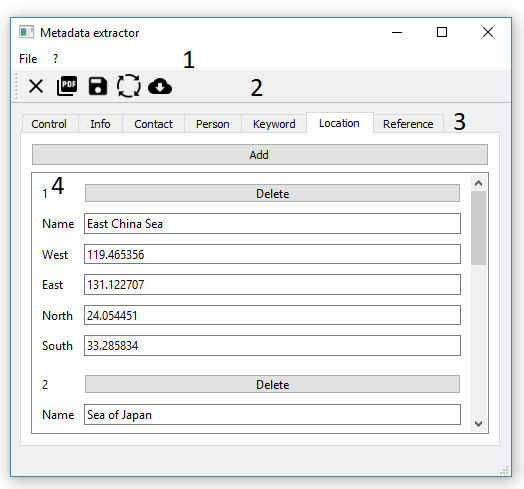
*“Рис. 4. Вкладка управления метаинформацией об авторах данных”*

1. Меню
2. Toolbar
3. Вкладки
4. Вкладка с информацией об авторах данных



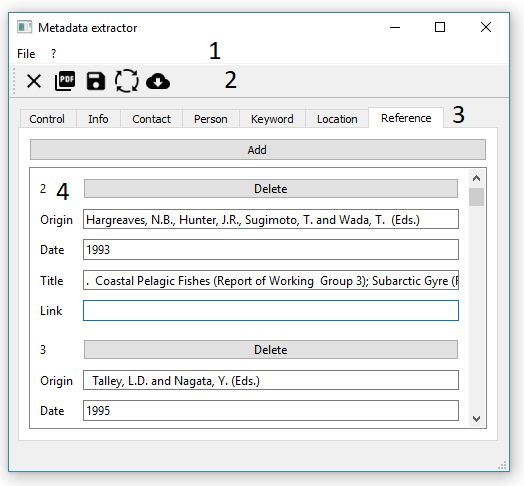
*“Рис. 5. Вкладка управления ключевыми словами и словосочетаниями”*

1. Меню
2. Toolbar
3. Вкладки
4. Вкладка с ключевыми словами и словосочетаниями



*“Рис. 6. Вкладка управления географическими объектами”*

1. Меню
2. Toolbar
3. Вкладки
4. Вкладка с информацией о координатах географических объектах



*“Рис. 7. Вкладка управления библиографическими ссылками”*

1. Меню
2. Toolbar
3. Вкладки
4. Вкладка с библиографическими ссылками

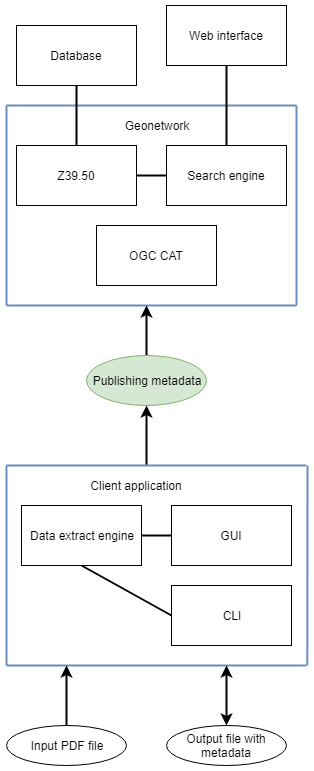
# **Проект**

В качестве языка программирования был выбран Python 3.5, так как он прост в использовании и у него много библиотек на все случаи жизни.

В качестве библиотеки машинного обучения была выбрана Tensorflow, так как у нее большое сообщество, а также отличная документация.

В качестве библиотеки для извлечения текста из pdf документа была выбрана pdfminer. Она хоть и медленнее альтернативных проектов, но зато строит дерево разбора, что помогает отчистить документ от лишних элементов, а это в свою очередь облегчает процесс извлечения метаданных.

Архитектура системы:



*“Рис. 8. Архитектура системы”*

Z39.50 – протокол для поиска и получения информации и удаленных баз данных.

OGC - протокол для публикации и доступа к цифровым каталогам геопространственных метаданных.

Для реализации системы требуется реализовать методы для извлечения информации, а также методы для представления ее в стандартизированном виде.

## **Описание метода**

### Извлечение именованных сущностей

Итак, мы имеем неструктурированный текст, и наша задача получить теги для слов из этого текста (I-PER, O, B-LOC и т. д.).

Пример:

|  |  |
| --- | --- |
| Токен | Тег |
| Inna | I-PER |
| V. | I-PER |
| Stonik | I-PER |
| lives | O |
| in | O |
| Vladivostok | B-LOC |

*“Табл. 13. Пример теггирования”*

Для обучения модели был взят размеченный корпус, в котором имеются следующие виды именованных сущностей: места, имена людей, организации, разное.

Данные для обучения представлены в следующем виде:

Phil NNP I-NP I-PER

Simmons NNP I-NP I-PER

took VBD I-VP O

four CD I-NP O

for IN I-PP O

38 CD I-NP O

on IN I-PP O

Friday NNP I-NP O

as IN I-PP O

Leicestershire NNP I-NP I-ORG

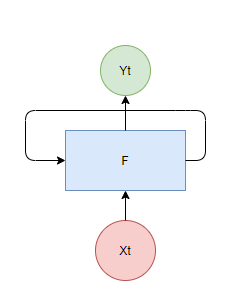
beat VBD I-VP O

Somerset NNP I-NP I-ORG

…

Проблема задачи определения тега слова в том, что нельзя рассматривать каждое слово в отдельности. Например, Russian Federation – имеет тег ‘место’ и состоит из двух слов. Следовательно, нужно, чтобы модель имела память, и при встрече слова ‘Federation’ понимала, что ‘Russian’ тоже является местом. С этой задачей справляются рекуррентные нейронные сети.

Рекуррентные нейронные сети - сети, содержащие обратные связи и позволяющие сохранять информацию. Наличие этой связи дает возможность передавать информацию от одного шага сети другому.



*“Рис. 9. Модель устройства рекуррентной сети”*

Здесь сеть принимает и возвращает .

Конкретно в реализованной модели используется модификация рекуррентной нейронной сети – LSTM.

Долгая краткосрочная память или Long short-term memory (LSTM) – разновидность рекуррентных нейронных сетей, способная к обучению долговременным зависимостям.

Её особенность в том, что что она может хранить состояние. В ходе работы LSTM её состояние может меняется. Это позволяет принимать решение о том, что нужно запомнить, а что можно забыть.

Для обучения нейронной сети, представим слова из обучающего корпуса в виде векторов . Для этого используется алгоритм GloVe.

GloVe - это алгоритм машинного обучения для получения векторных представлений для слов. Обучение выполняется по совокупной статистике совпадения слов из корпуса, а полученные представления показывают линейные зависимости векторного пространства слов.

Так же мы будем учитывать и символы из которых состоит слово, так как это тоже даёт важную информацию. (Например, если слово начинается с заглавной буквы, то возможно оно является именем или локацией.) Это будет вектор , он описывает символьные характеристики слова.

Итак:

Каждый символ представлен в виде вектора .

После работы LSTM над , происходит конкатенация:

.

Далее происходит объединение :

– это и есть слово, представленное в векторном виде.

Далее запускаем LSTM на полученном пространстве векторов и получаем новое пространство векторов .

Теперь, каждый вектор слова ассоциируется с вектором .

Чтобы сделать окончательное предсказание, выполним:

, где каждая компонента вектора – это вероятность в пользу того, что слово относится к определённому типу.

Теперь для окончательного предсказания нужно оценить вероятность последовательности слов.

Для этого используется алгоритм условно случайных полей (CRF).

Нужно оценить

– матрица перехода, – вектора оценок, которые фиксируют оценки первого или последнего заданного тега.

Матрица показывает линейную зависимость между соседними словами.

Таким образом происходит суммирование весов каждого слова и весов комбинации двух соседних слов.

Вектора тегов между оцениваются так, чтобы была максимальна.

Так как сложность растет экспоненциально в зависимости от длины последовательности для решения этой задачи используется метод динамического программирования.

Пусть уже известно решение для для шагов от до , для последовательности начинающейся с для всех классов тегов.

Тогда решение для последовательности, начинающейся с будет иметь вид:

В итоге общая вероятность для последовательности тэгов определяется с помощью softmax так:

, где – нормирующий коэффициент:

- сумма оценок всевозможных последовательностей.

Для вычисления так же используется метод динамического программирования.

Пусть - сумма оценок всевозможных последовательностей, начинающихся с тэга на шаге t.

Тогда:

Для оценки потерь используется кросс-энтропийная функция

### Оценка качества модели

При обучении алгоритма корпус разбивается на обучающую и тестовую выборку.

С помощью тестовой выборки можно оценить качество модели.

Оценка осуществляется с помощью следующих численных характеристик.

1. Accurancy:

Полнота присваивает всем классам одинаковый вес, что может быть не корректно в случае если распределение в обучающей выборке смещено в сторону каких-то классов.

Поэтому нельзя судить о качестве модели только по этой характеристике.

1. Точность:

Точность – это доля элементов, действительно принадлежащих данному классу относительно всех элементов, которые должны быть отнесены к этому классу.

1. Полнота:

Полнота – это доля найденных элементов, принадлежащих классу относительно всех документов этого класса в тестовой выборке.

 — истино-положительное решение;

 — истино-отрицательное решение;

 — ложно-положительное решение;

— ложно-отрицательное решение.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Истинные знач |  |
|  |  | Положительные | Отрицательные |
| Оценка сист. | Положительные | TP | FP |
|  | Отрицательные | FN | TN |

*“Табл. 14. Оценка качества модели”*

1. F-мера

Чем выше точность и полнота, тем лучше. Но в реальной жизни максимальная точность и полнота не достижимы одновременно и приходится искать некий баланс. Поэтому, хотелось бы иметь метрику, которая объединяла бы в себе информацию о точности и полноте нашего алгоритма. Именно такой метрикой является F-мера.

F-мера представляет собой [гармоническое среднее](http://bazhenov.me/blog/2012/05/05/harmonic-mean.html) между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю.

### Результаты

Реализованный алгоритм выделяет следующие сущности: географические наименования (LOC); названия организаций (ORG); имена люжей (PER), разное (MISC).

Получены следующие результаты:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Acc | Pred | Recall | FM |
| LSTM | 98.89 | 93.5 | 93.6 | 93.57 |
| GROBID | 98.97 | 90.72 | 90.18 | 90.45 |
| CERMINE | - | 92.49 | 93.83 | 93.16 |
| ParsCit | 95.7 | 95.7 | 95.7 | 95.7 |

*“Табл. 15. Результаты модели”*

Результаты по отдельным классам:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Acc | Pred | Recall | FM |
| PER | 99.2 | 94.8 | 93.5 | 94.15 |
| LOC | 98.4 | 93.5 | 93.4 | 93.45 |
| ORG | 98.2 | 92.51 | 92.3 | 92.4 |
| MISC | 92.34 | 88.05 | 89.5 | 88.77 |
| REF | 95.4 | 98.34 | 98.01 | 98.17 |

*“Табл. 16. Результаты модели по отдельным сущностям”*

### Извлечение библиографических ссылок

Так как библиографическая ссылка имеет стандартизированный формат решено извлекать их с помощью регулярных выражений.

В предоставленных отчётах используется гарвардский тип библиографических ссылок.

Каждую библиографическую ссылку можно разбить на составляющие:

список имен, заголовок, издание, год, количество страниц, серия.

Таким образом, можно составить регулярное выражение для каждой части, и их конкатенация будет давать полное регулярное выражение.

Полученные выражения:

Извлечение автора:

name = "(((de)|(De)|(da)|(van)|(zu)|(di)|(der)|(tot)|(thoe))(\\s\*)){0,1}([A-ZС]{1})([A-Za-z\\-С])+,(\\s\*)([A-Za-z\\.С]+)"

Пример: Miller, C.B.

Извлечение авторов:

names = "(" + self.regexName + "([\\s\*)[and,]\*(\\s\*)){1,}](file:///\\s*)[and,]*(\\s*))%7b1,%7d)"

Пример: Aydin, K.Y., McFarlane, G.A., King, J.R. and Megrey, B.A.

Извлечение eds. – другие авторы:

eds = "(\\(Ed\\.\\))|(\\(Eds\\.\\))"

Извлечение авторов вместе с eds:

names+eds = names + "(\\s\*)" + "("+eds+")\*"

Пример: Aydin, K.Y., McFarlane, G.A., King, J.R. and Megrey, B.A. (Eds.)

Извлечение года:

year = "[0-9]{4}\\.";

Пример:

2010.

Извлечение авторов вместе с eds и годом:

names+eds+year = names + "(\\s\*)" + "("+eds+")\*" + "(\\s\*)" + "("+year+"){1}"

Извлечение названия и количества страниц:

title+pages = "[A-Za-z0-9/\\.,\\s();\\-:–]{1,300}(([0-9]+[\\-:][0-9]\*\\.?)|([0-9]+[0-9\\-–\\n]\*(\\s\*)((pp)|(p))\\.?)|((\\s\*)[0-9]+[\\-,–\\n]+[0-9]+\\.?)|((\\s\*)((pp)|(p))\\.)(\\s\*)[0-9]+[0-9\\-–\\n]\*)"

Полное регулярное выражение для извлечения библиографических ссылок:

reference = names+eds+year + "(\\s\*)" + title+pages

Пример:

King, J.R. (Ed.) 2005. Report of the Study Group on the

Fisheries and Ecosystem Responses to Recent Regime

Shifts. PICES Sci. Rep. No. 28, 162 pp.

### Извлечение локаций с координатной привязкой

Для извлечения локаций используется метод с рекуррентной нейронной сети, описанный ранее.

Далее эти локации отсеиваются с помощью словаря с названиями объектов и их координатами. Для сравнения кандидатов со словарём используется алгоритм нечёткого сравнения строк.

В алгоритме используется коэффициент Танимото, который измеряет степень схожести двух множеств. Он используется для того, чтобы оценить подобие образцов, представленных списками свойств.

Коэффициент Танимото:

*a* – количество элементов в первом слове, *b* – количество элементов во втором слове, *с* – совпадающие элементы первого и второго слова.

В данной реализации в качестве элементов слов используется N-граммы.

Такой подход помогает идентифицировать слова в разных формах и слова содержащие ошибки.

### Извлечение ключевых слов и словосочетаний

Для извлечения ключевых слов и словосочетаний используется алгоритм RAKE (Rapid Automatic Keyword Extraction).

В нём происходит генерация всевозможных словосочетаний из предоставленных предложений. Подсчитывается частота этих словосочетаний в тексте. Далее рассчитывается граф частот соседних слов из которых состоят словосочетания. Далее для каждого слова рассчитывается его степень.

Степень – это сумма всех частот с которыми соседствует слово .

Чтобы вычислить окончательный приоритет словосочетания нужно для каждого слова вычислить его метрику, которая вычисляется так:

.

Для улучшения качества используется нормализация слов, а также фильтрация с помощью стоп слов и опорных слов.

Стоп слова – словосочетания с этими словами игнорируется.

Опорные слова – предложения с этими словами используются для генерации словосочетаний.

В результате получается список словосочетаний отсортированный по метрики .

**Заключение**

В ходе написания работы было написано более 4000 строк на языке Python (более 400 кб.).

Таким образом, в процессе выпускной квалификационной работы мною были:

* исследованы существующие методы извлечения информации;
* реализованы методы извлечения информации из неструктурированного текста:
* извлечение именованных сущностей с помощью рекуррентных нейронных сетей с долгой краткосрочной памятью, извлечение библиографических ссылок с помощью регулярных выражений, извлечение ключевых слов и словосочетаний с помощью алгоритма Rapid Automatic Keyword Extraction.
* проведено тестирование разработанных методов;
* выполнена реализация графического интерфейса пользователя с возможностью редактирования извлеченных метаописаний и возможностью публикации их в систему GeoNetwork;

**Список литературы**

1. Rose S. et al. Automatic keyword extraction from individual documents //Text Mining: Applications and Theory. – 2010. – С. 1-20.
2. Zhu X. Cs838-1 advanced nlp: Conditional random fields. – Technical report, The University of Wisconsin Madison, 2007.
3. Limsopatham N., Collier N. H. Bidirectional LSTM for named entity recognition in Twitter messages. – 2016.
4. McCallum A., Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons //Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. – Association for Computational Linguistics, 2003. – С. 188-191.
5. Klinger R., Friedrich C. M. Feature subset selection in conditional random fields for named entity recognition //Proceedings of the International Conference RANLP-2009. – 2009. – С. 185-191.
6. VanderPlas J. Python data science handbook: Essential tools for working with data. – " O'Reilly Media, Inc.", 2016.
7. Пескова О. В. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. – 2015.
8. Колмогорцев С. В., Сараев П. В. Извлечение библиографии из текстов регулярными выражениями //Новые информационные технологии в автоматизированных системах. – 2017. – №. 20.
9. Ванюшкин А. С., Гращенко Л. А. Методы и алгоритмы извлечения ключевых слов //Новые информационные технологии в автоматизированных системах. – 2016. – №. 19.
10. Антонова А. Ю., Соловьев А. Н. Метод условных случайных полей в задачах обработки русскоязычных текстов //Информационные технологии и системы//Труды международной научной конференции. – 2013. – С. 1-6.
11. Лабутин, И.А., Фирсов, А.Н., Чуприна, С.И. Распознавание именованных сущностей в текстах на естественном языке с использованием метода пробросо-цепочных условных случайных полей // Пермский государственный национальный исследовательский университет. – 8 с.