МИНОБРНАУКИ РОССИИ

федеральное государственное бюджетное образовательное

учреждение высшего образования

«Омский государственный университет им. Ф.М. Достоевского»

Кафедра программного обеспечения и защиты информации

РЕШЕНИЕ ЗАДАЧ NLP С ИСПОЛЬЗОВАНИЕМ БИБЛИОТЕКИ APACHE OPEN NLP

Курсовой проект   
по направлению Прикладная математика и информатика

Научный руководитель:

\_\_\_\_\_\_\_\_\_\_\_\_ Агафонов А.Л.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_ 2023 г.

Выполнил:

студент группы ММБ-004

\_\_\_\_\_\_\_\_\_\_\_ Васильев Д.О.

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_ 2023 г.

Омск

2023

Содержание

Введение ……………………………………………………………………………. 3

Глава 1. Установка OpenNLP………………………………………………………. 4

Глава 2. Обработка и разметка текста……………...……………………………....6

Глава 3. Извлечение информации из текста………..……………………………..10

Глава 4. Анализ полученный результатов .…………………………………….....13

Заключение ………………………………………………………………………....15

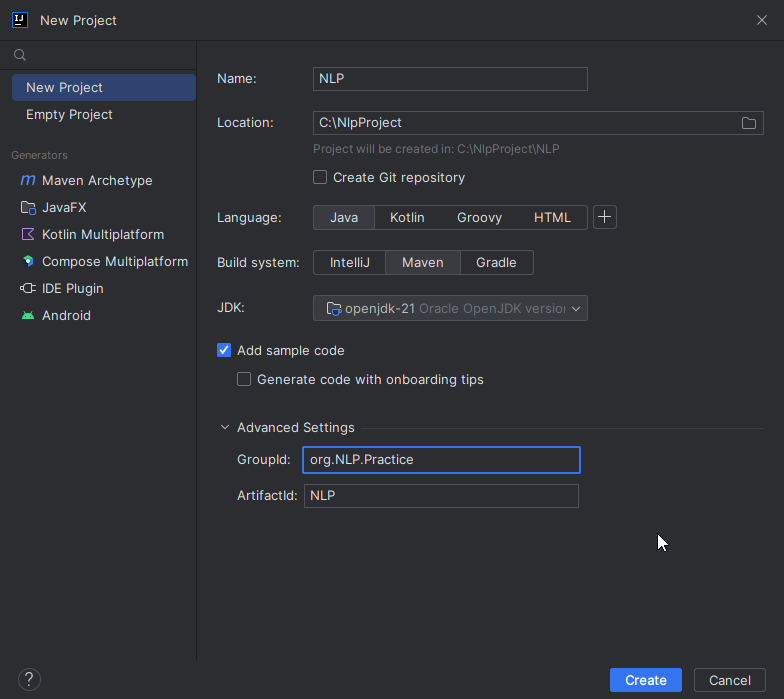
Введение

Apache OpenNLP — это мощная и широко используемая библиотека для обработки естественного языка (Natural Language Processing - NLP), основанная на машинном обучении. Она предоставляет инструменты и модели для выполнения различных задач, связанных с анализом и обработкой текста, включая обнаружение именованных сущностей, распознавание частей речи, анализ синтаксиса и многое другое.

Цель данной курсовой работы состоит в знакомстве и изучении библиотеки Apache OpenNLP. В ходе работы будет проведен обзор некоторых основных возможностей и функциональностей OpenNLP, основополагающих принципов его работы и архитектуры. В работе будет рассмотрена установка и настройка Apache OpenNLP, приведены примеры кода, демонстрирующие базовые операции с текстом. Будет также рассмотрено использование предварительно обученных моделей OpenNLP для выполнения задач анализа текста. Далее будут рассмотрены примеры применения Apache OpenNLP для решения конкретных задач в области NLP.

В заключение будут сделаны выводы о возможностях и перспективах Apache OpenNLP, а также оценена его эффективность и применимость в области анализа естественного языка. Работа будет обладать практической направленностью и представит реальные примеры использования Apache OpenNLP для решения задач NLP.

Глава 1. Установка OpenNLP

На первом шаге нам нужно создать наш java проект, используя возможности сборщика проекта maven:  
Рис. 1. Диалоговое окно Intellij IDEA по созданию проекта

Далее обратимся к файлу pom.xml и добавим главную зависимость к нашему проекту для возможности работы с библиотекой OpenNLP:

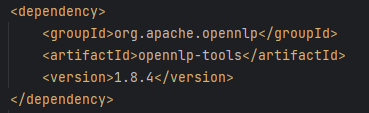


Рис. 2. Зависимость в файле pom.xml для OpenNLP

Для дальнейшего удобства тестирования в файле pom.xml так же пропишем зависимость к библиотекам junit:

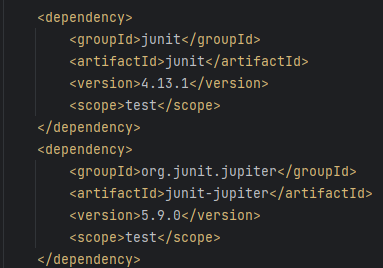


Рис. 3. Зависимость в файле pom.xml для junit

Создадим класс NLPpractice в подкаталоге main и TestNLPpractice в подкаталоге test:

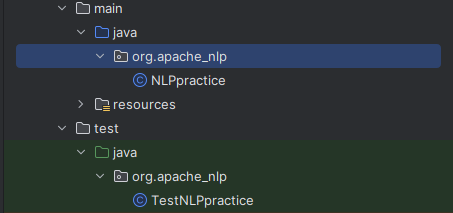


Рис. 4. Дерево проекта

На этом основная настройка проекта завершена, перейдем к библиотеке OpenNLP.

Глава 2. Обработка и разметка текста

Прежде чем мы начнем работать с библиотекой, нужно вспомнить, что OpenNLP – библиотека, основанная на машинном обучении. Библиотека не только предоставляет возможность пользователям самим создавать модели для обработки текста и обучать их по своему усмотрению, но также дает возможность пользоваться уже готовыми, обученными моделями, предоставляя доступ к их загрузке на своем официальном сайте. Загрузим несколько моделей и поместим их в папку resources нашего проекта:

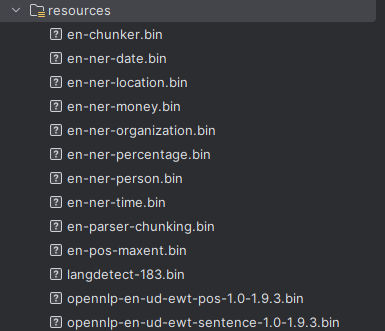


Рис. 5. Модели Apache OpenNLP в проекте

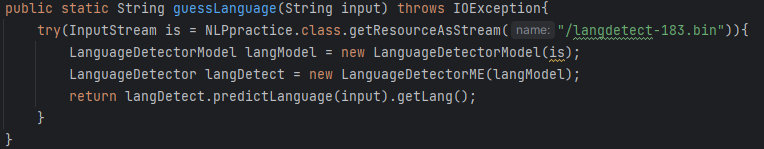
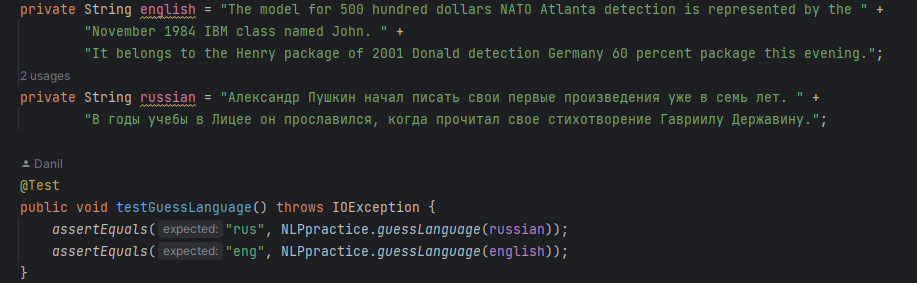
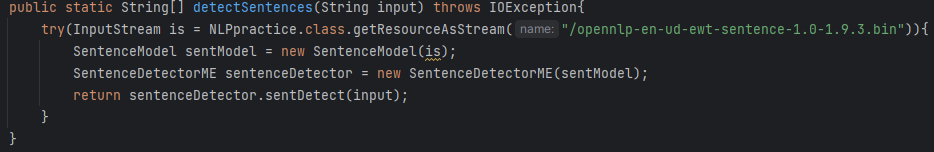
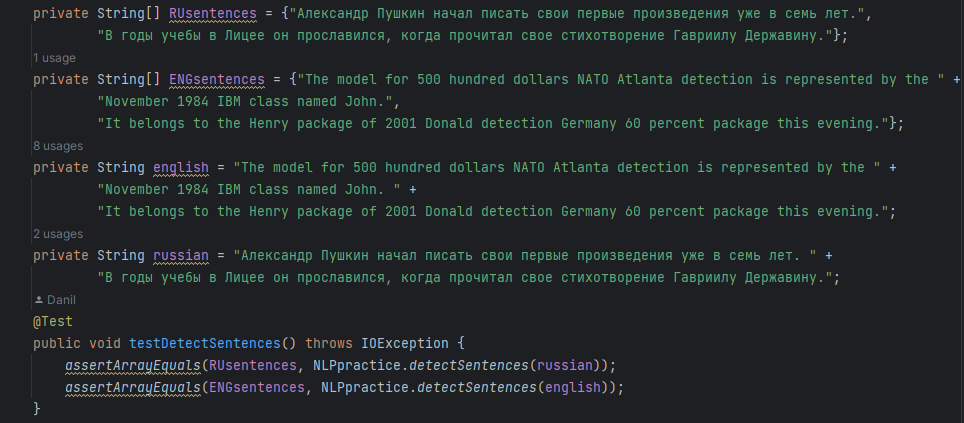
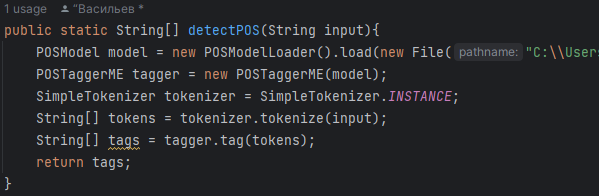
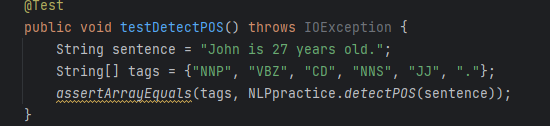
Перейдем к изучению возможностей библиотеки. Для начала напишем простой метод, который будет распознавать язык в тексте и протестируем его:   
  


Рис. 6, 7. Метод распознавания языка и его тест

На вход методу подается два текста, которые он успешно распознает как русский и английский. В этом методе и каждом последующем мы будем использовать ранее скаченные модели Apache из папки resources. Важное замечание, что для каждой операции с текстом предполагается использование своей модели. Теперь напишем метод, который в заранее необработанном тексте будет распознавать предложения. На вход метод будет получать текст, а возвращать массив из предложений:  
  
  
Рис. 8, 9. Метод обнаружения предложения и его тест

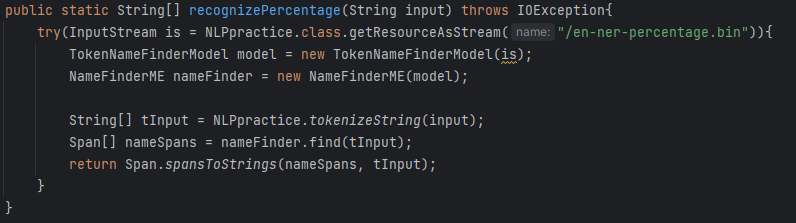
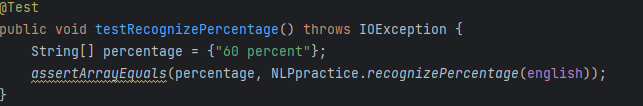
Т.к. методу на вход были даны тексты из двух предложений, то метод возвращает массив из 2 элемента – тех самых предложений. В заключении обработки и разметки текста напишем метод, который будет определять части речи в предложении:

  
  
Рис. 10, 11. Метод определения частей речи в предложении и его тест.

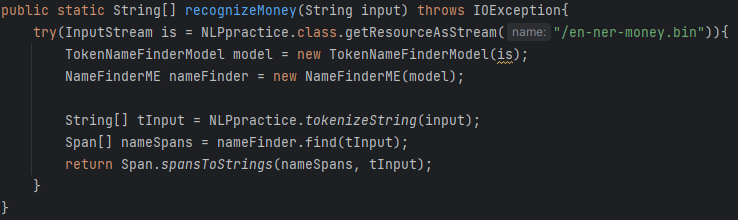
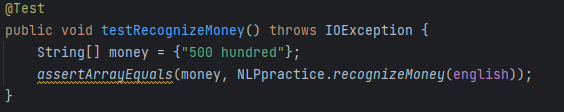
В данном тесте проверяется массив с тегами частей речи на каждом своем месте:

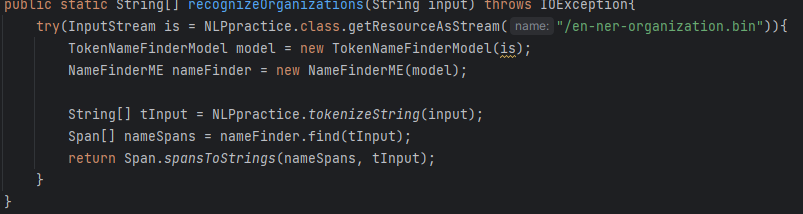
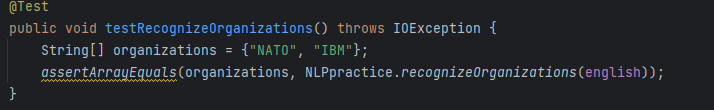
* NNP - proper noun, singular – существительное собственное
* VBZ - Verb, third person singular present – глагол в третьем лице, едн. Числе
* CD – Cardinal number – количественное числительное
* NNS - Noun, plural- существительное в множественном числе
* JJ – Adjective - прилагательное

Глава 3. Извлечение информации из текста

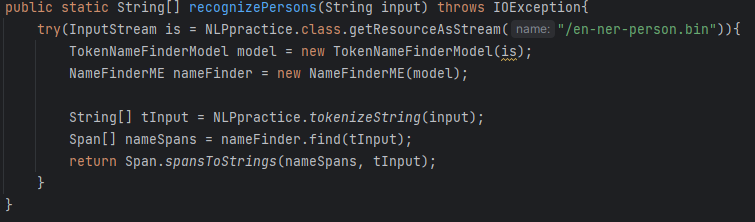
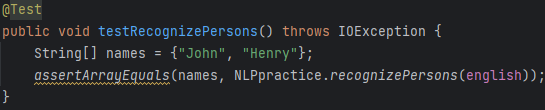
Перейдем к изучению написанию методов, которые помогут извлекать нужную нам информацию из текста. Напишем метод, который будет распознавать процент тексте:   
   
  
Рис. 12, 13. Метод для извлечения процента и его тест

Данному методу поступает на вход на английском языке, который использовался в прошлой главе: The model for 500 hundred dollars NATO Atlanta detection is represented by the November 1984 IBM class named John. It belongs to the Henry package of 2001 Donald detection Germany 60 percent package this evening. В данном тексте фигурирует процент – 60 percent, что успешно определила функция. Теперь напишем метод, который будет определять деньги в тексте:

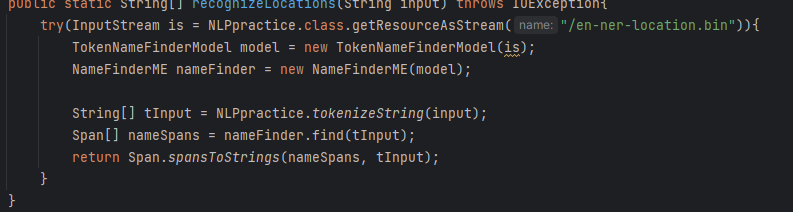
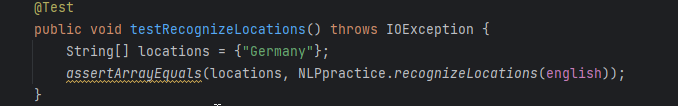
  
  
Рис. 14, 15. Метод для извлечения денег и его тест

И вправду, в тексте присутствует упоминание 500 hundred dollars, но модель не смогла определить валюту, только количество. Посмотрим как правильно готовая модель сможет определить наименования организаций в тексте:   
  
  
Рис. 16, 17. Метод распознавание организаций и его тест

Действительно, в тексте упоминаются такие организации как NATO и IBM, модель успешно справилась со своей работой. Перейдем к методам, с которыми готовые модели справились не совсем успешно. Рассмотрим метод распознавание имен:

  
  
Рис. 18, 19. Метод распознавание имен и его тест.

В тексте упоминается 3 имени: John, Henry, Donald, но модель смогла выявить только 2 из 3 (77,7…%). Это не является провальным результатом, но хотелось бы лучше. Далее рассмотрим метод распознавания мировых локаций:

  
  
Рис. 20, 21. Метод распознаваний мировых локаций и его тест

В тексте упоминается:

* Atlanta – город в США
* Germany – страна

Данный метод смог распознать только страну (50%), не самый желаемый результат работы. Рассмотрим последний метод, который распознает дату:

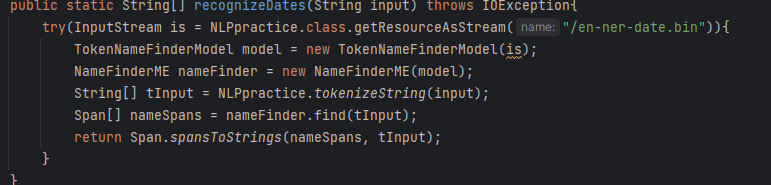
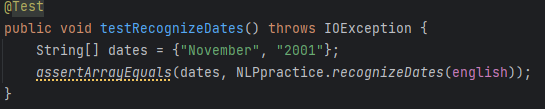
  


Рис. 22, 23. Метод распознавание даты и его тест

В тесте упоминаются 2 даты:

* November 1984
* 2001

Метод успешно распознает только одну последнюю дату, а также вводит в заблуждения, распознавая первую только как November. Эта самый худший результат среди всех методов.

Глава 4. Анализ полученных результатов

В прошлых главах мы написали 10 методов и протестировали их, но не все из них показались желаемый результат. Модель для распознавания денег не смогла в результате определить валюту. Метод для распознавания локаций не определил город Атланту, а метод по распознаванию даты не смог определить месяц и год как одну дату. Это может свидетельствовать о недоученности модели или иных упущениях. В основном методы библиотеки работали корректно и без нареканий, а большая часть моделей, на основе которых производился анализ, успешно распознала желаемый результат. Для более успешного результата можно обучать модель на основе своих желаний и использовать ее, вместо готовой от Apache, т.к. они очень редко обновляются.

Заключение

Apache OpenNLP представляет значительную ценность для разработчиков и исследователей в области обработки естественного языка. Эта библиотека предоставляет широкий спектр инструментов и функциональности, которые позволяют автоматизировать и улучшить процессы анализа текста, извлечения информации, классификации и определения частей речи. В ходе выполнения данной работы мы изучили основные принципы работы с Apache OpenNLP, включая установку, конфигурацию и использование различных модулей библиотеки. Также мы рассмотрели примеры использования Apache OpenNLP для решения практических задач. Однако, необходимо отметить, что при работе с Apache OpenNLP возможны некоторые вызовы и сложности, такие как подбор и обучение подходящих моделей. Эти аспекты требуют дополнительного исследования и опыта для достижения оптимальных результатов. В целом, изучение и использование Apache OpenNLP позволяет производить эффективный анализ текста на естественном языке с использованием современных методов и подходов. Эта библиотека является мощным инструментом в руках разработчика, который может значительно упростить и улучшить процессы обработки текстовых данных.