НАУЧНАЯ СТАТЬЯ
УДК 00.00
DOI 10.46698/VNC.2023.79.12.001

К.Д. Плиев

Обзор инструментов и методов парсинга и анализа полных имен

Константин Давидович Плиев
Российский университет дружбы народов, факультет физико-математических наук, магистрант, Москва, Россия, plievk@vk.com

Аннотация. *Основой настоящей статьи является обзор и анализ качества работы инструментов парсинга и анализа полных человеческих имен, обзор методов и технологий, применяемых в них. Для написания статьи и анализа качества работы использовался датасет, собранный с помощью открытых ресурсов и охватывающий различные культуры, языки и в целом случаи, имеющие нетривиальный характер. В результате были обобщены и систематизированы знания о существующих решениях и подходах; выработаны рекомендации заинтересованным лицам, планирующим работать в данном направлении.*
Ключевые слова: *NER, полное имя, анализ имени, интеллектуальный анализ текстов, парсинг имен*
Для цитирования: *Плиев К.Д. Обзор инструментов и методов парсинга и анализа полных человеческих имен // Вестник Владикавказского научного центра. 2023. Т. 23. № 2. С. 71–74. DOI 10.46698/VNC.2023.79.12.001*

Overview of tools and methods for parsing and analyzing full names

Konstantin D. Pliev
RUDN University, Faculty of Science, master's student, plievk@vk.com.

Abstract. *The basis of this article is a review and analysis of the quality of the parsing and analysis tools of full human names, an overview of the methods and technologies used in them. It was proposed to use dataset for effeciencyanalyzing of tools considered in the article. As a result, knowledge about existing solutions and approaches was generalized and systematized; recommendations were made to interested persons planning to work in this direction*
Keywords: *NER, full name, name analysis, text analysis, name parsers.*
For citation: *Pliev K.D. Overview of tools and methods for parsing and analyzing full names // Bulletin of the Vladikavkaz Scientific Center. 2023. Vol. 23. No 2. P. 71–74. DOI 10.46698/VNC.2023.79.12.001*

ВВЕДЕНИЕ

Данная работа посвящена обзору и тестированию инструментов и методов парсинга и анализа полных человеческих имен. Под парсингом в статье подразумевается обнаружение полного имени в неструктурированном тексте и его разбиение на субкомпоненты (фамилия, имя, отчество и т. д.) Для оценки качества работы был собран датасет из открытых источников, в который вошли случаи, охватывающие различные культуры, языки и другие данные, релевантные для каждой личности (страна, пол, возраст).

Целью интеллектуального анализа полных человеческих имен является получение информации о личности (гражданство, национальность, пол, примерный возраст, деление полного имени человека на субкомпоненты: имя, фамилия, отчество и т. п.) при наличии только полного имени человека.

Как известно, данные, подходящие для обучения моделей, пригодные для решения целого спектра задач, которые рассматриваются в данной статье (парсинг и анализ полных человеческих имен), уже бывают структурированы в силу самого характера цели их первоначального сбора. К примеру: перепись населения, записи в медицинских учреждениях и т. д.

Таким образом, можно сказать, что рассматриваемая тематика несколько отличается от традиционных задач по обработке массивов текста, т. к. чаще всего обрабатываются неструктурированные массивы текстов, выявляются закономерности между разными словами-сущностями и т. п.

Что же касается существующих методов, средств или продуктов, то такие черты, как универсальность (работа с многообразием культур, языков) и узконаправленность (неполный анализ по полному имени), стали неотъемлемыми свойствами всех доступных на сегодняшний день решений. К примеру, если взять продукт от компании Name API [1] (как один из наиболее универсально применяемых для различных культур), то можно явно увидеть, что данный коммерческий продукт поддерживает не все разнообразие языков и культур. В нем имеется поддержка 70 стран (хотя «страна» и «культура» понятия не тождественные). Что же касается менее локализованных продуктов и более узконаправленных, то можно констатировать, что все эти продукты требуют серьезной доработки.

С проблемой парсинга и анализа полных человеческих имен так или иначе сталкиваются многие организации, да и отдельные личности тоже. К примеру: различным организациям, агрегирующим научные документы, необходимо разбить список авторов на отдельных авторов, правильно поделить на субкомпоненты полное имя каждого автора. Это необходимо для правильного присвоения уникальных идентификаторов авторам документов, чтобы, к примеру, облегчить поиск по авторам внутри системы. На первый взгляд может показаться, что задача тривиальная, ведь, когда человек читает, то, с точки зрения науки ономастики, он совершает анализ, сравнивая увиденные на бумаге слова (лексемы) с теми, что есть в его словарном запасе, и применяя некие правила, которые различаются в зависимости от культуры и языка, с которыми знаком человек. Парсер – программа, позволяющая компьютеру проделать аналогичную работу.

На сегодняшний день главная проблема парсеров полных имен заключается в неполноценной поддержке культур и языков малых народов в связи с недостатком данных, необходимых для написания правил разбиения на субкомпоненты полного имени или же оформления в виде обучающего датасета для моделей машинного обучения.

Таким образом, данная тема актуальна, так как на текущий момент имеется некоторое разнообразие продуктов, технологий для решения вышеописанной проблемы, и хотелось бы иметь некоторое сравнение их работы и анализа функциональности, для правильного принятия решения по использованию тех или иных средств. Также важно выявить общие недостатки (или же изъяны, присущие конкретным средствам) для формирования рекомендаций по дальнейшему развитию технологий, применяемых в анализе и парсинге полных имен.

Целями настоящей работы являются: 1) разработка рекомендаций будущим заинтересованным лицам в разработке средств, технологий, продуктов, которые бы могли быть использованы в парсинге и анализе полных человеческих имен; 2) сравнение уже существующих средств для облегчения заинтересованным лицам принятия решения по использованию того или иного инструмента в своей работе.

Основными задачами данной работы являются следующие моменты:

1. Собрать датасет для тестирования существующих решений.

2. Собрать список и кратко охарактеризовать существующие технологии-средства парсинга и анализа полных имен.

3. Разработать рекомендации для будущих исследований.

ПРИНЦИПЫ И МЕТОДЫ СБОРА ДАТАСЕТА

Очевидно, что для сбора датасета, который можно применять для объективного оценивания работы существующих и будущих решений по парсингу и анализу полных имен, необходимо охватить как можно большее количество стран, языков и культур. Но мало лишь гнаться за простым числом охваченных стран или же языков, необходимо выявить в каждой культуре свои собственные «отклонения от нормы», иначе говоря, случаи, имеющие нетривиальный характер, даже для культуры-языка, которая их породила (подробнее об этом будет сказано чуть позже). Ведь именно характер решения таких нетривиальных задач средствами для оных проблем будет являться свидетельством, по которому можно будет судить о глубине и широте охвата тех или иных культур, языков или же стран, поддержка которых заявлена авторами-разработчиками средств, применяемых для парсинга и анализа полных имен.

Так как под анализом и парсингом полных имен в данной работе подразумевается: определение пола, гражданства, этнической принадлежности, примерного возраста, определение нерелевантных токенов в составе полного имени человека (например: Mr., Ms и т. п.), разбиение строки с несколькими именами на отдельные, разделение на субкомпоненты – то, конечно же, датасет для оценки качества работы тех или иных инструментов должен включать в себя все эти данные по каждому отдельному человеку. Очевидно, что масштабный сбор такой персональной информации без согласия носителей полных имен (и другой сопутствующей и необходимой нам информации) является неэтичным. Поэтому было принято решение использовать *wikidata* [2]. *Wikidata* – это совместно редактируемая база знаний, созданная Фондом Викимедиа. Схема сбора данных выглядит следующим образом:

1) Определение количества культур (национальностей или же языков), доступных через API.

2) Сбор данных об известных личностях – представителях каждой доступной по API культуре, языку или же национальности.

Также полезной будет информация, содержащая транслитерацию полного имени в английский и русский языки.

Получить список всех языковых версий Википедии, доступных через API (например, с помощью запроса *https://meta.wikimedia.org/w/api.php?action=sitematrix&format=json*).

Для получения необходимых данных можно использовать следующий запрос, где в поле P27 необходимо менять свойство таким образом, чтобы перебирать все страны (в том числе и не входящие в ООН для более полного охвата).

```
SELECT DISTINCT ?personLabel_en ?personLabel_ru ?genderLabel_en ?birthYear ?nativeName ?ethnic_groupLabel ?firstNameLabel ?lastNameLabel ?patronymNameLabel ?titulNameLabel WHERE {
  ?person wdt:P31 wd:Q5 ;
```

```
  wdt:P27 wd:Q30 .
    OPTIONAL {
      ?person rdfs:label ?personLabel_en.
      FILTER (lang(?personLabel_en) = 'en')
    }
    OPTIONAL {
      ?person rdfs:label ?personLabel_ru.
      FILTER (lang(?personLabel_ru) = 'ru')
    }
    OPTIONAL {
      ?person wdt:P569 ?birthDate .
  BIND(year(?birthDate) AS ?birthYear)
    }
    OPTIONAL {
      ?person wdt:P21 ?gender .
      ?gender rdfs:label ?genderLabel_en .
      FILTER (lang(?genderLabel_en) = 'en')
    }
    OPTIONAL {
      ?person wdt:P1559 ?nativeName .
    }

    OPTIONAL {
      ?person wdt:P172 ?ethnic_group .
      ?ethnic_grouprdfs:label ?ethnic_groupLabel .
      FILTER (lang(?ethnic_groupLabel) = 'en')
      SERVICE wikibase:label { bd:serviceParamwikib
ase:language "[AUTO_LANGUAGE],en". }

    }
    OPTIONAL {
      ?person wdt:P735 ?firstName .
      ?firstNamerdfs:label ?firstNameLabel
      FILTER (lang(?firstNameLabel) = 'en')
      }
    OPTIONAL {
      ?person wdt:P734 ?lastName .
      ?lastNamerdfs:label ?lastNameLabel
      FILTER (lang(?lastNameLabel) = 'en')
      }
    OPTIONAL {
      ?person wdt:P5056 ?patronymName .
     ?patronymNamerdfs:label ?patronymNameLabel
      FILTER (lang(?lastNameLabel) = 'en')
      }
    OPTIONAL {
      ?person wdt:P511 ?titulName .
      ?titulNamerdfs:label ?titulNameLabel
      FILTER (lang(?lastNameLabel) = 'en')
      }

  }
  LIMIT 50
```

Следует отметить, что перебирание по странам неизбежно приведет к тому, что многие этнические группы (культуры) не будут включены в датасет. Поэтому необходимо еще предпринять следующие шаги:

1) собрать перечень многонациональных стран;

2) вместо поля P27 использовать P19. И в качестве значения передавать названия национально-государственных образований внутри государств-членов ООН.

В итоге получился датасет, который размещен в публичном репозитарии на *GitLab* [3]. *Gitlab* – это веб-сервис для хостинга IT-проектов и их совместной разработки, позволяющий безвозмездно хранить свободно распространяющееся программное обеспечение и датасеты. Следует отметить, что для проведения качественного тестирования на собранном датасете всех общедоступных продуктов (сервисов) необходимо разработать соответствующее программное обеспечение. В силу трудозатратности и ограничений на объем публикации было принято решение посвятить отдельную статью разработке такого программного обеспечения и результатам его применения.

СПИСОК ПРОДУКТОВ И СЕРВИСОВ ДЛЯ ПАРСИНГА И АНАЛИЗА ПОЛНЫХ ЧЕЛОВЕЧЕСКИХ ИМЕН

На сегодняшний день существуют несколько инструментов и методов анализа полных человеческих имен. Все они имеют свои особенности и, соответственно, отличаются и по качеству работы, и по иным характеристикам.

Nameparser

Nameparser – это Python-библиотека, предназначенная для разделения полных имен на их составляющие части, такие как имя, отчество, фамилия и титулы. Она предоставляет простой и гибкий интерфейс для работы с именами, но может столкнуться с трудностями при обработке имен из неанглоязычных стран [4].

human-name

human-name – еще одна Python-библиотека для анализа и сравнения полных имен. «human-name» лучше всего работает с латинскими именами, то есть с данными из Северной или Южной Америки и/или Европы. Например, он не понимает форматы написания первой фамилии без запятых, распространенные в Восточной Азии: «Пак Кын Хе» будет проанализировано как имеющиеся имя «Пак» и фамилия «Кын Хе». И он не обрабатывает имена, состоящие из одного слова. Поскольку авторы этой библиотеки большое внимание уделяли именно такой функциональности, как сравнение имен, а также и эффективности использования памяти, анализируемые имена нормализованы в Unicode NFKD и пишутся с заглавной буквы обычным способом (обрабатывается «Mc» и несколько других крайних случаев), а необработанные входные данные не сохраняются [5].

Parserator

Parserator – инструмент для создания собственных

парсеров на основе наборов данных и правил. Его можно использовать для разработки специализированных парсеров полных имен, которые могут быть лучше адаптированы к обработке имен из разных стран и культур. Основным недостатком Parserator является необходимость вручную разрабатывать правила для парсинга имен и обучать модель на основе этих правил [6].

Rosette Name Indexer (API)
Rosette предоставляет API для извлечения информации из имени, такой как пол, а также для транслитерации имен [7]. В основном используется для верификации и сравнения имен.

Name API
Name API – это проприетарная сервисная платформа для работы с именами. Он предоставляет функциональность в виде веб-сервисов для синтаксического анализа имен, определения пола имен, сопоставления имен, форматирования имен и многого другого. В функционал входит:

1) извлечение имен и фамилий из строки полного имени;
2) определение пола человека;
3) профилирование культуры;
4) проведение различия между физическими и юридическими лицами (предприятиями);
5) идентификация отдельных людей в строке имен [8].

Namsor
Namsor API предлагает сервисы для определения гендерной структуры имени, оценки этнического происхождения и определения страны происхождения. Таксономия имеет 192 страны. По ним определяется происхождение человека. Этническая принадлежность имени лучше работает на мультикультурных странах [9]. Плохо определяет нетитульные этносы в постсоветских странах.

РАЗРАБОТКА РЕКОМЕНДАЦИЙ ДЛЯ БУДУЩИХ ИССЛЕДОВАНИЙ

На основе проведенного сравнительного анализа технологий для парсинга и анализа полных имен разрабатываются рекомендации для их дальнейшего использования и развития:

1. Для проектов с ограниченными ресурсами и временем, где основной фокус на англоязычные имена, можно использовать Nameparser или human-name. Однако следует учесть их ограничения в отношении имен из других культур и языков.

2. Для проектов, которым необходима более высокая точность и адаптивность к различным языкам и культурам, рекомендуется использовать Parserator. Он потребует больше времени и усилий для настройки и обучения, но в результате обеспечит более высокую точность и универсальность.

3. Возможным направлением для дальнейших исследований является интеграция существующих инструментов-парсеров с машинным обучением и искусственным интеллектом, чтобы повысить точность и эффективность обработки полных имен.

4. Разработка универсального инструмента для парсинга и анализа полных имен, который бы сочетал в себе преимущества всех рассмотренных инструментов и обеспечивал бы высокую точность для разных языков и культур, является перспективным направлением для будущих исследований.

ЗАКЛЮЧЕНИЕ

Парсинг и анализ полных человеческих имен может быть полезным для многих задач, включая исследования в области социальных сетей, гендерной и этнической статистики (тем более, как показали исследования, по фамилиям можно достаточно точно определять не только этническое происхождение, но и определять лингвистические особенности [10]), машинного обучения и т. д. В этой статье было рассмотрено несколько инструментов и методов для парсинга и анализа полных имен, такие как регулярные выражения, библиотеки для обработки естественного языка. Каждый из них имеет свои преимущества и недостатки, и выбор конкретного инструмента или метода зависит от конкретной задачи. Был собран датасет для тестирования различных инструментов и даны рекомендации по дальнейшим исследованиям.

ЛИТЕРАТУРА

1) NameAPI - Intelligence in Names URL: https://www.nameapi.org/ (дата обращения: 01.02.2023).
2) Wikidata // URL: https://www.wikidata.org/wiki/Wikidata:Main_Page (дата обращения: 18.04.2023).
3) https://gitlab.com/thekonstantin/nomanalyzer/-/blob/main/DataSetCountries.rar
4) Python Human Name Parser URL: https://nameparser.readthedocs.io/en/latest/?badge=latest (датаобращения: 18.04.2023).
5) human-name URL: https://github.com/djudd/human-name (дата обращения: 18.04.2023).
6) Parserator // Github URL: https://github.com/datamade/parserator (дата обращения: 18.04.2023).
7) Rosette Name Indexer (API) // Rosetter URL: https://www.rosette.com/ (дата обращения: 18.04.2023).
8) nameapi // nameapi URL: https://www.nameapi.org/ (дата обращения: 18.04.2023).
9) name checker for gender, origin and ethnicity determination // Namsor URL: https://namsor.app/ (дата обращения: 18.04.2023).
10) Jens Kandt , Paul A. Longley Ethnicity estimation using family naming practices // PLoS ONE, 2018. №13