Депонируемые материалы к программе “Классификатор медицинских услуг в сфере электронного здравоохранения для оптимизации взаимодействия участников модели e-health”

- Программа для ЭВМ с использованием компьютерного анализа и синтеза естественных языков (NLP) для оптимизации системы электронного здравоохранения в России.

Авторы:
Александра Бутнева | Aleksandra Butneva
Гюзель Гумерова | Gusel Gumerova
Стефан Хюзиг | Stephan Hüsig
Герхард Шеве | Gerhard Schewe
Эльмира Шаймиева | Elmira Shaimieva

Использованные языки программирования:
Python 3.8 (64-bit) for Windows 10
C++ Built Tools in Visual Studio
Command-line interface (CLI)

Содержание

1 Предисловие

Обработка естественного языка (англ. Natural Language Processing; NLP) представляет собой одно из передовых направлений искусственного интеллекта и математической лингвистики. Программы на основе NLP способны построить мост между языком компьютера и нашим пониманием языка как знаковой системы, служащей предметом общения между людьми. Таким образом, NLP позволяет обрабатывать и моделировать тексты с минимальным участием человека, что сокращает время на их анализ и классификацию.

Время – это, безусловно, один из ключевых факторов в вопросах здравоохранения, поэтому применение NLP в этой области особенно актуально. С учетом продолжающейся пандемии коронавируса мы хотим подчеркнуть, что поддержка инноваций и цифровых решений в здравоохранении – наша первичная цель при разработке компьютерных программ. Технологические решения, предложенные в данной работе, направлены на использование искусственного интеллекта с устойчивым и осознанным подходом к существующим проблемам в э-здравоохранении.

Также стоит отметить, что русский язык – один из самых сложных языков для изучения и преподавания, по мнению лингвистов. Это объясняется грамматическими и лексическими особенностями русской речи. Однако в XXI веке не только человек сталкивается со сложностями при овладевании новым языком, но и компьютер. Создавая лингвистическую модель, которая способна работать со специфической научной (медицинской) терминологией, мы предлагаем технологическое решение данной проблемы в области здравоохранения.

Наша международная исследовательская группа состоит из пяти ученых, работающих в различных дисциплинах социальных наук: политологии, экономики, бизнес-администрирования, педагогики и цифрового менеджмента. За 2019-2020 гг. мы опубликовали теоретические статьи, учебные пособия и монографии в области инновационного менеджмента электронного здравоохранения.

Ниже представлены основные наши работы, которые находятся в свободном доступе на сайте научной электронной библиотеки elibrary:

- Шеве Г., Гумерова Г.И., С. Хюзиг, Шаймиева Э.Ш. Менеджмент цифровой экономики. Менеджмент 4.0. Монография // М.: КНОРУС, М.: КНОРУС, 2019. 232 с.
- Шеве Г., Гумерова Г.И., С. Хюзиг, Шаймиева Э.Ш. Менеджмент 4.0 цифровой экономики Германии: опыт и инструменты для цифровой экономики России // Познание КИУ, 2020. 75 с.
- Шеве Г., Гумерова Г.И., С. Хюзиг, Шаймиева Э.Ш. Менеджмент организаций цифроовй экономики : учебное пособие // М.: КНОРУС. 2020. 384 с.

- Гумерова Г.И., Шаймиева Э.Ш., Бутнева А.Ю., Рафикова Н.Н. Сбор и переработка отходов (пластика) как социальная проблема городов. Развитие социальной политики на основе изменений социальных механизмов и использования цифровых технологий // Государственное управление. Электронный вестник. 2020. № 81. С. 233-259

Вышеперечисленные работы послужили теоретической основой для создания нашей программы с учетом опыта немецкой Индустрии 4.0. Функционал программы, руководство по ее установке, возможные области применения, а также графические изображения подробно описаны в разделах 2–5.

Авторами отмечается, что в связи с ограничениями, налагаемыми на максимальный размер программы Федеральным Институтом Промышленной Собственности, виртуальная среда не была включена в машиночитаемый код, т.е. все необходимые пакеты должны быть загружены вручную. Виртуальная среда также предоставляется правообладателями по требованию. Авторы обращают особое внимание на то, что программа не содержит конфиденциальной информации и/или личные данные пользователей.

2 Инструментарий программы

2.1 Технические характеристики

- *Основной язык программирования*: Python 3.8 (64-bit)
  Среда разработки: PyCharm 2020 Community
  Утилиты: C++ Build Tools в Visual Studio, CMD
- *Операционная система разработчика*: Windows 10
- *Использованные NLP библиотеки*: SpaCy, Stanza
- *Использованные пакеты Python*: os, spacy, torch, stanza, spacy_stanza, __future__, collections, pandas, glob, textblob, plotly.express, plotly.offline, numpy
- *Средняя скорость программы*: 2 мин / статья (10-15 стр.); варьируется в зависимости от мощности процессора компьютерного устройства и размера текста
- *Продукты программы*: документы в HTML-формате, интерактивные HTML-графики, рамки данных (csv)
- *Особенности программы*: чувствительность к пустым строкам в txt-файлах; полная поддержка инпут-файлов **только** в формате txt (веб- и пдф-скрейпинг вводных файлов возможен в индивидуальном порядке); для комфортной работы в интерфейсе программы требуется знание английского языка на начальном/среднем уровне.

2.2 Качественные характеристики

В Таблице 1 представлена характеристика программы для ЭВМ “Классификатор медицинской терминологии в сфере электронного здравоохранения для оптимизации взаимодействия участников модели e-health”. Авторами отмечается, что наличие знаний в области NLP не является обязательным для пользователей программы. В связи с новизной используемой терминологии и спецификой компьютерной лингвистики авторами были подготовлены ответы на семь наиболее часто задаваемых вопросов в разделе 2.4. Более опытным пользователям также рекомендуется ознакомиться с ответами на вопросы перед использованием программы.

Программа создана на основе двух бесплатных библиотек для NLP на языке Python 3.8: SpaCy и Stanza (предыдущее название библиотеки – StanfordNLP). Каждая библиотека содержит пре-программированные лингвистические модели для обработки естественных языков. В целях улучшения качества обработки русского языка, лингвистические модели были расширены авторами вручную. Научные работы в области телемедицины и электронного здравоохраения были проанализированы авторами программы, что позволило внести следующие дополнения во встроенный классификатор текста:

1. Аббревиатура наиболеее распространенных телемедицинских понятий (*Источник: на основе работ А.В.Владзимирского*):
   АД – артериальное давление
   АМН – академия медицинских наук
   АН – академия наук
   БРТМ – биорадиотелеметрия

Таблица 1: Характеристика программы “Классификатор медицинской терминологии в сфере электронного здравоохранения для оптимизации взаимодействия участников модели e-health”

Показатель	**Описание**
Объект	система российского электронного здравоохранения
Предмет	Взаимоотношения участников системы российского электронного здравоохранения: - получатели медицинской услуги (пациенты – P), - исполнители (медицинский и административный персонал – D, - страховые компании – I.
Цели	оптимизация взаимоотношений между тремя участниками модели российского здравоохранения (P, D, I)
Целевая группа программы	получатели медицинской услуги (пациенты – P), исполнители, т.е. медицинский и административный персонал (D), страховые компании (I)
Задачи	1. Усовершенствование NLP библиотеки в целях распознавания русского языка на высоком уровне (научная литература, медицинские акты) 2. Дефиниция основных сокращений в области телемедицины на основе работ, опубликованных в журналах: “Врач и информационные технологии” (RSCI, ВАК), “Менеджер здравоохранения” (ВАК), “Журнал телемедицины и электронного здравоохранения” (РИНЦ) 3. Дефиниция основных участников э-здравоохранения на основе теоретических работ авторов программы 4. Визуализация обработанных текстов и семантических трендов для получения медицинским персоналом информации о содержании текста
Этапы	Авторами выделены четыре этапа оптимизации взаимоотношений между P, D и I. Настоящая программа технически поддерживает и автоматизирует первый этап по направлениям “Телемедицина”,“Телеконсультирование”, “Теледиагностика” , “Веб-порталы здоровья”. Базой данных научной литературы, медицинские актов, дефиниций выступили статьи, опубликованные в журналах: “Врач и информационные технологии”, “Менеджер здравоохранения”, “Журнал телемедицины и электронного здравоохранения”. Критериями отбора направлений первого этапа являются: - актуальное положение системы э-здравоохранения; - кол-во научных статей в журналах “Врач и информационные технологии”, “Менеджер здравоохранения”, “Журнал телемедицины и электронного здравоохранения”. На этапах 2-4 планируется расширить программу для внедрения инновационных решений NLP по следующим 11 направлениям электронного здравоохранения, с включением новых терминов: телелаборатория, телемониторинг, электронное назначение лекарств, медицинская документация, персональный менеджер здоровья, социальные сети здравоохранения, профессиональная служба каталогов, е-обучение, электронное фактурирование, е-оплата, е-врачебное письмо, е-документы по выписке, планирование ресурсов и сроков. (*Источник: разработки авторов*)

ВК(С) – видеоконференция, видеоконференцсвязь
ВОЗ – Всемирная организация здравоохранения
д.б.н. – доктор биологических наук
ДДЦ – дистанционный диагностический центр
д.м.н. – доктор медицинских наук
ДО – дистанционное обучение
д.п.н. – доктор педагогических наук
д.т.н. – доктор технических наук
ИТ - информационные технологии
к.б.н. – кандидат биологических наук
КГР - кожно-гальванические реакции
к.м.н. – кандидат медицинских наук
КПК – карманный персональный компьютер
к.п.н. – кандидат педагогических наук
КТ – компьютерная томограмма (томография)
к.т.н. – кандидат технических наук
ЛИС – лабораторная информационная система
ЛПУ – лечебно-профилактическое учреждение
МЗ – министерство здравоохранения
МИС – медицинская информационная система
МО – медицинская организация (объединение)
МОЭТ - метод оценки эффективности телемедицины
МРТ – магнитно-резонансная томограмма (томография)
НИИ – научно-исследовательский институт
НПО – научно-производственное объединение
ОВП - общее виртуальное пространство
ОДС – опорно-двигательная система
ОКБ – областная клиническая больница
ОКД – областной кардиологический диспансер
ОНМК - острое нарушение мозгового кровообращения
ПГ – пневмограмма
ПО – программное обеспечение
РАЕН – Российская академия естественных наук
РИС – радиологическая информационная система
РКД – Республиканский кардиологический диспансер
СКГ – сейсмокардиограмма
СКТ – спиральная компьютерная томография
СМИ - система медицинских исследований
СМК - система медицинского контроля

СМП – скорая медицинская помощь
ТК, ТМК – телемедицинская консультация
ТМП - телемедицинский пункт
ТМРС – телемедицинская рабочая станция
ТМЦ – телемедицинский центр
УЗИ – ультразвуковое исследование
ЦРБ – центральная районная больница
ЦФ – цифровая фотосъемка
ЧД – частота дыхания
ЧСС – частота сердечно-сосудистых сокращений
ЭВП - электронная виртуальная перчатка
ЭИБ – электронная история болезни
ЭКГ – электрокардиограмма (графия)
ЭКС – электрокардиосигнал
ЭОГ – электроокулограмма
ЭСП - электронная сенсорная перчатка
ЭЭГ – электроэнцефалограмма
3G – Third Generation
x(A)DSL – (Asymmetric) Digital Subscriber Line
AVI - Audio Video Interleave
CD/DVD-ROM - Compact Disc /Digital Versatile Disc -Read-Only Memory
CDMA - Code Division Multiple Access
DICOM - Digital Imaging and Communications in Medicine
EDGE - Enhanced Data rates for GSM Evolution
FTP - file transfer protocol
GPRS - General Packet Radio Service
GSM - Global System for Mobile Communications
HL7 – Health Level 7
IBID – от лат. Ibidem – «там же», «в том же месте», означает, что текст цитируется по вышепреведенному источнику
IP - Internet Protocol
IrD - Infrared Data Association
ISDN - Integrated Services Digital Network
ISO – International Standart Organisation
JPEG – Joint Photographic Experts Group
MP3 - Moving Pictures Experts Group-1/2/2.5 Layer
MMS - Multimedia Messaging Service
MPEG - Moving Pictures Experts Group
NASA – The National Aeronautics and Space Administration

NIHSS - National Institutes of Health Stroke Scale
NTSC - National Television Standards Committee
PACS - Picture Archiving and Communication System
PAL - Phase-Aternating Line
PDA – Personal Digital Assistant
PDF – Portable Document Format
RGB - Red, Green, Blue
RTF - Rich Text Format
SCG-ECG - Standard Communication Protocol - Computer-Assisted Electrocardiography
SMS - Short Message Service
TIFF - Tagged Image File Format
TFT - Thin Film Transistor
USB - Universal Serial Bus
VoIP – Voice Over Internet Protocol
VPN - Virtual Private Network
WAP - Wireless Application Protocol
WAV – сокращение от Wave
WMA - Windows Media Audio

2. Участники бизнес-модели электронного здравоохранения в Российской Федерации (*Источник: на основе работ авторов программы*) в форме лемм [леммы указываются в квадратных скобках]:
   - Пациент [PATIENT] – получатель медицинской услуги
   - Врач [DOCTOR] – исполнитель медицинской услуги
   - Государственная или частная страховая компания [INSURANCE] – поставщик медицинской услуги

3. Категории для классификации документов, относящихся к четырем этапам развития электронного здравоохранения в РФ (*Источник: разработки авторов*):
   - Телемедицина [TELEMED]
   - Телеконсультирование [TELECONSULT]
   - Теледиагностика [TELEDIAG]
   - Телелаборатория [TELELAB]
   - Телемониторинг [TELEMONIT]
   - Электронное назначение лекарств [EMEDICINE]
   - Медицинская документация [MEDDOC]
   - Веб-порталы здоровья [WEBHEALTH]
   - Персональный менеджер здоровья [HEALTHMANAG]

- Социальные сети здравоохранения [SNETWORK]
- Профессиональная служба каталогов: регистры, список медикаментов, реестр медикаментов, врачебный список [CATALOG]
- E-обучение [ELEARN]
- Электронное фактурирование (предоставление электронного счета и его оплата) [EBILL]
- E-оплата, E-врачебное письмо, e-документы по выписке [EPAY]
- Планирование ресурсов и сроков: электронная коммуникация заказов, планирование лечения, планирование операций, планирование персонала, планирование сроков, онлайн-согласования сроков [EPLAN]

Вышеперечисленная терминология была добавлена в базовые лингвистические модели в качестве именованных сущностей (англ. entities [ENT]) для дальнейшего использования в развитии российского электронного здравоохранения.

2.3 Области применения программы

Авторами рекомеднованы следующие области применения программы:

- Распознавание аббревиатур на уровне документа, т.е., каждый медицинский работник получит HTML-документ с выделенными аббревиатурами и их значением, что ускорит процесс анализа документа. Ниже предлагается код для получения доступа к именованным сущностям, идентифицированным программой в документе.

```
# импортируйте встроенный визуализатор displacy
from spacy import displacy

# создайте HTML-документ
displacy.serve(doc, style="ent")

# пройдите по ссылке ниже для просмотра HTML-документа
# http://127.0.0.1:5000
```

- Категоризация медицинской документации и научных статей в области телемедицины, т.е., каждый участник бизнес-модели электронного здравоохранения сможет оценить значимость каждой именованной сущности по кол-ву ее упоминаний в тексте, подсчитанному программой; идентичная описательная статистика может быть подсчитана для аббревиатур.

- Основанные на предлагаемой авторами усовершенствованной лингвистической модели (а) визуализация би-грамм, (б) сентимент-анализ, (в) идентификация тематики документов, (г) анализ трендов и (д) статистическое прогнозирование могут быть использованы участниками бизнес-модели электронного здравоохранения для автоматизированного

анализа документации в области e-health и ее интерактивного графического представления с минимальным участием человека. Примеры графического анализа на основе NLP приведены в разделе 4.

2.4 Словарь NLP

В Таблице 2 представлены основные английские понятия из активного вокабуляра NLP и их русские эквиваленты.

Таблица 2: Базовые понятия в области NLP

Английская аббревиатура	**Русский перевод**	**Описание / Примеры**
NLP (Natural Language Processing)	обработка естественного языка	метод, основанный на искуственном интеллекте, который позволяет компьютеру распознавать и анализировать письменную речь
ENT (entity)	сущность	категория предметов или обобщающее понятие (напр., “город”, “время года”, “профессия”)
POS (part of speech)	часть речи	категория слов с общими морфологическими и синтаксическими признаками
LEMMA (lemma)	лемма	каноническая (“базовая”) форма слова
NER (Named Entity Recognition)	определение именованных сущностей	компьютерный метод распознавания обобщающих понятий в тексте
token	токен	последовательность символов в лексическом анализе в информатике, соответствующая лексеме
label	лейбл	наименование для определенной группы слов или символов
sentence	предложение	законченная единица языка и речи
DEP (dependency)	зависимость	зависимость слов в предложении (словосочетания)
pipeline	канал	виртуальное хранилище компонентов программы, которое управляется пользователем
component	компонент	функция программы, которая модет быть (де)активирована или добавлена вручную
displacy	дисплей	визуализатор библиотеки SpaCy; функция, позволяющая создавать веб-объекты на основе текста, обработанного с помощью NLP (*Источник: разработки авторов*)

2.5 Ответы на часто задаваемые вопросы

1. **Какие действия стоит предпринять пользователям без базовых знаний компьютерной лингвистики для работы с программой?**
   Отсутствие базовых знаний не является проблемой. Разработчики библиотеки SpaCy

подготовили интернет-гид на английском языке по интерфейсу SpaCy и ее основным функциям. Также рекомендуется пройти бесплатный онлайн-курс по основам работы с естественными языками и ознакомиться с NLP словарем в разделе 2.3.

2. **Какими функциями обладает программа?**
   Основные функции программы – это (а) распознавание и обработка русского языка на высоком уровне; (б) разложение цельного текста на предложения; (в) токенизация текста; (г) определение именованных сущностей в соответствии с предложенной классификацией; (д) идентификация значимых понятий из области телемедицины и э-здравоохранения; (е) графическая визуализация грамматических структур предложений, описательной статистики, телемедицинских служб и терминов. Подробно с областями применения и функциями программы можно ознакомиться в разделе 2.2.

3. **Что понимается под "текстом" для программы?**
   Исходным текстом для программы послужили научные статьи, монографии и пособия в области инновационного менеджмента электронного здравоохранения. Однако возможности программы не ограничиваются автоматизацией документации в рамках здравоохранения, поскольку, благодаря встроенной функции расширения словарей вручную, пользователи могут работать с русскоязычными текстами на любую тематику.

4. **Каким образом программа способствует развитию электронного здравоохранения в России?**
   Программа оптимирует взаимодействие между участниками модели электронного здравоохранения в России (клиентом, врачом и государственной/частной страховой компанией). В частности, NLP обработка одной статьи занимает ок. 2 минут, что позволяет значительно ускорить анализ текста.

5. **Могут ли пользователи программы наряду с научными текстами из журналов использовать терминологию из врачебной практики? Если "да", что в этом случае получит каждый из ее пользователей?**
   Да, программа может быть использована в целях получения информации из учебных и методических материалов.

   В зависимости от запроса, пользователь категории D (доктор/медицинский персонал) получит информацию в виде HTML-документа с выделенными именованными сущностями "Телемедицина", "Телеконсультирование", "Теледиагностика", что соответствует первому этапу разработки настоящей программы, а также термины из раздела 2.1 (аббревиатура наиболеее распространенных телемедицинских понятий). Также пользователь сможет активировать функцию дефиниции понятий и расширять словарь вручную, напр. при появлении новых терминов для обозначения инновации или ранее не исследованного заболевания, как показано ниже.

```
# Расширение словаря вручную
# импортируйте пакеты
from spacy.matcher import PhraseMatcher
from spacy.tokens import Span

# определите понятие, которое Вы хотите добавить в словарь
ddc = ['ДДЦ'] # аббревиатура инновации

# определите паттерн (комбинация знаков для поиска совпадений)
pattern = [nlp(ddc) for ddc in ddcs]

# добавьте понятие в компонент matcher
matcher = PhraseMatcher(nlp.vocab)
matcher.add('Дистанционный Диагностический Центр', None, *pattern)

# проведите NLP анализ
doc = nlp(text)
matches = matcher(doc)
for match_id, start, end in matches:
    # создайте новый спан и используйте ID в качестве лейбла
    span = Span(doc, start, end, label=match_id)
    # добавьте спан к именованным сущностям документа
    doc.ents = list(doc.ents) + [span]

# проверьте наличие понятия в ENT
print([(ent.text, ent.label_) for ent in doc.ents])
```

Пользователь категории I (страховая компания) получит информацию, аналогичную пользователю D. Также пользователь I сможет лемматизировать текст (см. код ниже) и посчитать статистику упоминаний аббревиатур, интересующих страховую компанию. Также страховая компания сможет оценить географию своих клиентских запросов, доминирующие темы и проблемы э-здравоохранения. Результаты оценки могут быть использованы I для оптимизации спроса и предложения впоследствии.

```
# Лемматизация текста с помощью рамки данных pandas
# импортируйте необходимые дополнительные пакеты
import pandas as pd
import string
import numpy as np
```

```
# проведите NLP анализ документа
nlp = StanzaLanguage(snlp)
doc = nlp(text)

# создайте лемму для каждого слова в документе
joint = ' '.join(token.lemma_ for token in doc)
str(joint)
lemmas = word_count(joint)

# создайте рамку данных с столбцами ``Лемма" и ``Частота употребления"
df = pd.DataFrame(list(lemmas.items()), columns=['Лемма',
'Частота употребления'])
df['Лемма'] = df['Лемма'].apply(remove_punctuation)

# удалите пустые ряды и пунктуацию (по желанию)
df['Лемма'].replace('', np.nan, inplace=True)
df.dropna(subset=['Лемма'], inplace=True)

# сохраните Вашу рамку данных в формате csv для дальнейшего анализа
df.to_csv('ваш путь к файлу/df.csv', encoding="utf-8")
```

Пользователь категории P (пациент) также получит доступ к вышеописанной информации. В частности, результаты NLP анализа пациентской карты в виде HTML-документа с выделенными именованными сущностями помогут пациенту разобраться в аббревиатурах из области телемедицины, напр., напротив “ДДЦ” программа поставит расшифровку “Дистанционный Диагностический Центр”. Это избавит пациента от дополнительных затрат времени на поиски необходимых определений в сети Интернет.

6. **Каков размер программы?**
   Размер программы, с учетом виртуальной среды и библиотек, составляет 154 байт. Размер машиночитаемого кода без учета виртуальной среды – 57,7 Кб.

7. **Какие возможности для контакта существуют в случае возникновения технических проблем во время установки и/или использования программы?**
   По техническим вопросам, а также по вопросам, связанным с визуализацией данных и интерактивными инструментами рекомендуется обращаться к разработчику программы – Александре Юрьевне Бутневой – по электронной почте (alexabutneva@mail.ru; aleksandra.butneva@gmx.de). Электронная переписка может осуществляться на русском, английском и немецком языках по желанию пользователя.

8. **Какие возможности для контакта существуют по вопросам сотрудничества и применения программы в цифровом менеджменте российского здравоохранения?**
   По вопросам сотрудничества и использования программы в области электронного здравоохранения и цифрового менеджмента рекомендуется обращаться к А.Ю. Бутневой (см. Вопрос 7).

3 Руководство по установке программы

Как было отмечено ранее, программа написана в операционной системе Windows 10 на языке Python 3.8 (64-bit) с поддержкой C++ Build Tools в Visual Studio и командной строки. Программа полностью опирается на данные ресурсы, поэтому их наличие в операционной системе Windows и компиляция с PyCharm необходимы для успешного запуска программы. Рекомендуется придерживаться пошагового руководства по установке для избежания технических проблем. В шагах 1-5 пользователю предлагается установить необходимые пакеты и библиотеки в интерфейсе командной строки. Последующие шаги предусматривают работу в PyCharm.

- Шаг 1. Удостоверьтесь, что на Вашем компьютере установлены PyCharm 2020 и Python 3.8 (64-bit). Если данные программы отсутствуют, установите PyCharm 2020 Community Version здесь и Python 3.8 for Windows 10 (64-bit) здесь.

- Шаг 2. Удостоверьтесь, что на Вашем компьютере поддерживается pip для Python 3.8. Для этого откройте командную строку нажатием комбинации клавиш WIN+R. В открывшемся окне запуска приложений введите “cmd” и нажмите “Ok”. В окне командной строки введите:

```
pip help
#при запуске кода не должна отображаться ошибка
```

- Шаг 3. Установите C++ Build Tools в Visual Studio (Community) по ссылке. Убедитесь, что все расширения C++ отмечены ✓при установке. Если на данном этапе возникают проблемы, установите полную бесплатную версию Visual Studio (Community). Поддержка C++ необходима для оптимизации работы библиотек NLP.

- Шаг 4. Установите NLP библиотеку SpaCy через командную строку:

```
pip install -U spacy
```

- Шаг 5. Конфигурируйте пакет torch, выбрав один из предложенных способов, и установите расширение spacy-stanza:

```
# вариант 1 - клонирование репозитория
git clone https://github.com/torch/distro.git ~/torch --recursive

# вариант 2 - классическая установка через pip
pip install torch==1.6.0+cpu torchvision==0.7.0+cpu -f
https://download.pytorch.org/whl/torch_stable.html

# установите расширение spacy-stanza
pip install spacy-stanza
```

- Шаг 6. В интерфейсе PyCharm импортируйте установленные пакеты (Script 1):

```
import spacy
# если возникают проблемы, добавьте gpu ниже
spacy.prefer_gpu()

# загрузите библиотеку русского языка
from spacy.lang.ru import Russian

# конфигурируйте установленный torch
import torch
from torch.autograd import Variable
import torch.nn as nn
import torch.nn.functional as F

# импортируйте расширение spacy-stanza
import stanza
from spacy_stanza import StanzaLanguage

# загрузите существующую модель русского языка
stanza.download('ru')

# конфигурируйте каналы на русском языке
snlp = stanza.Pipeline(lang="ru")
```

- Шаг 7. Загрузите файл в формате txt для NLP анализа и создайте NLP объект:

```
nlp = StanzaLanguage(snlp)
import os # импортируйте пакет os
```

```
# определите путь к файлу для анализа
abspath = os.path.abspath("ваш путь к файлу")
os.chdir(abspath)
# откройте файл
with open('ваш файл.txt', 'rt', encoding="utf-8") as file:
    text = file.read().replace('\n', ' ')
# проведите NLP анализ
doc = nlp(text)
```

После успешной установки программы пользователь получает NLP файл (doc) со встроенными функциями лемматизации, токенизации, выделения именованных сущностей и разложения текста на предложения:

```
# разложение текста на предложения
sentences = str([sent.text for sent in doc.sents])

# выделение именованных сущностей
entities = [(ent.text, ent.label_) for ent in doc.ents]

# токенизация (разложение на отдельные слова)
for token in doc:
    print(token.text,       # текст
          token.lemma_,     # лемма
          token.pos_,       # часть речи
          token.dep_)       # лейбл
```

4 Графические изображения

4.1 Пояснение к интерактивной визуализации текста на основании результатов программы

Согласно руководству по установке и использованию программы, одна из ключевых функций программы – это распознавание границ предложений в тесте, т.е. продуктом программы в данном случае является ряд отдельных предложений. Предложение в аналитике данных – это ценная единица для компьютерного анализа, которая содержит сведения о контексте, грамматической структуре, эмоциональной окраске и др. характеристиках высказывания. В целях демонстрации функций программы авторами осуществлена визуализация текстов статей трех ведущих российских журналов в области э-здравоохранения: “Врач и информационные технологии”; “Менеджер здравоохранения”; “Журнал телемедицины и электронного здравоохранения”.

Интерактивные графики созданы при поддержке пакетов Plotly и Pandas в виртуальной среде программы. После дифференциации цельного текста до уровня отдельных предложений формируется “рамка данных” (англ. data frame) класса Pandas. Другими словами, для каждой статьи из научного журнала программой автоматически создается пронумерованный лист, единицей в котором является предложение.

Далее рамки данных получают уникальный идентификатор (анг. ID) по номеру статьи. Индивидуальные рамки формируют единую таблицу, содержащую разложенные по предложениям тексты статей с присвоенными ID. На основе единой таблицы данные преобразуются в интерактивные графики в интерфейсе Plotly, отражая семантические особенности текстов. В связи со статичностью формата PDF интерактивные и кликабельные версии HTML-графиков прилагаются в качестве отдельных файлов (см. "Программа/Графики/interactive").

Интерактивные графики поддерживаются интернет-браузерами компьютерного устройства. Посли открытия графика в браузере, пользователь имеет возможность регулировать изображение при помощи панели управления, доступной в правом верхнем углу (см. Рис. 1). Панель управления обладает следующими функциями (слева направо):

- прямое сохранение графика в формате PNG;
- автоматический цифровой зум;
- смещение фокуса графика;
- выбор бокса (сегмент графика, имеющий четырехугольную форму);
- выбор лассо (сегмент графика, не имеющий углов);
- увеличение (+) и уменьшение (−) фокуса графика вручную;
- автоскалирование;
- сброс (восстановление первоначальной разметки) осей графика;
- включение и выключение пунктирных контуров;
- отображение наиболее близких значений;
- сравнение данных;
- переадресация на сайт Plotly.

Рис. 1: Панель управления Plotly

4.2 Индекс читабельности текста статей в области э-здравоохранения

Средняя длина читабельного предложения в русском языке составляет от 15 до 20 слов. На Рис. 2 представлено графическое сравнение журналов “Врач и информационные технологии”; “Менеджер здравоохранения”; “Журнал телемедицины и электронного здравоохранения” на

интервале 10-25 слов, что позволяет оценить среднюю читабельность статей.

Рис. 2: Средняя читабельность статей из журналов “Врач и информационные технологии”, “Менеджер здравоохранения”, “Журнал телемедицины и электронного здравоохранения” на момент разработки настоящей программы

Источник: разработка авторов на основе официального сайта Научной электронной библиотеки elibrary

Как показано на Рис. 2, изученные статьи журнала “Врач и информационные технологии” более сбалансированны и читабельны для широкой аудитории. Данный вывод сделан на основе распределения длин предложений из выборки статей журнала.

4.3 География телемедицины

Рис. 3 – отражение территориальных единиц, упомянутых в выборке статей. Рис. 3 создан программой на основе метода определения именованных сущностей без участия человека. Таким образом, слова и словосочетания, образующие категорию “location” – прямой продукт искусственного интеллекта. Аналогичный код может быть использован для визуализации любой категории текста с помощью метода присвоения лейблов словам или расширения словарей вручную.

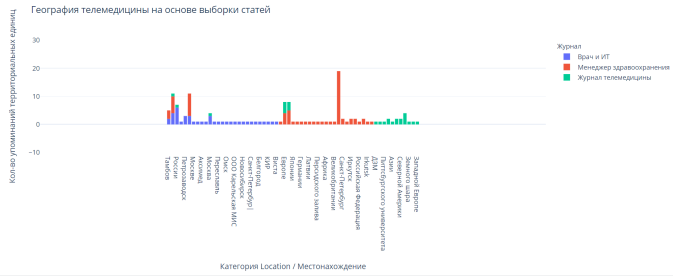

Рис. 3: География телемедицины

Источник: разработка авторов на основе статей из журналов “Врач и информационные технологии”, “Менеджер здравоохранения”, “Журнал телемедицины и электронного здравоохранения” на момент разработки настоящей программы, официального сайта Научной электронной библиотеки elibrary

4.4 Описательный анализ на уровне лемм

Анализу на Рис. 4 предшествует автоматизированное разложение текста статей до уровня базовых форм слов (лемм) – процедура под названием "лемматизация". Сначала программа преобразует словоформы в лексемы и создает индивидуальный словарь для каждого текста. Затем программа трасформирует словарь в рамку данных класса Pandas и рассчитывает количество упомянутых в документе лексем в зависимости от выбора пользователя. Рис. 4 – визуализация лемм, упоминающих трех участников бизнес-модели э-здравоохранения: пациента, врача и страховой компании.

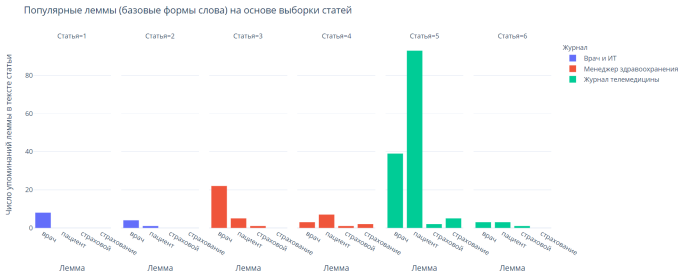

Рис. 4: Лемматизация выборки статей

Источник: разработка авторов на основе статей из журналов “Врач и информационные технологии”, “Менеджер здравоохранения”, “Журнал телемедицины и электронного здравоохранения” на момент разработки настоящей программы, официального сайта Научной электронной библиотеки elibrary

Аналогичный анализ ключевых слов первого этапа оптимизации взаимоотношений между P, D и I был проведен авторами на Рис 5. По результатам анализа наибольшее количество ключевых слов было выявлено в статье Владзимирского А. В. “История телемедицины первые 150 лет”.

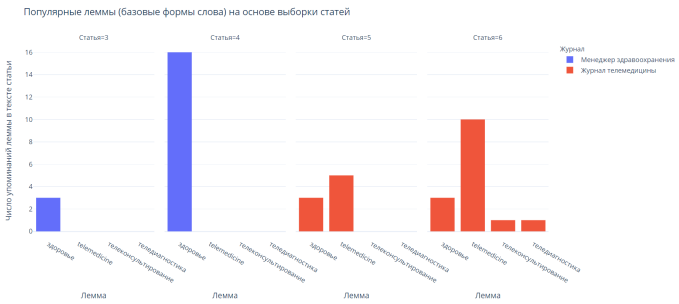

Рис. 5: Лемматизация ключевых слов выборки статей

Источник: разработка авторов на основе статей из журналов "Врач и информационные технологии", "Менеджер здравоохранения", "Журнал телемедицины и электронного здравоохранения" на момент разработки настоящей программы, официального сайта Научной электронной библиотеки elibrary

Поиск леммы по запросу пользователя внутри рамки данных осуществляется следующим образом:

```
# создайте лист лемм для поиска
lemma_list = ['лемма','лемма2', 'лемма3','лемма4']

# создайте новую рамку данных, содержащую совпадения
filtered_df = df[df['Лемма'].isin(lemma_list)]

# создайте гистограмму
fig = px.histogram(filtered_df, x="Лемма", y="Частота употребления",
                   color='Журнал',
                   facet_col='Статья',
                   title='Популярные леммы на основе выборки статей'
                   )
# по желанию, измените название оси y
fig.update_layout(yaxis_title_text='Число упоминаний леммы в тексте статьи')

# по желанию, включите режим офлайн
plotly.offline.plot(fig)
```

5 Ресурсы для апробации программы

5.1 Критерии отбора материалов

Для отбора материалов в целях апробации модели и визуализации результатов описательного анализа авторами были разработаны следующие критерии:
- количество цитирований в elibrary;
- рейтинг (импакт-фактор) научного журнала/учебного пособия/монографии;
- общая тематика (медицина и здравоохранение);
- профиль (электронное здравохранение, телемедицина, цифровой менеджмент, информационные системы);
- доступность статьи в формате PDF.

5.2 Статьи в научных журналах

- Владзимирский А. В. История телемедицины первые 150 лет // Журнал телемедицины и электронного здравоохранения. 2015. №1. URL: https://cyberleninka.ru/article/n/istoriya-telemeditsiny-pervye-150-let (дата обращения: 19.09.2020). [индекс цитирования: 12]

- Владзимирский А.В. Первичная телемедицинская консультация «Пациент-врач»: первая систематизация методологии // Журнал телемедицины и электронного здравоохранения. 2017. №2 (4). URL: https://cyberleninka.ru/article/n/pervichnaya-telemeditsinskaya-konsultatsiya-patsient-vrach-pervaya-sistematizatsiya-metodologii (дата обращения: 19.09.2020). [индекс цитирования: 14]

- Гусев А. В. Рынок медицинских информационных систем: обзор, изменения, тренды // Врач и информационные технологии. 2012. №3. URL: https://cyberleninka.ru/article/n/rynok-meditsinskih-informatsionnyh-sistem-obzor-izmeneniya-trendy (дата обращения: 19.09.2020). [индекс цитирования: 44]

- Пивень Д. В., Кицул И. С. О формировании новой системы контроля качества и безопасности медицинской деятельности в здравоохранении Российской Федерации // Менеджер здравоохранения. 2013. №2. URL: https://cyberleninka.ru/article/n/o-formirovanii-novoy-sistemy-kontrolya-kachestva-i-bezopasnosti-meditsinskoy-deyatelnosti-v-zdravoohranenii-rossiyskoy-federatsii (дата обращения: 19.09.2020). [индекс цитирования: 55]

- Фролов С. В., Фролова М. С. Мировые проблемы при выборе медицинского изделия для учреждения здравоохранения // Менеджер здравоохранения. 2013. №11. URL: https://cyberleninka.ru/article/n/mirovye-problemy-pri-vybore-meditsinskogo-izdeliya-dlya-uchrezhdeniya-zdravoohraneniya (дата обращения: 19.09.2020). [индекс цитирования: 54]

- Фролов С. В., Фролова М. С., Потлов А. Ю. Рациональный выбор медицинской техники для лечебно-профилактического учреждения на основе системы поддержки принятия решений // Врач и информационные технологии. 2014. №3. URL: https://cyberleninka.ru/article/n/ratsionalnyy-vybor-meditsinskoy-tehniki-dlya-lechebno-profilakticheskogo-\uchrezhdeniya-na-osnove-sistemy-podderzhki-prinyatiya (дата обращения: 19.09.2020). [индекс цитирования: 86]

5.3 Монографии

- Владзимирский А.В., Лебедев Г.С. Телемедицина // М., изд-во «ГЭОТАР-М.», 576 с. [кол-во цитирований 22] https://elibrary.ru/item.asp?id=36375052 (дата обращения: 19.09.2020)
- Владзимирский А.В. Телемедицина // Донецк, изд-во «Цифровая типография», 2011. 437 с. [кол-во цитирований 95] https://elibrary.ru/item.asp?id=26373429 (дата обращения: 19.09.2020)

6 Компоненты программы в виртуальной среде PyCharm

Программа состоит из 15 скриптов (кодовых файлов), отсортированных в алфавитном порядке:

1. Additional functions (Script 4)
2. Built-in visualization of findings with displacy (Script 5)
3. Creating Pandas for articles
4. Creating Pandas for entities
5. Creating Pandas for lemmas
6. Installation and preliminaries (Script 1)
7. Introduction of abbreviations (Script 3)
8. Introduction of e-health actors (Script 2)
9. Merging df
10. Merging df for entities
11. Merging df for lemmas
12. Vis of absolute length of sentences
13. Vis of e-health geography

14. Vis of text lemmas
15. Vis of text mean readability

Авторами рекомендован следующий порядок использования скриптов: **6 8 7 1 2 3 4 5 9 10 11**; далее – визуализация (**12 13 14 15**) в зависимости от запросов пользователя.