

«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ
КАФЕДРА ИНФОКОГНИТИВНЫХ ТЕХНОЛОГИЙ

Автоматическое рубрицирование текстов на основе word2vec.

Расчетно-пояснительная записка
курсового проекта по дисциплине
«Методы работы с большими данными»
студент группы 211-322
Сундуреев Тамир Юрьевич

Преподаватели:
проф. Ю.Н. Филиппович
асс. Н.Г.Воробьев

Москва, 2024 г.

ВВЕДЕНИЕ

Цель и задачи курсового проекта

Цель курсового проекта: приобретение навыков обработки и использования естественно-языковых текстовых данных.

Задачи курсового проекта:

- изучение материалов лекционных и практических занятий из курса LMS «Введение в интеллектуальные диалоговые системы» (разделы лекционных курсов — методы сбора данных и анализа запросов, методы формализации и оценки естественно языковых данных, методы технической реализации глубокого обучения и др.; практические занятия — практическое задание «Bag of words», практическое задание «Word2Vec» и др.; специальное информационное программное обеспечение — среда разработки IntelliJ IDEA, PyCharm или Visual Studio Code, текстовые файлы статей на тему дипломной работы; основная и дополнительная литература — см. список программ дисциплин);

- приобретение навыков анализа и изучение методов и приемов исследования ЕЯ описания предметной области (ПО) в процессе выполнения заданий курсовой работы;

- приобретение знаний и навыков интегрированного использования программного обеспечения (текстовых процессоров, электронных таблиц, специального программного обеспечения и других программных средств) для проведения обработки и использования ЕЯ ресурсов, характеризующих ПО;

- приобретение знаний и навыков по оформлению результатов анализа и исследования ПО при оформлении курсовой работы.

Описание предметной области

В качестве информационного ресурса предметной области были выбраны две медицинские темы — диабет 2-го типа (Type 2 Diabetes Mellitus) и кандидоз (Candidiasis). Они являются значимыми медицинскими проблемами, оказывающими существенное влияние на здоровье населения. Диабет характеризуется нарушением обмена глюкозы в организме, что приводит к серьезным последствиям для здоровья. Кандидоз — это грибковая инфекция, которая может поражать различные части тела, особенно у людей с ослабленной иммунной системой.

Дополнительно была выбрана тема для выпускной квалификационной работы — "Спортивное приложение для студентов". Приложение предназначено для организации тренировок, контроля физической активности и поддержки здорового образа жизни среди студентов.

Язык анализируемого текста — английский для медицинских тем и русский язык для темы выпускной квалификационной работы. Статьи выдержаны в официально-деловом и публицистическом стиле.

1 СБОР ЕЯ ТЕКСТОВОГО ОПИСАНИЯ ПО

При сборе ЕЯ текстовых описаний ПО были выполнены следующие работы:

- с помощью предложенных ресурсов найдены 20 научных статей, соответствующих подкатегориям выбранной категории ПО «Медицина»;
- с помощью предложенных ресурсов найдены 20 научных статей, соответствующих подкатегориям выбранной категории ПО бакалаврской ВКР «Спортивное приложение для студентов»;
- сохранен список названий статей и ссылок на них в виде таблицы;
- проведено предварительное ручное рубрицирование текстов.

Для ПО «Медицина» были выбраны подкатегории: «Type 2 Diabetes», «Candidiasis». Для ПО ВКР подкатегории: «здоровье», «спорт». Для статей медицинской тематики получили две основных статьи, остальные статьи, на которые ссылается основные. Для статей тематики ВКР воспользовались поиском в научной электронной библиотеке cyberleninka.ru.

При проведении ручного рубрицирования текстов изучили их содержание. Для каждого из них были обозначены ключевые слова. После выполнения первого задания была заполнена таблица, содержащая следующие столбцы: название статьи, ссылка на статью, дата обращения к ресурсу, ключевые слова, рубрика. Подкатегории ПО ВКР представлены на втором листе таблицы. Фрагмент таблицы представлен на рисунке 1. Таблица хранится в файле «Задание 1.xlsx».

F2						
Type 2 Diabetes						
	A	B	C	D	E	F
1	№	Название статьи	Ссылка на статью	Дата обращения	Ключевые слова	Подкатегория
		Daily Fasting Blood Glucose Rhythm in Male Mice: A Role of the Circadian Clock in the				Type 2 Diabetes
2	1	Liver	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Circadian Clock, blood glucose	
3	2	CLOCK and BMAL1 regulate M	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	circadian clock myofilaments mitochondria	
4	3	Circadian and feeding rhythms	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	circadian rhythms ribosome profiling mRNA translat	
5	4	Phototransduction by Retinal	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	retinal ganglionic cells, circadian clock	
6	5	Disruption of Circadian Insulin	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Type 2 Diabetes, circadian insulin secretion	
7	6	Evidence for a Circadian Rhythm	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Circadian Rhythm, Insulin Sensitivity	
8	7	Projection of the year 2050 bu	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	diabetes	
9	8	Beta-Cell Deficit and Increased	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Type 2 Diabetes, Beta-cell apoptosis	
10	9	Adverse Metabolic Consequen	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	metabolic syndrome, diabetes, circadian disruption	
11	10	Reciprocal Regulation of Brain	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	circadian rhythm, circadian clock, diabetes	
12	1	Clinical Practice Guideline for	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	candidemia; invasive candidiasis; fungal diagnostics; az	Candidiasis
13	2	Candida ESOPHAGITIS: SPECIE	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Candida; Opportunistic infections; Esophagitis; Endosco	
14	3	CANDIDA ESOPHAGITIS A pros	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	CANDIDA ESOPHAGITIS	
15	4	Candida Esophagitis in Achalas	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Achalasia cardia, Candida esophagitis, double contrast	
16	5	Candida Esophagitis: Feathery	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Barium esophagogram; Candida esophagitis; feathery	
17	6	An Elderly Case of Type 2 Diab	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	denture, candidiasis, hyperosmolar syndrome, ketoacid	
18	7	Guidelines for Treatment of Ca	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Candidiasis	
19	8	Long-Term Trends in Esophage	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Esophageal Candidiasis	
20	9	Fatal Esophageal Perforation	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Invasive Candidiasis	
21	10	Surgical Management of Necro	https://doi.org/10.1016/j.cmet.2012.01.001	12.01.2025	Candida Esophagitis	

Рисунок 1 – Сбор ЕЯ текстового описания ПО

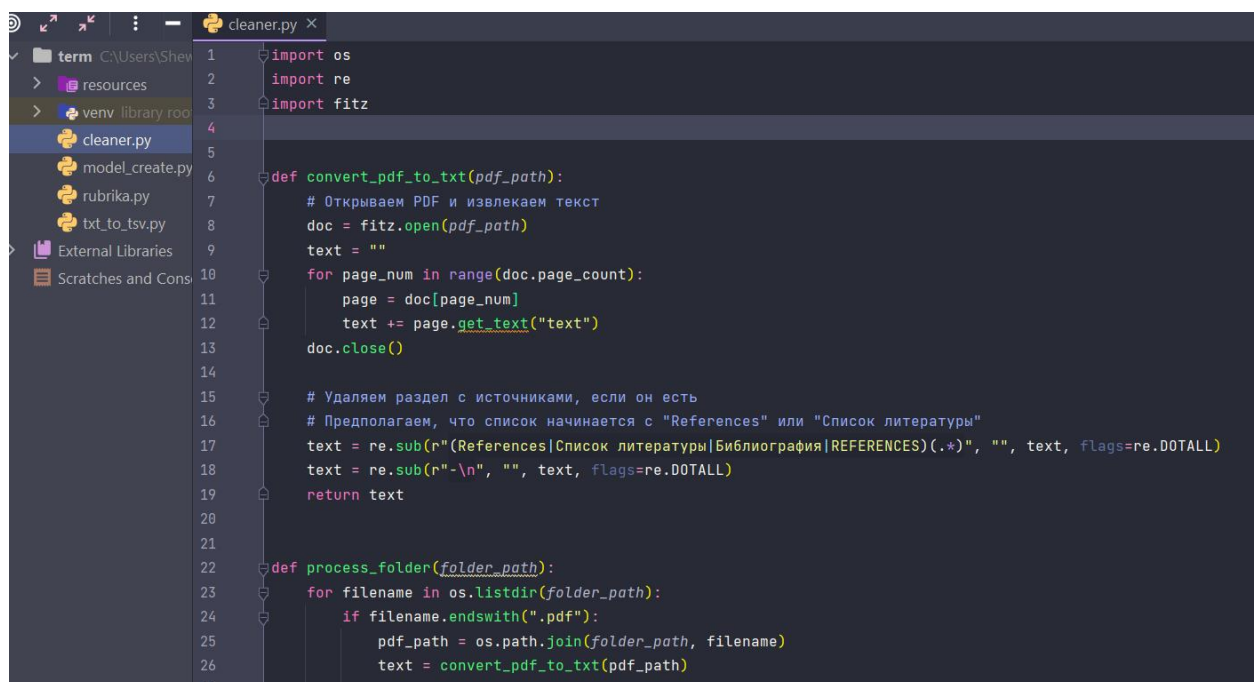
2 СОСТАВЛЕНИЕ ДАТАСЕТА ДЛЯ ДООБУЧЕНИЯ ВЕКТОРНОЙ МОДЕЛИ

Создание обучающего датасета состоит из обработки текстовых данных, с целью подведения их к формату, пригодному для использования в программных средствах. Для этого произвели следующие действия:

- скачаны все статьи;
- произведена предварительная обработка текстов для улучшения качества обучения с помощью файла очистки;
- совмещены тексты в один файл .TXT;
- конвертирован файл в формат .TSV.

Работа выполнялась отдельно по медицинской ПО и ПО для ВКР.

С помощью файла очистки статьи формата .PDF конвертировали в .TXT и удалили раздел «Список литературы» для того, чтобы сделать модель более качественной. Фрагмент файла очистки представлен на рисунке 2.



```

1  import os
2  import re
3  import fitz
4
5
6  def convert_pdf_to_txt(pdf_path):
7      # Открываем PDF и извлекаем текст
8      doc = fitz.open(pdf_path)
9      text = ""
10     for page_num in range(doc.page_count):
11         page = doc[page_num]
12         text += page.get_text("text")
13     doc.close()
14
15     # Удаляем раздел с источниками, если он есть
16     # Предполагаем, что список начинается с "References" или "Список литературы"
17     text = re.sub(r"(References|Список литературы|Библиография|REFERENCES)(.*)", "", text, flags=re.DOTALL)
18     text = re.sub(r"\n", "", text, flags=re.DOTALL)
19     return text
20
21
22 def process_folder(folder_path):
23     for filename in os.listdir(folder_path):
24         if filename.endswith(".pdf"):
25             pdf_path = os.path.join(folder_path, filename)
26             text = convert_pdf_to_txt(pdf_path)

```

Рисунок 2 – Фрагмент файла очистки

После этого тексты статей категории были объединены в один файл .TXT. Конвертация в формат .TSV выполнялась с помощью онлайн-конвертера aspose.app. Интерфейс программы конвертера представлен на рисунке 3. В результате были получены два файла: med.tsv, vkr.tsv.

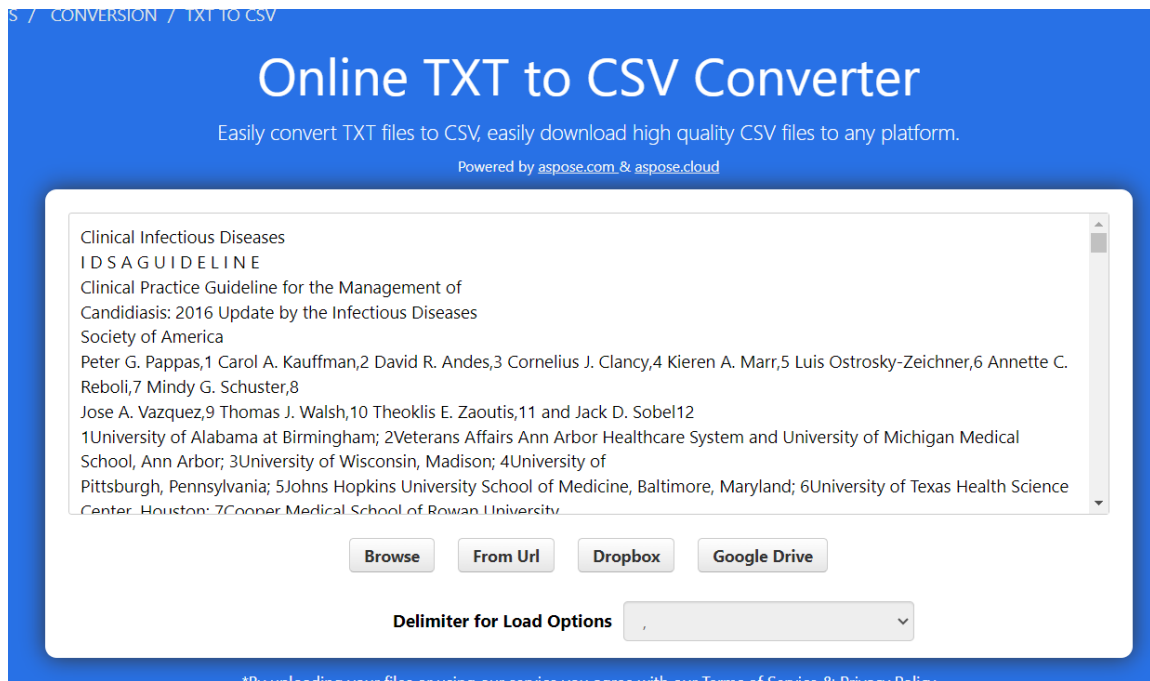
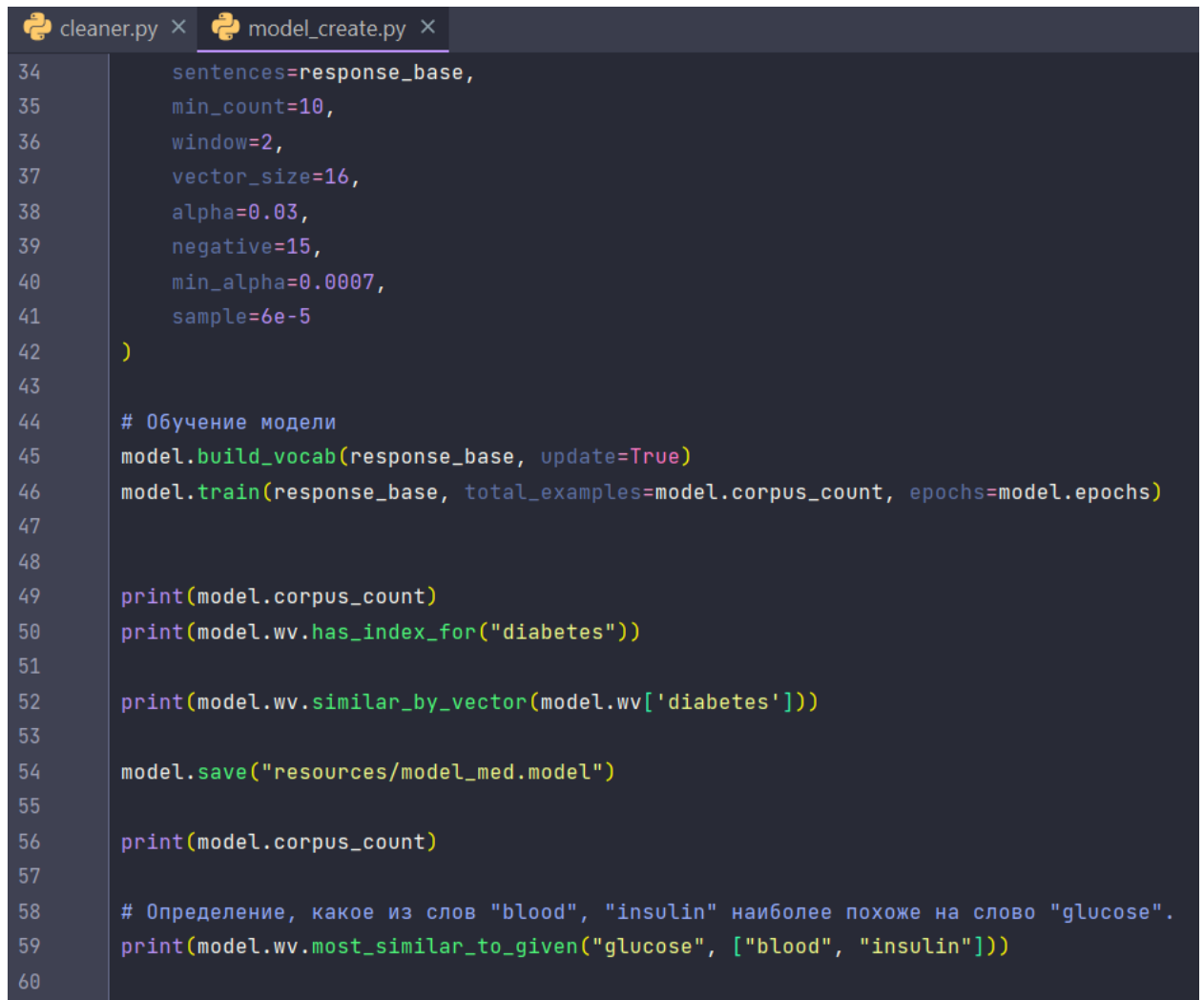


Рисунок 3 – Интерфейс онлайн программы-конвертера aspose.app

3 ОБУЧЕНИЕ МОДЕЛИ WORD2VEC

Для обучения моделей word2vec был получен файл шаблона у преподавателя. Для корректной работы шаблонного проекта были установлены библиотеки (genism, pandas), используя средства среды разработки. Далее был создан проект, в который импортирован файл шаблона и собранные датасеты. Настроили параметры для обучения модели и сохранили в проекте. Фрагмент кода для создания модели для медицинской тематики приведен на рисунке 4. Полный код можно увидеть в файле «model_create.py»



```
34     sentences=response_base,
35     min_count=10,
36     window=2,
37     vector_size=16,
38     alpha=0.03,
39     negative=15,
40     min_alpha=0.0007,
41     sample=6e-5
42 )
43
44 # Обучение модели
45 model.build_vocab(response_base, update=True)
46 model.train(response_base, total_examples=model.corpus_count, epochs=model.epochs)
47
48
49 print(model.corpus_count)
50 print(model.wv.has_index_for("diabetes"))
51
52 print(model.wv.similar_by_vector(model.wv['diabetes']))
53
54 model.save("resources/model_med.model")
55
56 print(model.corpus_count)
57
58 # Определение, какое из слов "blood", "insulin" наиболее похоже на слово "glucose".
59 print(model.wv.most_similar_to_given("glucose", ["blood", "insulin"]))
60
```

Рисунок 4 – Фрагмент кода создания модели для медицинской тематики

Для доступа к тексту ПО необходимо указать путь к файлу .TSV, а также формат кодирования, чтобы избежать ошибки при работе программы. Для проверки модели проверили наличие слова в модели, а также вывели близкие слова по вектору. И еще три слова для сравнения близости их векторных представлений.

В итоге получили две модели: model_med.model, model_vkr.model. Вывод программы для ПО представлен на рисунках 5, 6.

```

37 print(model.wv.has_index_for("diabetes"))
38
39 print(model.wv.similar_by_vector(model.wv['diabetes']))
40
41 model.save("resources/model_med.model")
42
43 print(model.corpus_count)
44 print(model.wv.most_similar_to_given("glucose", ["blood", "insulin"]))
45

```

Process finished with exit code 0

Рисунок 5 – Результат обучения model_med

```

36 print(model.corpus_count)
37 print(model.wv.has_index_for("здоровье"))
38
39 print(model.wv.similar_by_vector(model.wv['здоровье']))
40
41 model.save("resources/model_vkr.model")
42
43 print(model.corpus_count)
44 print(model.wv.most_similar_to_given("спорт", ["культура", "молодежь"]))
45

```

Process finished with exit code 0

Рисунок 6 – Результат обучения model_vkr

4 РУБРИЦИРОВАНИЕ ЕЯ ТЕКСТОВЫХ ДАННЫХ ОПИСАНИЯ ПО

Для проведения автоматического рубрицирования была написана программа, реализующая алгоритм из задания.

Для того, чтобы загрузить готовую модель, использовали конструкцию «w2v_model = Word2Vec.load("путь_к_файлу.model")».

Для проверки наличия слова в модели – «w2v_model.wv.has_index_for(word)».

Для получения нормализованного вектора по слову «w2v_model.wv.get_vector(word).sum()».

Алгоритм для рубрицирования:

- текст разбивается на массив слов;

Использовали line.split()

- для каждого слова в массиве был получен соответствующий ему нормализованный вектор из модели;

- для слов, вектор которых неизвестен, значение вектора считали равным 0;

Для этих пунктов проверили наличие слова в модели, затем получили его вектор, просуммировали и увеличили счетчик слов.

- найдено среднее арифметическое полученных векторов;

Таким образом каждому тексту выборки соответствует числовое значение вектора.

Далее было произведено рубрицирование:

- отсортировали значения векторов;

- используя отсортированные значения, разделили тексты на 2 категорий, основываясь на близости значений векторов (числовые характеристики, соответствующие темам текстов, должны иметь минимальное отклонение между собой относительно всего датасета);

- определили среднее арифметическое числового значения для каждой категории (это число – числовое значение категории);

- получили набор ближайших по вектору слов из созданной модели (эти слова являются условно-ключевыми);

Для получения ближайших слов использовали «model.wv.similar_by_vector(vector=np.array(your_word_vector), topn=5)», где второй параметр обозначает число получаемых слов. Код алгоритма приведен на рисунке 7.


```

rubrika.py x
12 for j in range(1, 21):
13     result = 0
14     words = 0
15     result1 = 0
16     with open(f"clean_result/med/text_{j}.txt", 'r', encoding="utf8") as file:
17         for line in file:
18             line = re.sub(patterns, ' ', line)
19             for word in line.split():
20                 if w2v_model.wv.has_index_for(word):
21                     result += w2v_model.wv.get_vector(word).sum()
22                     result1 += w2v_model.wv.get_vector(word)
23             words += 1
24     print(result/words, j)
25     array.append(result/words)
26     array1.append(result1/words)
27
28 sorted_array = sorted(array1, key=np.linalg.norm)
29
30 mid_index = len(sorted_array) // 2
31 first_half = sorted_array[:mid_index]
32 category1_vector = np.mean(first_half, axis=0)
33 second_half = sorted_array[mid_index:]
34 category2_vector = np.mean(second_half, axis=0)
35
36 print("Категория 1:", category1_vector.sum())
37 print("Категория 2:", category2_vector.sum())
38
39 print(w2v_model.wv.similar_by_vector(vector=category1_vector, topn=5))
40 print(w2v_model.wv.similar_by_vector(vector=category2_vector, topn=5))

```

Рисунок 7 – Алгоритм автоматического рубрицирования

Результат вывода для ПО представлен на рисунках 8,9.

```

C:\Users\Shewa\PycharmProjects\term\venv\Scripts\python.exe C:/Users/Shewa/PycharmProjects/term/rubrika.py
1.2867878565462655 1
1.0517785819794168 2
1.1194191124202528 3
0.9698850885088826 4
1.0086789170272343 5
1.1178840139529087 6
0.9516935409255248 7
1.3273923803759133 8
1.4159866583239038 9
1.2135951687001671 10
1.37478430225766 11
1.0197104430841533 12
1.1394853617351402 13
1.174124672457024 14
1.1766192985126396 15
1.2768447952652724 16
1.3838685251239489 17
0.86451420808975091 18
1.1588156508538308 19
1.4130978993591832 20
Категория 1: 1.0401857
Категория 2: 1.3043017
[('flucytosine', 0.998669445514679), ('history', 0.9986093640327454), ('receptor', 0.9985940456390381), ('performed', 0.9985873699188232), ('identified', 0.9985445737838745)]
[('flucytosine', 0.9987562298774719), ('performed', 0.9986156225204468), ('history', 0.9985690712928772), ('identified', 0.9985049366950989), ('identical', 0.9984852671623253)]

```

Рисунок 8 – Категории для медицинской ПО

```

C:\Users\Shewa\PycharmProjects\term\venv\Scripts\python.exe C:/Users/Shewa/PycharmProjects/term/rubrika.py
0.08664417178953192 1
0.13649479988647303 2
0.11631484888873372 3
0.11087843412844891 4
0.1172859813603902 5
0.13122163227754444 6
0.06748101873028439 7
0.09665056954980474 8
0.0893394037091788 9
0.10964769528186072 10
0.09879347757648549 11
0.13673874317762558 12
0.09438608724295217 13
0.12258782493276002 14
0.1286248070793816 15
0.13241013443702793 16
0.12241886970880149 17
0.1241873209175028 18
0.11510169251358336 19
0.11083848380793132 20
Категория 1: 0.098523706
Категория 2: 0.12627286
[('на', 0.9867351651191711), ('студентов', 0.9855296611785889), ('and', 0.9841821789741516), ('of', 0.9833813985715942), ('спорта', 0.980512261390686)]
[('and', 0.9859327673912048), ('на', 0.9853858947753906), ('of', 0.984419047832489), ('студентов', 0.9842018485069275), ('спорта', 0.9833254218101501)]

```

Рисунок 9 – Категории для ПО ВКР

Слова, найденные таким образом слишком похожи. Улучшить выдачу слов не получилось, самостоятельно выделили ключевые слова и названия категорий, основываясь на текстах, входящих в категории. Для ПО по ВКР получили подкатегории: «здоровье», «спорт». Для медицинской ПО выделили подкатегории: «diabetes», «candidiasis». Результаты в виде таблицы с распределением подкатегорий, ключевыми словами представлены в файле «Задание 4.xlsx». Фрагменты таблицы изображены на рисунках 10, 11.

	А	В	С	Д	Е
1	№	Название статьи	Числовое значение	Ключевые слова	Подкатегория
2	1	Здоровые студенты - здоровая нация	0,10	Игра, здоровье, ж	здоровье
3	2	К проблеме формирования здорового образа жизни подро	0,10	жизни,	здоровье
4	3	ЗДОРОВЫЙ ОБРАЗ ЖИЗНИ В СОВРЕМЕННОМ МИРЕ	0,10	здоровье, здоров	здоровье
5	4	ПИТАНИЕ КАК АКТУАЛЬНАЯ ПРОБЛЕМА ПОВЫШЕНИЯ ЭФФ	0,10	питание, спорт, з	здоровье
6	5	ТЕНДЕНЦИИ РАЗВИТИЯ ПРЕДСТАВЛЕНИЙ О ЗДОРОВОМ ОБ	0,10	ЗОЖ, молодежь, з	здоровье
7	6	ОТНОШЕНИЕ СТУДЕНТОВ К ФОРМИРОВАНИЮ ЗДОРОВОГО	0,10	образ жизни, здо	здоровье
8	7	Здоровье и здоровый образ жизни молодёжи	0,10	здоровье, здоров	здоровье
9	8	Отношение молодежи к информационно-коммуникацион	0,10	здоровье, здоров	здоровье
10	9	Здоровье студенческой молодежи:ценностные установки	0,10	здоровье, молод	здоровье
11	10	Здоровый образ жизни молодежи	0,10	здоровый образ	здоровье
12	11	ЦИФРОВЫЕ ТЕХНОЛОГИИ В СПОРТЕ	0,13	цифровые технол	спорт
13	12	ЦИФРОВЫЕ ТРЕНДЫ В СФЕРЕ ФИЗИЧЕСКОЙ КУЛЬТУРЫ И С	0,13	цифровая трансф	спорт
14	13	КАК ЦИФРОВЫЕ ТЕХНОЛОГИИ ВЛИЯЮТ НА СПОРТ?	0,13	спортивная инфо	спорт
15	14	ЦИФРОВЫЕ ТЕХНОЛОГИИ В СФЕРЕ ФИЗИЧЕСКОЙ КУЛЬТУРЫ	0,13	цифровые технол	спорт
16	15	ЦИФРОВОЙ ПОДХОД В ОРГАНИЗАЦИИ ФИЗИЧЕСКОЙ КУЛЬ	0,13	студенты, занятия	спорт
17	16	Тенденции и потенциал развития технологичных видов сп	0,13	высокотехнологи	спорт
18	17	ОСНОВНЫЕ НАПРАВЛЕНИЯ ЦИФРОВОЙ ТРАНСФОРМАЦИИ	0,13	цифровизация, ц	спорт
19	18	НОВЫЕ ЭЛЕКТРОННЫЕ И ЦИФРОВЫЕ СЕРВИСЫ ПО ФИЗИЧЕ	0,13	цифровые платфор	спорт
20	19	ЦИФРОВАЯ ОБРАЗОВАТЕЛЬНАЯ СРЕДА ПО ФИЗИЧЕСКОЙ КУ	0,13	информационны	спорт
21	20	ЦИФРОВЫЕ РЕШЕНИЯ АКТУАЛЬНЫХ ВОПРОСОВ В СФЕРЕ ФИ	0,13	физическая культ	спорт

Рисунок 10 – Категории для ПО ВКР

	A	B	C	D	E
1	№	Название статьи	Числовое значение	Ключевые слова	Подкатегория
2	1	Daily Fasting Blood Glucose Rhythm in Male Mice: A Role of the	1,040	Circadian Clock, bl	diabetes
3	2	CLOCK and BMAL1 regulate MyoD and are necessary for maintenanc	1,040	circadian clock m	diabetes
4	3	Circadian and feeding rhythms differentially affect rhythmic mRNA t	1,040	circadian rhythms	diabetes
5	4	Phototransduction by Retinal Ganglion Cells That Set the Circadian C	1,040	retinal ganglian ce	diabetes
6	5	Disruption of Circadian Insulin Secretion Is Associated With Reduced	1,040	Type 2 Diabetes, c	diabetes
7	6	Evidence for a Circadian Rhythm of Insulin Sensitivity in Patients Wit	1,040	Circadian Rhythm,	diabetes
8	7	Projection of the year 2050 burden of diabetes in the US adult popul	1,040	diabetes	diabetes
9	8	Beta-Cell Deficit and Increased Beta-Cell Apoptosis in Humans With	1,040	Type 2 Diabetes, B	diabetes
10	9	Adverse Metabolic Consequences in Humans of Prolonged Sleep Res	1,040	metabolic syndron	diabetes
11	10	Reciprocal Regulation of Brain and Muscle Arnt- Like Protein 1 and P	1,040	circadian rhythm, c	diabetes
12	11	Clinical Practice Guideline for the Management of Candidiasis: 2016	1,304	candidemia; invas	candidiasis
13	12	Candida ESOPHAGITIS: SPECIES DISTRIBUTION AND RISK FACTORS F	1,304	Candida; Opportun	candidiasis
14	13	CANDIDA ESOPHAGITIS A prospective study of 27 cases	1,304	CANDIDA ESOPHA	candidiasis
15	14	Candida Esophagitis in Achalasia Cardia: Case Report and Review of	1,304	Achalasia cardia, C	candidiasis
16	15	Candida Esophagitis: Feathery Appearance as a New Sign on Barium	1,304	Barium esophagog	candidiasis
17	16	An Elderly Case of Type 2 Diabetes which Developed in Association v	1,304	denture, candidias	candidiasis
18	17	Guidelines for Treatment of Candidiasis	1,304	Candidiasis	candidiasis
19	18	Long-Term Trends in Esophageal Candidiasis Prevalence and Associa	1,304	Esophageal Candid	candidiasis
20	19	Fatal Esophageal Perforation Caused by Invasive Candidiasis	1,304	Invasive Candidias	candidiasis
21	20	Surgical Management of Necrotizing Candida Esophagitis	1,304	Candida Esophagit	candidiasis

Рисунок 11 – Категории для медицинской ПО

5 СОСТАВЛЕНИЕ НАБОРА СПРАВОЧНЫХ ДАННЫХ, ОПИСЫВАЮЩИХ РУБРИКИ ИССЛЕДУЕМОЙ ПО

Для сбора справочных данных были произведены следующие действия для каждой из категорий (2 медицинские и 2 по теме ВКР):

- выбрано описание нескольких самых важных тем, фигурирующих в подкатегории;
- собраны описания вместе и сопоставлены их числовому значению подкатегории;

На основе полученных описаний была составлена таблица, содержащая поля: числовое значение категории, ключевые слова, справочная информация.

Для каждой из категорий выбраны 5 терминов / понятий / важных для понимания групп сведений, найдены для них определения. Составлена таблица, содержащая эти определения. Термины для «здоровье»: здоровый образ жизни, правильное питание, нормированный распорядок дня, ментальное здоровье, физическая активность. Термины для «спорт»: физкультурное образование, мониторинг здоровья, двигательная активность, спортивная информатика, спортивные технологии. Термины для «diabetes»: insulin, blood glucose, hyperglycemia, hyperglycemic clamps, metabolic syndrome. Термины для «candidiasis»: candidemia, yeasts, candida esophagitis, endoscopy, fungal infection.

Набор справочных данных для каждой из категории содержится в файле «Задание 5.xlsx». Также выделены отдельные таблицы с полями: термин, определение. Фрагменты таблиц представлены на рисунках 12, 13.

А		В	С
1	Числовой вектор категории	Ключевые слова	Справочная информация
2	0,10	здоровье, здоровый образ жизни, молодежь	правильное питание - организация режима и рациона питания, обеспечивающая организм всеми необходимыми питательными веществами в оптимальных количествах и соотношениях для поддержания здоровья, работоспособности и профилактики заболеваний.
3			здоровый образ жизни - комплексное поведение и привычки, направленные на поддержание и улучшение физического, психического и социального благополучия, включая рациональное питание, физическую активность, отказ от вредных привычек и соблюдение режима дня.
4			нормированный распорядок дня - планирование и распределение времени суток, включающее периоды работы, отдыха, сна, приема пищи и физической активности, с целью оптимизации работоспособности и общего самочувствия.
5			ментальное здоровье - состояние психического благополучия, при котором человек способен реализовывать свои способности, справляться с жизненными трудностями, эффективно работать и вносить вклад в общественную жизнь.
6			физическая активность - любые движения тела, выполняемые скелетными мышцами, которые требуют расхода энергии и включают разнообразные виды активности, такие как ходьба, бег, плавание и упражнения.
7	0,13	спорт, физическая культура, тренировка	Физкультурное образование - процесс обучения и воспитания, направленный на формирование знаний, умений и навыков в области физической культуры, развитие физических качеств и воспитание устойчивой потребности в здоровом образе жизни.
8			Мониторинг здоровья - регулярное наблюдение за состоянием организма с использованием медицинских, технологических и самоконтрольных методов для своевременного выявления изменений, профилактики и лечения заболеваний.
9			Двигательная активность - любые целенаправленные или спонтанные движения, выполняемые человеком в течение дня, которые способствуют укреплению здоровья, повышению выносливости и работоспособности.
			Спортивная информатика - научная и прикладная дисциплина, изучающая методы и средства использования информационных технологий и систем в области спорта для анализа, планирования и оптимизации тренировочного

Рисунок 12 – Справочная информация по теме ВКР

	A	B
1	Термин	Определение
2	Insulin	A peptide hormone produced by the beta cells of the pancreas, essential for regulating blood glucose levels. Insulin facilitates the uptake of glucose by cells for energy production or storage as glycogen, and it plays a critical role in lipid and protein metabolism.
3	Blood glucose	The amount of glucose present in the bloodstream, typically measured in milligrams per deciliter (mg/dL). Normal fasting blood glucose levels range from 70 to 99 mg/dL, and glucose serves as a primary energy source for cellular processes, especially in the brain and muscles.
4	Hyperglycemia	A condition where blood glucose levels are higher than normal, typically above 125 mg/dL when fasting. Chronic hyperglycemia is a hallmark of diabetes and can lead to complications such as neuropathy, retinopathy, and cardiovascular diseases.
5	Hyperglycemic clamps	A precise experimental technique used in metabolic research to assess beta-cell function and insulin secretion. It involves infusing glucose intravenously to maintain a stable elevated blood glucose concentration, while measuring the amount of insulin the pancreas releases in response.
6	Metabolic syndrome	A combination of metabolic disorders, including central obesity, insulin resistance, dyslipidemia (elevated triglycerides and reduced HDL cholesterol), high blood pressure, and elevated fasting blood glucose. This syndrome significantly increases the risk of developing type 2 diabetes, cardiovascular disease, and stroke.
	Candidemia	A severe bloodstream infection caused by Candida species, which are opportunistic pathogens. Candidemia can result from invasive medical procedures, the use of central venous catheters, or

Рисунок 13 – Термины по медицинской теме

6 СОЗДАНИЕ АВТОМАТИЧЕСКОЙ СПРАВОЧНОЙ СИСТЕМЫ

Для создания автоматической справочной системы преобразовали таблицу, полученную в Задании 5 в базу данных в виде словаря.

Справочная система представлена в виде бд с ключом – термин, а значение – определение. Формат базы данных приведен на рисунке 14.

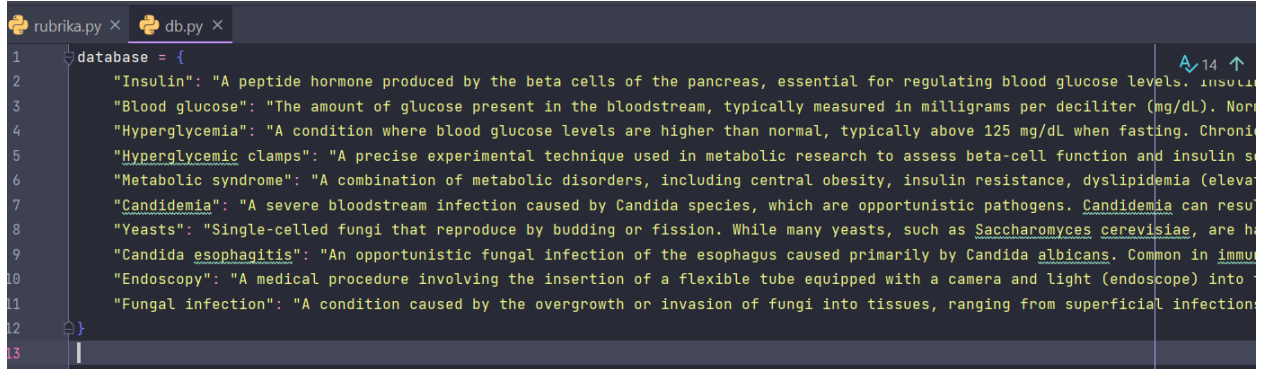


Рисунок 14 – База данных справочной системы

Программа работает следующим образом:

1. На вход поступает запрос на тему определений.
2. Запрос преобразуется в массив слов, а затем для каждого слова находится нормализованный вектор из подгруженной модели.
3. Собирается массив из векторов запроса и находится среднее арифметическое. Затем в цикле рассчитывается средний вектор для определения и сравнивается с вектором запроса путем косинусного сходство для определения наиболее близкого совпадения.

Код справочной системы представлен на рисунке 15.

```

49     closest_key = None
50     closest_similar = -1
51     similar = -1
52
53     for key, description in database.items():
54         key_vector = vector_num(key)
55         if (len(key_vector) > 0) & (len(average_vector) > 0):
56             similar = cosine_similarity([average_vector], [key_vector])[0][0]
57             if similar > closest_similar:
58                 closest_similar = similar
59                 closest_key = key
60
61         description_vector = vector_num(description)
62         if (len(description_vector) > 0) & (len(average_vector) > 0):
63             similar = cosine_similarity([average_vector], [description_vector])[0][0]
64             if similar > closest_similar:
65                 closest_similar = similar
66                 closest_key = key
67
68     if closest_key is not None:
69         print("Описание:", database[closest_key])
70     else:
71         print("Описание не найдено.")
72
73     user_query = input("Введите ваш запрос: ")
74     directory_query(user_query)

```

Рисунок 15 – Код справочной системы

Результат работы программы представлен на рисунке 16.

```
C:\Users\Shewa\PycharmProjects\term\venv\Scripts\python.exe C:/Users/Shewa/PycharmProjects/term/task_6_vkr.py
Введите ваш запрос: что такое здоровый образ жизни
Описание: комплексное поведение и привычки, направленные на поддержание и улучшение физического, психического и социального благополучия, включая рациональное питание, физич

Process finished with exit code 0
|

C:\Users\Shewa\PycharmProjects\term\venv\Scripts\python.exe C:/Users/Shewa/PycharmProjects/term/task_6_med.py
Введите ваш запрос: what is a fungal infection
Описание: A condition caused by the overgrowth or invasion of fungi into tissues, ranging from superficial infections to systemic infections.

Process finished with exit code 0

C:\Users\Shewa\PycharmProjects\term\venv\Scripts\python.exe C:/Users/Shewa/PycharmProjects/term/task_6_med.py
Введите ваш запрос: what are hyperglycemic clamps used for?
Описание: A precise experimental technique used in metabolic research to assess beta-cell function and insulin secretion.

Process finished with exit code 0
|

C:\Users\Shewa\PycharmProjects\term\venv\Scripts\python.exe C:/Users/Shewa/PycharmProjects/term/task_6_vkr.py
Введите ваш запрос: С помощью чего измеряется мониторинг здоровья
Описание: регулярное наблюдение за состоянием организма с использованием медицинских, технологических и самоконтрольных методов для своевременного выявления изменений, профи

Process finished with exit code 0

C:\Users\Shewa\PycharmProjects\term\venv\Scripts\python.exe C:/Users/Shewa/PycharmProjects/term/task_6_vkr.py
Введите ваш запрос: Из чего состоит физическая активность
Описание: любые движения тела, выполняемые скелетными мышцами, которые требуют расхода энергии и включают разнообразные виды активности, такие

Process finished with exit code 0
```

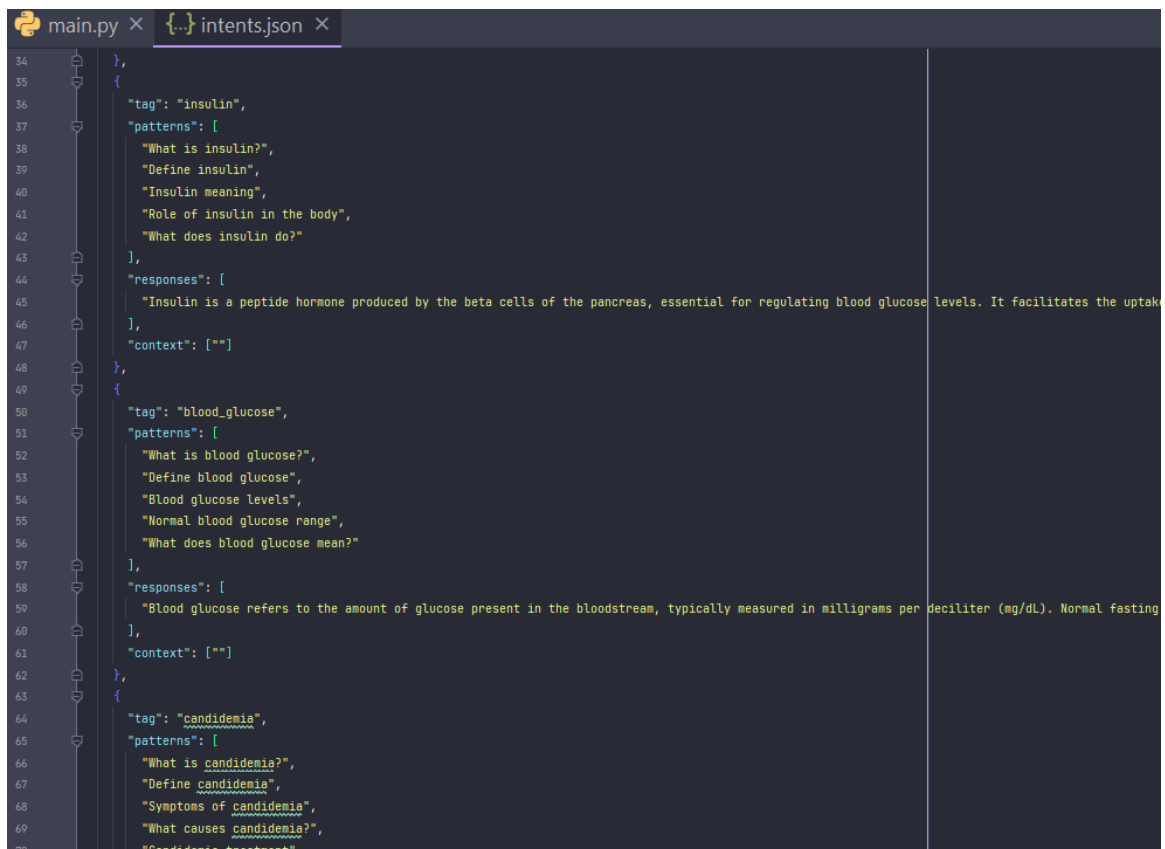
Рисунок 16 – Работа справочной системы

7 СОЗДАНИЕ АВТОМАТИЧЕСКОЙ ДИАЛОГОВОЙ СИСТЕМЫ

Для создания автоматической диалоговой системы был получен шаблон. Шаблон готов к использованию. Рядом с файлом кода необходимо было расположить файл `intents.json`. Данный файл заполняется на основе наполнения подкатегорий предметных областей.

Были созданы четыре intents для каждой из подкатегорий (по 2 для медицинской тематики и по 2 для тематики ВКР). В них были указаны в `patterns` различные формулировки вопросов, а в `responses` несколько вариантов ответа чат-бота. После завершения обучения система должна отвечать на контекстуально близкие запросы заготовленными ответами.

Структура файла `intents.json` с заполненными вопросами и ответами для подкатегорий представлена на рисунке 17.

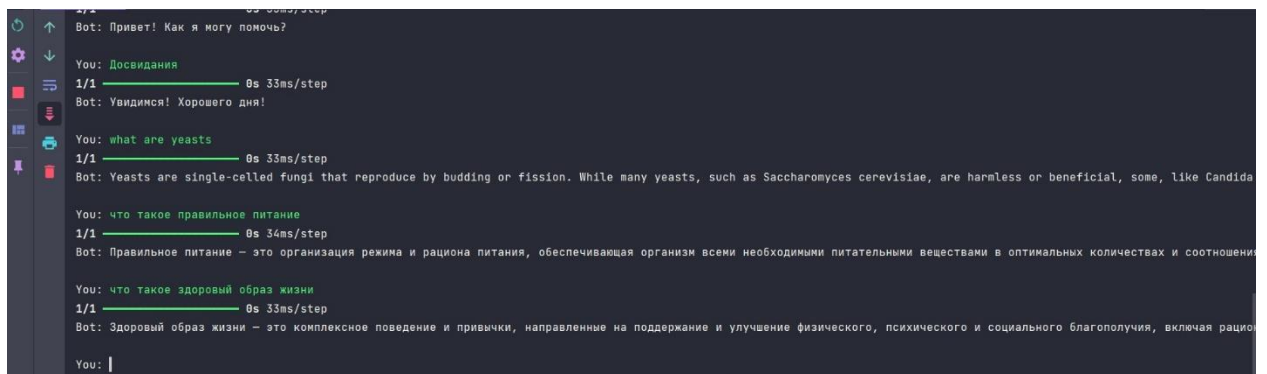


```

34  },
35  {
36    "tag": "insulin",
37    "patterns": [
38      "What is insulin?",
39      "Define insulin",
40      "Insulin meaning",
41      "Role of insulin in the body",
42      "What does insulin do?"
43    ],
44    "responses": [
45      "Insulin is a peptide hormone produced by the beta cells of the pancreas, essential for regulating blood glucose levels. It facilitates the uptake
46    ],
47    "context": [""]
48  },
49  {
50    "tag": "blood_glucose",
51    "patterns": [
52      "What is blood glucose?",
53      "Define blood glucose",
54      "Blood glucose levels",
55      "Normal blood glucose range",
56      "What does blood glucose mean?"
57    ],
58    "responses": [
59      "Blood glucose refers to the amount of glucose present in the bloodstream, typically measured in milligrams per deciliter (mg/dL). Normal fasting
60    ],
61    "context": [""]
62  },
63  {
64    "tag": "candidemia",
65    "patterns": [
66      "What is candidemia?",
67      "Define candidemia",
68      "Symptoms of candidemia",
69      "What causes candidemia?",
70      "Candidemia treatment"
  
```

Рисунок 17 – Структура файла `intents.json`

Система была обучена и протестирована. Результат представлен на рисунке 18.



```

272  Bot: Привет! Как я могу помочь?

You: Досвидания
1/1 ██████████ 0s 33ms/step
Bot: Увидимся! Хорошего дня!

You: what are yeasts
1/1 ██████████ 0s 33ms/step
Bot: Yeasts are single-celled fungi that reproduce by budding or fission. While many yeasts, such as Saccharomyces cerevisiae, are harmless or beneficial, some, like Candida

You: что такое правильное питание
1/1 ██████████ 0s 34ms/step
Bot: Правильное питание – это организация режима и рациона питания, обеспечивающая организм всеми необходимыми питательными веществами в оптимальных количествах и соотношении

You: что такое здоровый образ жизни
1/1 ██████████ 0s 33ms/step
Bot: Здоровый образ жизни – это комплексное поведение и привычки, направленные на поддержание и улучшение физического, психического и социального благополучия, включая рацион

You: |
  
```

Рисунок 18 – Работа диалоговой системы

ТЕХНОЛОГИЯ ПРОВЕДЕНИЯ ИССЛЕДОВАНИЯ

При проведении данного исследования на его этапах активно использовались функциональные возможности программы Microsoft Excel. Для составления электронных таблиц, содержащих данные до и после анализа.

Для поиска статей медицинской тематики использовался SciArticleBot. Для поиска статей по тематике ВКР использовался сайт электронной научной библиотеки cyberleninka.

Для автоматизации рубрицирования, обучения модели, создания автоматических справочной и диалоговой систем использовали код на языке Python. Среда разработки – PyCharm Community.

Основа исследования – работа с моделью на основе нейронных сетей word2vec. Активное использование библиотек gensim, pandas. Для вычисления косинусного сходства в справочной системе использовалась библиотека scikit-learn. Библиотеки nltk и tensorflow были необходим для создания чат-бота, также модуль pickle для сериализации объектов.

Помимо данных автоматизированных технологий применялись и ручные, такие как изучение текста статьи и отдельных его фрагментов, выявление основной смысловой составляющей, понимание предметной области и её естественно-языковых особенностей.

ЗАКЛЮЧЕНИЕ

Основными результатами проведенного автоматического рубрицирование текстов на основе word2vec явились следующие:

1. Собраны ЕЯ текстовые описания ПО.
2. Составлен датасет для дообучения векторной модели.
3. Обучена модель word2vec с помощью собранных данных.
4. Произведено рубрицирование ЕЯ текстовых данных описания ПО.
5. Составлен набор справочных данных, описывающих рубрики, исследуемой ПО.
6. Создана автоматическая справочная система.
7. Создана автоматическая диалоговая система, которая в будущем может быть интегрирована в систему ВКР.

ЛИТЕРАТУРА

1. ГОСТ 7.32-2017. Система стандартов по информации, библиотечному и издательскому делу. Отчет о научно-исследовательской работе. Структура и правила оформления. – [Электронный ресурс] – URL: <https://docs.cntd.ru/document/1200157208> (дата обращения: 2.12.2024).
2. Курс LMS «Введение в интеллектуальные диалоговые системы». – [Электронный ресурс] – URL: <https://online.mospolytech.ru/course/view.php?id=12946> (дата обращения: 2.12.2024).
3. Word2Vec: как работать с векторными представлениями слов. – [Электронный ресурс] – URL: <https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlja-mashinnogo-obuchenija/> (дата обращения: 2.12.2024).
4. Natural Language Toolkit. Documentation. – URL: <https://www.nltk.org/> (дата обращения: 2.12.2024).

ПРИЛОЖЕНИЯ

Результат сбора ЕЯ текстового описания ПО находится в файле «Задание 1.xlsx», который прикреплен вместе с отчетом.

Статьи ПО прикреплены в виде папки `clean_result` вместе с отчетом.

Обученные модели `model_med.model`, `model_vkr.model` прикреплены вместе с отчетом в папке «models».

Проект из PyCharm Community со всеми вложениями представлен в папке «проект».

Результат автоматического рубрицирования представлен в файле «Задание 4.xlsx», который прикреплен вместе с отчетом.

Справочные данные по подкатегориям хранятся в файле «Задание 5.xlsx», который прикреплен вместе с отчетом.