МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра Теоретической и прикладной информатики
(полное название кафедры)

УТВЕРЖДАЮ

Зав. кафедрой Чубич В.М.
(фамилия, имя, отчество)

_____
(подпись)

«_____»_____ 2021 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Ждановой Любови Олеговны
(фамилия, имя, отчество студента – автора работы)

Исследование функций расстояния между строками для объединения источников статистических данных
(тема работы)

Факультет Прикладной математики и информатики
(полное название факультета)

Направление подготовки: 02.03.03. Математическое обеспечение и администрирование информационных систем
(код и наименование направления подготовки бакалавра)

Руководитель от НГТУ	**Автор выпускной квалификационной работы**
Тимофеева А.Ю. (фамилия, имя, отчество)	Жданова Л.О. (фамилия, И.О.)
к. э. н. (ученая степень, ученое звание)	ФПМИ, ПМИ-92 (факультет, группа)
.....	.....

(подпись, дата) (подпись, дата)

Новосибирск, 2023 г.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра Теоретической и прикладной информатики
(полное название кафедры)

УТВЕРЖДАЮ

Зав. кафедрой Чубич В.М.
(фамилия, имя, отчество)

«15» марта 2023 г.

(подпись)

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ БАКАЛАВРА

студентке ***Ждановой Любови Олеговне***
(фамилия, имя, отчество студента)

Направление подготовки *02.03.03. Математическое обеспечение и администрирование информационных систем*

Факультет Прикладной математики и информатики

Тема ***Исследование функций расстояния между строками для объединения источников статистических данных***

Исходные данные (или цель работы):

Изучение алгоритмов вычисления различных строковых метрик для нахождения похожих строк с одинаковыми показателями из различных баз данных. Сравнение алгоритмов и выделение наиболее эффективного. Реализация пользовательского приложения для нахождения похожих строк из двух файлов.

Структурные части работы:

1.Изучение теории о алгоритмах вычисления различных строковых метрик.

2.Реализация приложений для определения похожих строк по различным алгоритмам.

3. Сравнение результатов.

4. Выделение наиболее эффективных алгоритмов.

Задание согласовано и принято к исполнению.

Руководитель от НГТУ	**Студент**
Тимофеева А.Ю.	*Жданова Л.О.*
(фамилия, имя, отчество)	(фамилия, имя, отчество)
к. э. н.	*ФПМИ, ПМИ-92*
(ученая степень, ученое звание)	(факультет, группа)
22.03.2023 г.	*22.03.2023 г.*
(подпись, дата)	(подпись, дата)

Тема утверждена приказом по НГТУ № *1248/2* от «*15*» *марта* *2023* г.

ВКР сдана в ГЭК № _____, тема сверена с данными приказа

15 марта 2023 г.
(подпись секретаря экзаменационной комиссии по защите ВКР, дата)

Филиппова Е.В.
(фамилия, имя, отчество секретаря экзаменационной комиссии по защите ВК

АННОТАЦИЯ

Отчет 40 с., 7 ч., 4 рис., 6 табл., 5 источников, 1 прил.

МЕТРИКА РАССТОЯНИЯ, АЛГОРИТМ, СТРОКА, ЖАККАР, КОЭФФИЦИЕНТ, КОСИНУСНОЕ РАССТОЯНИЕ, ЕВКЛИДОВОЕ РАССТОЯНИЯ, МАНХЭТЕВСКОЕ РАССТОЯНИЕ

Объектом исследования являются алгоритмы вычисления строковых метрик для определения сходства строк.

Цель работы заключается в изучении различных методов определения сходства строк и их применении при анализе текстовых данных.

В процессе работы были рассмотрены основные типы строковых метрик, такие как коэффициент Жаккара, косинусное расстояние и другие. Были изучены принципы и методы работы каждой метрики, а также их применимость в различных сферах.

В результате исследований была проведена аналитическая работа по описанию различных строковых метрик и создано пользовательское приложение для поиска похожих строк на основе более точных метрик.

Благодаря выбранным алгоритмам успешно были найдены 82% строк с одинаковыми показателями из базы данных по Скандинавии и России.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время огромную роль играет обработка и анализ текстовой информации, особенно в условиях быстрого развития

информационных технологий. Одной из задач в этой области является нахождение похожих строк в больших массивах данных, таких как текстовые документы или код программ. Данная работа посвящена исследованию строковых метрик и разработке алгоритмов для нахождения похожих строк. В работе проводится анализ различных методов измерения сходства строк и их применимости в различных сферах. Также описывается разработка и реализация программного комплекса для решения задачи нахождения похожих строк на примере текстовых документов.

Реализованная мной программа позволяет отыскать похожие строки, используя алгоритмы строковых метрик.

1 ПОСТАНОВКА ЗАДАЧИ

В ходе анализа и сопоставления динамики социально-экономических показателей по странам Скандинавии и регионам Сибири возникли сложности по сбору сходной статистической информации по этим двум мегарегионам. Для этого необходимо было произвести сопоставление

названий показателей из двух статистических баз. Данные представлены в виде файлов:

1. Файл EMISS_indices_all - показатели из российской базы ЕМИСС.

2. Файл Nordic_stat_indices - показатели из базы по Скандинавии.

Первые строки этих файлов можно посмотреть на рисунке 1 и рисунке 2, соответственно.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		x											
2	1	Наличие сельскохозяйственной техники (окончательные итоги)											
3	2	Общая земельная площадь (окончательные итоги)											
4	3	Площади многолетних плодовых насаждений и ягодных культур (окончательные итоги)											
5	4	Посевная площадь по видам сельскохозяйственных культур (окончательные итоги)											
6	5	Средний размер хозяйств по общей площади и сельскохозяйственных угодий, посевной площади (окончательные итоги)											
7	6	Средний размер хозяйств по поголовью скота (окончательные итоги)											
8	7	Средний размер хозяйств по численности работников (окончательные итоги)											
9	8	Структура многолетних плодовых насаждений и ягодных культур (окончательные итоги)											
10	9	Структура посевных площадей по видам сельскохозяйственных культур (окончательные итоги)											
11	10	Численность работников, занятых в организациях (хозяйствах) (окончательные итоги)											
12	11	Число организаций (хозяйств) (окончательные итоги)											
13	12	Уровень грамотности взрослого населения											
14	13	Численность населения по итогам Всероссийской переписи населения											

Рисунок 1 - Файл EMISS_indices_all

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		eng	rus												
2	1	AGRI10: A	AGRI10: Пахотные земли и постоянные пастбища с разбивкой по стране, площади, культурам и времени.												
3	2	AGRI11: C	AGRI11: Растениеводство с пахотных земель в разбивке по странам, культурам и времени												
4	3	AGRI12: C	AGRI12: Урожайность с убранной площади в разбивке по стране, культурам и времени												
5	4	OEKO01: (	OEKO01: Органическое сельское хозяйство по стране, единице измерения и времени												
6	5	FISH11: Fi	FISH11: Рыболовство, годовой улов по странам, представившим информацию, видам, единицам измерения и времени.												
7	6	FISH12: Fi	FISH12: Рыболовство, годовой улов по странам, представляющим отчетность, районам промысла, видам и времени.												
8	7	FISH13: Ac	FISH13: Аквакультура, производство по странам, представляющим информацию, видам, единицам измерения и времени												
9	8	FISH14: Fi	FISH14: Рыболовный флот по стране, представившей отчет, длина, возраст, валовая вместимость, единица измерения и время.												
10	9	FORE07: P	FORE07: Производство древесины, 1000 кубических метров или 1000 тонн, по стране, продукту, единице измерения и времени.												
11	10	FORE08: F	FORE08: Рубки леса и прирост по странам, представившим информацию, содержанию и времени												
12	11	ENTP01: S	ENTP01: Структурная статистика предприятий по видам деятельности, стране отчетности, индикатору и времени												
13	12	ENTP03: N	ENTP03: Создание новых предприятий, индекс (1 кв. 2013 г. = 100) по странам и периодам времени												

Рисунок 2 – Файл Nordic_stat_indices

В этих файлах отражены цены на продукцию (затраты, услуги) инвестиционного назначения с 2017 года, потребительские цены на товары и услуги, оперативные данные площади жилых домов, средние потребительские цены (тарифы) на товары и услуги.

В рамках выпускной квалификационной работы требуется подобрать метрики расстояния для поиска одинаковых показателей из двух баз данных. Также необходимо сравнить используемые метрики и выделить наиболее эффективные. На основе лучших метрик написать программу пользовательского приложения.

2 АЛГОРИТМЫ РЕШЕНИЯ

2.1 Выбор алгоритмов решения

В своей дипломной работе я буду исследовать алгоритм Жаккара, косинусное, евклидовое и манхэттенское расстояния, поскольку эти метрики являются наиболее распространёнными, благодаря своей быстрой работе и эффективности по нахождению похожих строк.

2.2 Описание алгоритмов

2.2.1 Алгоритм Жаккара

Алгоритм Жаккара – это математический метод нахождения коэффициента сходства между двумя или более множествами. Он основывается на вычислении отношения пересечения (рисунок 3).

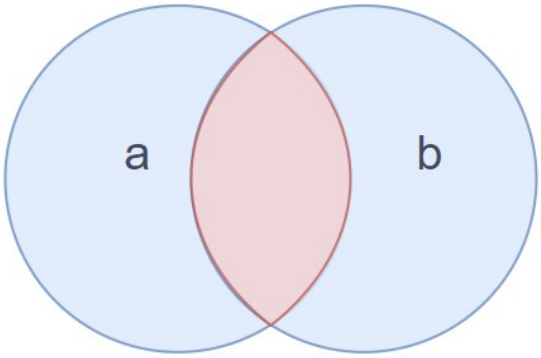

Рисунок 3 – Отношение пересечения пространств

Расчет коэффициента производится по формуле:

.

Алгоритм Жаккара находит широкое применение во многих областях, например:

- в биоинформатике для сравнения последовательностей нуклеотидов и аминокислот;
- в поисковых системах для определения релевантности поискового запроса;
- в социологии для анализа социальных сетей;
- в криптографии для проверки идентичности документов и файлов.

Ниже перечислены некоторые плюсы и минусы использования алгоритма Жаккара.

Плюсы:

- простота реализации;
- высокая скорость работы;
- позволяет быстро определить степень сходства между двумя наборами данных;

- может применяться для решения широкого спектра задач, таких как классификация, поиск информации и других.

Минусы:

- не всегда точен, поскольку не учитывает важность элементов множества и их порядок;
- не учитывает контекст и смысл элементов множества;
- может дать неправильные результаты в случае большого количества элементов.

2.2.2 Алгоритм косинусного расстояния

Косинусное расстояние – это мера сходства между двумя множествами, которая определяется с помощью косинуса угла между ними в n-мерном пространстве. Чтобы рассчитать косинусное расстояние нам необходимо представить наши строки в виде векторов, а затем произвести расчет по знакомой формуле:

,

где и – вектор первой и второй строки, соответственно.

Плюсы косинусного расстояния:

- широко используется в информационном поиске, обработке естественного языка, машинном обучении, тематическом моделировании и других областях;
- один из наиболее эффективных подходов для выявления похожих элементов в больших наборах данных, так как он игнорирует масштаб и фоновое распределение и фокусируется только на направлении векторов;
- косинусное расстояние может работать с разнородными данными, такими как текст, многомерные массивы или фотографии.

Минусы косинусного расстояния:

- оно основано на предположении, что объекты представлены в виде векторов в n-мерном пространстве, и не учитывает контекст и неявные связи между объектами;

– не учитывает порядок элементов в множествах, что может быть нежелательным для некоторых задач;

- результаты косинусного расстояния могут быть не интерпретируемыми, особенно если векторы значительно отличаются по длине.

Косинусное расстояние используется в различных задачах, таких как:

- рекомендательные системы для поиска похожих по вкусовым предпочтениям пользователей;

- кластеризация текстов для группирования по тематикам;

- определение сходства между изображениями в задачах компьютерного зрения;

- оценка сходства траекторий и маршрутов в задачах мобильной локализации и анализа транспортных потоков.

2.2.3 Алгоритм Евклидового расстояния

Евклидово расстояние - это метрика, используемая для измерения расстояния между двумя множествами в многомерном пространстве. Оно также может использоваться для определения сходства между наборами данных, что делает его полезным в анализе данных и машинном обучении.

Для работы этого алгоритма нам также понадобиться превратить строки в вектора, а затем по формуле рассчитать схожесть пар строк.

.

Чем выше значение, тем выше сходство.

Евклидово расстояние используется в:

- кластерном анализе;

- машинном обучении;

- анализе изображений;

- распознавании образов;
- анализе данных социальных сетей.

2.2.3 Алгоритм манхэттенского расстояния

Манхэттенское расстояние (также называемое "таксикабовское расстояние" или "расстояние L") - это метрика, которая измеряет расстояние между двумя точками на координатной плоскости. В случае строк, это можно использовать для определения степени похожести между ними.

Для определения манхэттенского расстояния между двумя строками, нужно посчитать сумму модулей разниц между соседними символами в каждой:

,

где a[i] и b[i] - символы строк a и b на позиции i.

Плюсы использования манхэттенского расстояния:

- пространственная эффективность: в отличие от других метрик расстояний, таких как евклидово расстояние, манхэттенское расстояние не требует вычисления корней или степеней;
- хорошее соответствие человеческому мнению: манхэттенское расстояние часто используется для измерения физических расстояний в городских районах, где важно учитывать препятствия на пути (например, перекрестки, углы домов). В этом смысле, оно ближе к тому, как человек оценивает расстояния.

Минусы использования манхэттенского расстояния:

- не учитывает длину дуги: если символы находятся на одной горизонтальной или вертикальной линии, то манхэттенское расстояние считает, что расстояние между ними составляет одну единицу, даже если фактически расстояние между ними больше;
- не учитывает порядок символов: манхэттенское расстояние не учитывает порядок символов при сравнении строк, поэтому две строки, которые

отличаются только порядком символов, будут считаться максимально отличными друг от друга.

Манхэттенское расстояние используется:

- в компьютерной графике: для обнаружения коллизий между объектами, отображаемыми на экране;
- в машинном обучении: как метрика расстояний в кластеризации данных;
- в анализе текста: может быть использовано для сравнения двух строк и оценки их похожести.

2.3 Описание алгоритмов для программной реализации

2.3.1 Сценерий работы программы по алгоритму Жаккара

Для того, чтобы определить одинаковые показатели из двух файлов необходимо работать по следующему алгоритму:

1) в программу загружаются два файла. Все символы текста приводятся к нижнему регистру, удаляются знаки пунктуации, различные небуквенные символы и цифры;

2)также перед работой алгоритма Жаккара стоит разбить наши слова на n-граммы, поскольку слова в предложениях могут содержать разное окончание или префикс. Я в своем примере буду использовать триграммы;

3)для каждой строки из первого файла сопоставлять каждую строку из второго файла;

4)для каждой пары таких строк необходимо применять алгоритм сходства Жаккара;

5) вывести похожие строки и их коэффициент в файл с результатом.

2.3.2 Сценарий работы программы с использованием алгоритма косинусного, Евклидового и манхэттенского расстояния

Для подсчета косинусного расстояния мы работаем по похожему принципу:

1) в программу загружаются два файла. Все символы текста приводятся к нижнему регистру, удаляются знаки пунктуации, различные небуквенные символы и цифры;

2) также перед работой алгоритма разбиваем наши слова на триграммы;

3) формируем веса по алгоритму TF*IDF для всех строк из набора и преобразовываем их в вектора;

4) для каждой строки из первого файла сопоставлять каждую строку из второго файла;

5) для каждой пары таких строк необходимо посчитывать косинусное, Евклидовое и манхэттенское расстояние.

3 ПРЕДВАРИТЕЛЬНАЯ ПОДГОТОВКА ИСХОДНЫХ ДАННЫХ

Для того, чтобы работать с данными из файла nordic_stat_indices, все строки были переведены на русский язык и частично интерпретированы. Были удалены некоторые незначащие обороты, такие как “с разбивкой по стране, площади, культурам и времени ”.

Для работы алгоритмов косинусного, Евклидового и манхэттенского расстояния необходимо превратить наши строки в вектора. Мы можем этого добиться при помощи алгоритма IF*IDF:

,

где q – количество данной триграммы в последовательности, t – количество триграмм в последовательности.

,

где N-количество последовательностей, N(q)-количество триграммы во всех последовательностях.

Переход к векторизации показано в формуле (8):

4 РЕЗУЛЬТАТЫ РАБОТЫ ПРОГРАММ

4.1 Работа программы по алгоритму Жаккара

Для проверки и сравнения работы алгоритмов был использован файл, в котором содержались 97 пар верно сопоставленных строк.

К неверно найденным строкам относится, например, “Дети в детских садах - Количество абонентских станций (устройств), подключённых к сетям

подвижной радиотелефонной связи в стандарте GSM ”, поскольку алгоритм нашел в них одинаковые триграммы: “тск”, “ски”, “ких”.

Верное сопоставление должно было выглядеть так: “Дети в детских садах - численность детей, посещающих дошкольные образовательные учреждения”, но в нем только лишь одна общая триграмма “дет”.

Если бы мы заменили строку “численность детей, посещающих дошкольные образовательные учреждения” на “ численность детей, посещающих детские сады”, алгоритм сопоставил их верно.

Можно сделать вывод, что алгоритм хорошо справляется с поиском похожих строк. Но, как мы и убедились, высокий коэффициент не всегда означает, что строки верно сопоставлены по смыслу.

В результате работы выяснилось, что по алгоритму Жаккара верно сопоставилась 71 пара строк, что составляет примерно 73% от необходимого объёма (см. таблица 1).

Таблица 1 – Верно найденные сопоставления по алгоритму Жаккара

№	Строки из файла EMISS	Строки из файла Nordic_stat	Коэффициент Жаккара
1	Растениеводство с пахотных земель	Производство основных видов продукции в натуральном выражении	0.0909
2	Урожайность с убранной площади	Урожайность сельскохозяйственных культур (в расчете на убранную площадь)	0.2500
3	Органическое сельское хозяйств	Внесено сельскохозяйственными организациями органических удобрений на 1 га посева	0.2778
4	Аквакультура, производство	Производство (выращивание) товарной рыбы и других объектов товарного рыбоводства (аквакультуры)	0.2500

Продолжение таблицы 1

5	Структурная статистика предприятий	Количество предприятий и организаций	0.1607
6	Кровати в коллективных средствах размещения	Число мест в коллективных средствах размещения	0.6250
7	Число ночевок пребывания гостей	Число ночевок в коллективных средствах размещения	0.1803
8	Зарегистрированные автомобильные транспорты	Наличие автомобильного транспорта	0.3148
9	Библиотеки и тома	Выдано экземпляров из библиотечного фонда библиотек Минкультуры России	0.1077

10	Театры	Количество театров Минкультуры России	0.0811
11	Музей	Число музеев Минкультуры России	0.0645
12	Книги	Количество вновь вышедших книжных изданий	0.0244
13	Население на 1 января по стране	Численность постоянного населения на 1 января	0.2979
14	Браки и разводы	Число браков за год	0.0741
15	Показатели рождаемости	Суммарный коэффициент рождаемости	0.2093
16	Общий коэффициент прерываний беременности и количество искусственных абортов	Число прерываний беременности	0.2703
17	Количество искусственных прерываний беременности	Число прерываний беременности	0.4001
18	Прибывшие	Число прибывших	0.3529
19	Выбывшие	Число выбывших	0.3125
20	Чистая миграция	Миграционный прирост	0.1786
21	Умершие	Число зарегистрированных умерших	0.1212
22	Средняя численность населения	Среднегодовая численность населения	0.5676
23	Прямые иностранные инвестиции в процентах от ВВП	Доля прямых иностранных инвестиций в российской экономике к ввп	0.2568
24	Годовой рост реального валового внутреннего продукта	Индексы физического объёма валового регионального продукта в основных ценах	0.2021
25	Валовой внутренний продукт в PPS/евро	Валовой внутренний продукт в рыночных ценах в соответствии с методологией снс 1993 с частичными отклонениями	0.2347
26	Валовая добавленная стоимость (ESA2010)	Валовая добавленная стоимость в основных ценах в соответствии с методологией снс 1993 с частичными отклонениями	0.2551
27	Расходы на конечное потребление домохозяйств в миллионах евро	Расходы на конечное потребление домашних хозяйств	0.5556
28	Количество мест, занимаемых женщинами в национальных парламентах	Доля мест, занимаемых женщинами в национальных парламентах	0.7258
29	Выбросы загрязнителей воздуха (тонны)	Выбросы загрязняющих атмосферу веществ, отходящих от стационарных источников, за i полугодие	0.1327

Продолжение таблицы 1

30	Воздействие твердых частиц	Среднегодовой уровень содержания мелких твердых частиц (рм2,5 и рм10) в атмосфере городов гусиноозерск, иркутск, казань, улан-удэ	0.0973
31	Валовое внутреннее потребление энергии	Производство и потребление электроэнергии в Российской Федерации	0.1948
32	Первичное производство энергии по стране, продукту и времени	Производство и потребление электроэнергии в Российской Федерации	0.2135

33	Нормальная температура в странах Северной Европы	средняя месячная температура воздуха в июле (фактическая)	0.1646
34	Число смертей на 100 000 человек, вызванных злокачественными новообразованиями, в разбивке по полу, возрасту, времени и стране, представившей отчет	Смертность населения от злокачественных новообразований, на 100 тыс. населения	0.2199
35	Число смертей на 100 000 человек, вызванных сердечно-сосудистыми заболеваниями, в разбивке по полу, возрасту, времени и стране, представившей отчет	Смертность от сердечно-сосудистых заболеваний, злокачественных новообразований, сахарного диабета, хронических респираторных заболеваний	0.1579
36	Количество смертей на 100 000 населения в результате самоубийств	Смертность от самоубийств	0.2000
37	Количество смертей на 100 000 человек в результате несчастных случаев в разбивке по стране, времени, полу и возрасту	Количество пострадавших со смертельным исходом в результате зарегистрированных несчастных случаев на производстве	0.2789
38	Разрешенные больничные койки по специальности, стране, представившей отчет, отделению и времени	Обеспеченность больничными койками на 10 тыс. населения	0.1468
39	Рождаемая продолжительность жизни в разбивке по стране, возрасту, полу и времени жизни	Ожидаемая продолжительность жизни при рождении (прогноз)	0.3747
40	Подтвержденное число новых случаев инфекционных заболеваний ВИЧ/СПИДа	Число зарегистрированных случаев инфекционных заболеваний (человек)	0.3837
41	Доля населения, которое не работает и не учится (в процентах)	Доля молодежи (в возрасте от 15 до 24 лет), которая не учится, не работает и не приобретает профессиональных навыков	0.2054
42	Занятость	Уровень занятости	0.3158
43	Экспорт, 1000 евро	Стоимость экспортных (импортных) операций по субъектам Российской Федерации	0.0781
44	Импорт, 1000 евро	Стоимость экспортных (импортных) операций по субъектам Российской Федерации	0.0625
45	Занятые в возрасте 15–64 лет (1000 человек)	Численность занятых по полу и возрастным группам	0.1525
46	Время (часы), потраченное на работу по дому помимо работы	Доля времени, затрачиваемого на неоплачиваемые работу по дому и труд по уходу за членами домохозяйства и семьи	0.1316
47	Средний заработок в PPS/евро	Средний почасовой заработок женщин и мужчин в разбивке по группам занятий и возрасту	0.1566

Окончание таблицы 1

48	Возобновляемые источники энергии	Мощность генерирующих объектов, функционирующих на основе использования возобновляемых источников энергии (без учета гидроэлектростанций установленной мощностью свыше 25 мвт)	0.1643
49	Энергоемкость	Энергоемкость ввп	0.6111
50	Занятость	Уровень занятости	0.3158
51	Расходы на исследования и разработки	Внутренние текущие затраты на научные исследования и разработки	0.3333

52	Индекс цифровой экономики и общества (DESI)	Внутренние затраты на развитие цифровой экономики за счет всех источников	0.1765
53	Продолжительность жизни	Ожидаемая продолжительность жизни при рождении (прогноз)	0.4082
54	Самостоятельная оценка здоровья	Оценка респондентами состояния своего здоровья	0.2241
55	Экономическое неравенство (коэффициент Джини)	Коэффициент джини (индекс концентрации доходов)	0.2500
56	Риск бедности и социальной изоляции - уровень бедности	Уровень бедности	0.1500
57	Национальные индексы потребительских цен, индекс (2015 г. = 100)	индексы потребительских цен к среднегодовому значению 2010 г.	0.3692
58	Индексы цен на жилье (2015 г. = 100)	Индексы цен на рынке жилья	0.4286
59	Продажа алкогольных напитков (литры чистого спирта на человека)	- продажа алкогольных напитков в абсолютном алкоголе на душу населения	0.2892
60	Профицит, дефицит и долг сектора гос. управления	Дефицит/профицит консолидированного бюджета субъекта РФ и территориального государственного внебюджетного фонда	0.1081
61	Налоги и чистые налоги	Начисление и поступление налогов, сборов и иных обязательных платежей в бюджетную систему Российской Федерации	0.0600
62	Количество выпускников докторантуры или эквивалентного уровня (МСКО 2011)	Выпуск из докторантуры в отчетном году	0.1923
63	Расходы на исследования и разработки	Внутренние текущие затраты на научные исследования и разработки	0.3333
64	Интернет и широкополосный доступ в домохозяйствах	Число домохозяйств, имеющих широкополосный доступ к сети «интернет	0.5303
65	Телефонные подписки	Количество абонентских станций (устройств), подключённых к сетям подвижной радиотелефонной связи в стандарте GSM	0.0900
66	Эквивалентный средний доход в PPS/евро	Медианный эквивалентный располагаемый денежный доход	0.2381
67	коэффициент Джини	Коэффициент джини (индекс концентрации доходов)	0.3333
68	Риск бедности	Уровень бедности	0.2857
69	Общее количество пенсионеров	Общая численность пенсионеров	0.2326
70	Численность лиц, получающих финансовую социальную помощь	Численность лиц, имеющих право на получение государственной социальной помощи в виде набора социальных услуг, состоящих на учете в системе пенсионного фонда российской федерации	0.1595

4.2 Работа программы по алгоритму косинусного расстояния

В результате работы по алгоритму косинусного расстояния нашлась 80 пар строк, что составило 82% от необходимого объема. Также, выяснилось, что две строки, найденные алгоритмом Жаккара, не были найдены косинусным расстоянием. Это пары строк: “Растениеводство с пахотных

земель - Производство основных видов продукции в натуральном выражении” и “Налоги и чистые налоги - Зачисление и поступление налогов, сборов и иных обязательных платежей в бюджетную систему Российской Федерации”. Объединив алгоритм Жаккара и косинусное расстояние, мы получим 82 пары верно сопоставленных строк, что составит 85% от необходимого объёма (см. таблица 2).

Таблица 2 – Верно сопоставленные пары строк по косинусному расстоянию

№	Строки из файла EMISS	Строки из файла Nordic_stat	Косинусное расстояние
1	Урожайность с убранной площади	Урожайность сельскохозяйственных культур (в расчете на убранную площадь)	0.4141
2	Органическое сельское хозяйств	Внесено сельскохозяйственными организациями органических удобрений на 1 га посева	0.3615
3	Аквакультура, производство	Производство (выращивание) товарной рыбы и других объектов товарного рыбоводства (аквакультуры)	0.3556
4	Структурная статистика предприятий	Количество предприятий и организаций	0.2887
5	Кровати в коллективных средствах размещения	Число мест в коллективных средствах размещения	0.7289
6	Число ночевок пребывания гостей	Число ночевок в коллективных средствах размещения	0.2851
7	Зарегистрированные автомобильные транспорты	Наличие автомобильного транспорта	0.4866
8	Библиотеки и тома	Выдано экземпляров из библиотечного фонда библиотек Минкультуры России	0.3641
9	Театры	Количество театров Минкультуры России	0.2188
10	Музей	Число музеев Минкультуры России	0.1648
11	Книги	Количество вновь вышедших книжных изданий	0.0651
12	Население на 1 января по стране	Численность постоянного населения на 1 января	0.4208
13	Браки и разводы	Число браков за год	0.1466
14	Показатели рождаемости	Суммарный коэффициент рождаемости	0.2849
15	Общий коэффициент прерываний беременности и количество искусственных абортов	Число прерываний беременности	0.4301

Продолжение таблицы 2

16	Количество искусственных прерываний беременности	Число прерываний беременности	0.5714
17	Прибывшие	Число прибывших	0.5322
18	Выбывшие	Число выбывших	0.4887
19	Чистая миграция	Миграционный прирост	0.3236
20	Умершие	Число зарегистрированных умерших	0.2974
21	Смертность	Смертность от самоубийств	0.3278
22	Средняя численность населения	Среднегодовая численность населения	0.5796

23	Доля населения	Доля городского населения в общей численности населения на 1 января	0.3461
24	Прямые иностранные инвестиции в процентах от ВВП	Доля прямых иностранных инвестиций в российской экономике к ввп	0.3115
25	Годовой рост реального валового внутреннего продукта	Индексы физического объёма валового регионального продукта в основных ценах	0.2951
26	Валовой внутренний продукт в PPS/евро	Валовой внутренний продукт в рыночных ценах в соответствии с методологией снс 1993 с частичными отклонениями	0.3636
27	Валовая добавленная стоимость (ESA2010)	Валовая добавленная стоимость в основных ценах в соответствии с методологией снс 1993 с частичными отклонениями	0.4182
28	Расходы на конечное потребление домохозяйств в миллионах евро	Расходы на конечное потребление домашних хозяйств	0.6226
29	Количество мест, занимаемых женщинами в национальных парламентах	Доля мест, занимаемых женщинами в национальных парламентах	0.8455
30	Выбросы загрязнителей воздуха (тонны)	Выбросы загрязняющих атмосферу веществ, отходящих от стационарных источников, за i полугодие	0.2478
31	Воздействие твердых частиц	Среднегодовой уровень содержания мелких твердых частиц (рм2,5 и рм10) в атмосфере городов гусиноозерск, иркутск, казань, улан-удэ	0.1802
32	Счета выбросов в атмосферу	выбросы загрязняющих атмосферу веществ, отходящих от стационарных источников, за i полугодие	0.2285
33	Валовое внутреннее потребление энергии	Производство и потребление электроэнергии в Российской Федерации	0.3055
34	Первичное производство энергии по стране, продукту и времени	Производство и потребление электроэнергии в Российской Федерации	0.2873
35	Доля энергии из возобновляемых источников	доля производства электрической энергии генерирующими объектами, функционирующими на основе использования возобновляемых источников энергии, в совокупном объеме производства электрической энергии (без учета гидроэлектростанций установленной мощностью свыше 25 мвт)	0.3311
36	Расход удобрений, 1000 тонн чистого удобрения	внесено минеральных удобрений	0.3347

Продолжение таблицы 2

37	Развитие производства и обработки бытовых отходов	Вывезено твердых коммунальных отходов на объекты, используемые для обработки отходов (тысяча кубических метров, значение показателя за год)	0.20447
38	Нормальная температура в странах Северной Европы	средняя месячная температура воздуха в июле (фактическая)	0.2514
39	Число смертей на 100 000 человек, вызванных злокачественными новообразованиями, в разбивке по полу,	Смертность населения от злокачественных новообразований, на 100 тыс. населения	0.3178

	возрасту, времени и стране, представившей отчет		
40	Число смертей на 100 000 человек, вызванных сердечно-сосудистыми заболеваниями, в разбивке по полу, возрасту, времени и стране, представившей отчет	Смертность от сердечно-сосудистых заболеваний, злокачественных новообразований, сахарного диабета, хронических респираторных заболеваний	0.2959
41	Количество смертей на 100 000 населения в результате самоубийств	Смертность от самоубийств	0.4059
42	Количество смертей на 100 000 человек в результате несчастных случаев в разбивке по стране, времени, полу и возрасту	Количество пострадавших со смертельным исходом в результате зарегистрированных несчастных случаев на производстве	0.4141
43	Расходы на здравоохранение	расходы средств обязательного медицинского страхования в расчете на 1 жителя вместо «расходы консолидированного бюджета субъекта российской федерации на здравоохранение (средства обязательного медицинского страхования) в расчете на 1 жителя»	0.2654
44	Разрешенные больничные койки по специальности, стране, представившей отчет, отделению и времени	Обеспеченность больничными койками на 10 тыс. населения	0.2123
45	Рождаемая продолжительность жизни в разбивке по стране, возрасту, полу и времени жизни	Ожидаемая продолжительность жизни при рождении (прогноз)	0.5114
46	Подтвержденное число новых случаев инфекционных заболеваний ВИЧ/СПИДа	Число зарегистрированных случаев инфекционных заболеваний (человек)	0.5243
47	Доля населения, которое не работает и не учится (в процентах)	Доля молодежи (в возрасте от 15 до 24 лет), которая не учится, не работает и не приобретает профессиональных навыков	0.3587
48	Занятость	Уровень занятости	0.3857
49	Население	Доля городского населения в общей численности населения	
50	Экспорт, 1000 евро	Стоимость экспортных (импортных) операций по субъектам Российской Федерации	0.2248
51	Импорт, 1000 евро	Стоимость экспортных (импортных) операций по субъектам Российской Федерации	0.2022
52	Занятые в возрасте 15–64 лет (1000 человек)	Численность занятых по полу и возрастным группам	0.2090
53	Время (часы), потраченное на работу по дому помимо работы	Доля времени, затрачиваемого на неоплачиваемые работу по дому и труд по уходу за членами домохозяйства и семьи	0.2074

Продолжение таблицы 2

54	Работа и безработица	численность безработных граждан, зарегистрированных в государственных учреждениях службы занятости и получающих пособие по безработице на конец отчетного периода	0.2566
55	Дети в детских садах	численность детей, посещающих дошкольные образовательные учреждений	0.0699
56	Средний заработок в PPS/евро	Средний почасовой заработок	0.2428

		женщин и мужчин в разбивке по группам занятий и возрасту	
57	Возобновляемые источники энергии	Мощность генерирующих объектов, функционирующих на основе использования возобновляемых источников энергии (без учета гидроэлектростанций установленной мощностью свыше 25 мвт)	0.2988
58	Птицы в сельском хозяйстве	поголовье скота и птицы в хозяйствах всех категорий	0.29851
59	Энергоемкость	Энергоемкость ввп	0.2985
60	Занятость	Уровень занятости	0.3946
61	Расходы на исследования и разработки	Внутренние текущие затраты на научные исследования и разработки	0.4403
62	Индекс цифровой экономики и общества (DESI)	Внутренние затраты на развитие цифровой экономики за счет всех источников	0.2850
63	Продолжительность жизни	Ожидаемая продолжительность жизни при рождении (прогноз)	0.5022
64	Самостоятельная оценка здоровья	Оценка респондентами состояния своего здоровья	0.3307
65	Экономическое неравенство (коэффициент Джини)	Коэффициент джини (индекс концентрации доходов)	0.3952
66	Риск бедности и социальной изоляции - уровень бедности	Уровень бедности	0.1978
67	Национальные индексы потребительских цен, индекс (2015 г. = 100)	индексы потребительских цен к среднегодовому значению 2010 г.	0.5186
68	Индексы цен на жилье (2015 г. = 100)	Индексы цен на рынке жилья	0.4946
69	Продажа алкогольных напитков (литры чистого спирта на человека)	- продажа алкогольных напитков в абсолютном алкоголе на душу населения	0.4677
70	Профицит, дефицит и долг сектора гос. управления	Дефицит/профицит консолидированного бюджета субъекта РФ и территориального государственного внебюджетного фонда	0.2428
71	Количество патентных заявок	коэффициент изобретательской активности (число отечественных патентных заявок на изобретения, поданных в россии в расчете на 10 тыс. человек населения)	0.2877
72	Количество выпускников докторантуры или эквивалентного уровня (МСКО 2011)	Выпуск из докторантуры в отчетном году	0.3524
73	Расходы на исследования и разработки	Внутренние текущие затраты на научные исследования и разработки	0.4349
74	Интернет и широкополосный доступ в домохозяйствах	Число домохозяйств, имеющих широкополосный доступ к сети «интернет	0.6379

Окончание таблицы 2

75	Телефонные подписки	Количество абонентских станций (устройств), подключённых к сетям подвижной радиотелефонной связи в стандарте GSM	0.1679
76	Эквивалентный средний доход в PPS/евро	Медианный эквивалентный располагаемый денежный доход	0.3164
77	коэффициент Джини	Коэффициент джини (индекс концентрации доходов)	0.5191
78	Риск бедности	Уровень бедности	0.3113
79	Общее количество пенсионеров	Общая численность пенсионеров	0.39314

80	Численность лиц, получающих финансовую социальную помощь	Численность лиц, имеющих право на получение государственной социальной помощи в виде набора социальных услуг, состоящих на учете в системе пенсионного фонда российской федерации	0.2723

4.3 Работа программы по алгоритму Евклидова расстояния

В результате работы оказалось, что по Евклидовому расстоянию нашлись те же пары строк, что и по косинусному(см. таблица 3).

Таблица 3 – Верно сопоставленные пары строк по Евклидовому расстоянию

№	Строки из файла EMISS	Строки из файла Nordic_stat	Евклидово расстояние
1	Урожайность с убранной площади	Урожайность сельскохозяйственных культур (в расчете на убранную площадь)	1.0825
2	Органическое сельское хозяйств	Внесено сельскохозяйственными организациями органических удобрений на 1 га посева	1.1301
3	Аквакультура, производство	Производство (выращивание) товарной рыбы и других объектов товарного рыбоводства (аквакультуры)	1.1352
4	Структурная статистика предприятий	Количество предприятий и организаций	1.1927
5	Кровати в коллективных средствах размещения	Число мест в коллективных средствах размещения	0.7363
6	Число ночевок пребывания гостей	Число ночевок в коллективных средствах размещения	1.1957
7	Зарегистрированные автомобильные транспорты	Наличие автомобильного транспорта	1.0133
8	Библиотеки и тома	Выдано экземпляров из библиотечного фонда библиотек Минкультуры России	1.12782
9	Театры	Количество театров Минкультуры России	1.2499
10	Музей	Число музеев Минкультуры России	1.2925
11	Книги	Количество вновь вышедших книжных изданий	1.3674

Продолжение таблицы 3

12	Население на 1 января по стране	Численность постоянного населения на 1 января	1.0763
13	Браки и разводы	Число браков за год	1.3064
14	Показатели рождаемости	Суммарный коэффициент рождаемости	1.1958
15	Общий коэффициент прерываний беременности и количество искусственных абортов	Число прерываний беременности	1.0677
16	Количество искусственных прерываний	Число прерываний беременности	0.9258

	беременности		
17	Прибывшие	Число прибывших	0.9672
18	Выбывшие	Число выбывших	1.0112
19	Чистая миграция	Миграционный прирост	1.1631
20	Умершие	Число зарегистрированных умерших	1.1855
21	Смертность	Смертность от самоубийств	1.1594
22	Средняя численность населения	Среднегодовая численность населения	0.9169
23	Доля населения	Доля городского населения в общей численности населения на 1 января	1.1436
24	Прямые иностранные инвестиции в процентах от ВВП	Доля прямых иностранных инвестиций в российской экономике к ввп	1.1734
25	Годовой рост реального валового внутреннего продукта	Индексы физического объёма валового регионального продукта в основных ценах	1.1872
26	Валовой внутренний продукт в PPS/евро	Валовой внутренний продукт в рыночных ценах в соответствии с методологией снс 1993 с частичными отклонениями	1.1282
27	Валовая добавленная стоимость (ESA2010)	Валовая добавленная стоимость в основных ценах в соответствии с методологией снс 1993 с частичными отклонениями	1.0787
28	Расходы на конечное потребление домохозяйств в миллионах евро	Расходы на конечное потребление домашних хозяйств	0.8687
29	Количество мест, занимаемых женщинами в национальных парламентах	Доля мест, занимаемых женщинами в национальных парламентах	0.5558
30	Выбросы загрязнителей воздуха (тонны)	Выбросы загрязняющих атмосферу веществ, отходящих от стационарных источников, за i полугодие	1.22649
31	Воздействие твердых частиц	Среднегодовой уровень содержания мелких твердых частиц (рм2,5 и рм10) в атмосфере городов гусиноозерск, иркутск, казань, улан-удэ	1.2804
32	Счета выбросов в атмосферу	выбросы загрязняющих атмосферу веществ, отходящих от стационарных источников, за i полугодие	1.24212
33	Валовое внутреннее потребление энергии	Производство и потребление электроэнергии в Российской Федерации	1.1785
34	Первичное производство энергии по стране, продукту и времени	Производство и потребление электроэнергии в Российской Федерации	1.1939

Продолжение таблицы 3

35	Доля энергии из возобновляемых источников	доля производства электрической энергии генерирующими объектами, функционирующими на основе использования возобновляемых источников энергии, в совокупном объеме производства электрической энергии (без учета гидроэлектростанций установленной мощностью свыше 25 мвт)	1.1566

36	Расход удобрений, 1000 тонн чистого удобрения	внесено минеральных удобрений	1.1535
37	Развитие производства и обработки бытовых отходов	Вывезено твердых коммунальных отходов на объекты, используемые для обработки отходов (тысяча кубических метров, значение показателя за год)	1.26136
38	Нормальная температура в странах Северной Европы	средняя месячная температура воздуха в июле (фактическая)	1.2236
39	Число смертей на 100 000 человек, вызванных злокачественными новообразованиями, в разбивке по полу, возрасту, времени и стране, представившей отчет	Смертность населения от злокачественных новообразований, на 100 тыс. населения	1.1681
40	Число смертей на 100 000 человек, вызванных сердечно-сосудистыми заболеваниями, в разбивке по полу, возрасту, времени и стране, представившей отчет	Смертность от сердечно-сосудистых заболеваний, злокачественных новообразований, сахарного диабета, хронических респираторных заболеваний	1.1867
41	Количество смертей на 100 000 населения в результате самоубийств	Смертность от самоубийств	1.0899
42	Количество смертей на 100 000 человек в результате несчастных случаев в разбивке по стране, времени, полу и возрасту	Количество пострадавших со смертельным исходом в результате зарегистрированных несчастных случаев на производстве	1.0825
43	Расходы на здравоохранение	расходы средств обязательного медицинского страхования в расчете на 1 жителя вместо «расходы консолидированного бюджета субъекта российской федерации на здравоохранение (средства обязательного медицинского страхования) в расчете на 1 жителя»	1.2121
44	Разрешенные больничные койки по специальности, стране, представившей отчет, отделению и времени	Обеспеченность больничными койками на 10 тыс. населения	1.2551
45	Рождаемая продолжительность жизни в разбивке по стране, возрасту, полу и времени жизни	Ожидаемая продолжительность жизни при рождении (прогноз)	0.9885
46	Подтвержденное число новых случаев инфекционных заболеваний ВИЧ/СПИДа	Число зарегистрированных случаев инфекционных заболеваний (человек)	0.9754
47	Доля населения, которое не работает и не учится (в процентах)	Доля молодежи (в возрасте от 15 до 24 лет), которая не учится, не работает и не приобретает профессиональных навыков	1.1325
48	Занятость	Уровень занятости	1.1084
49	Население	Доля городского населения в общей численности населения на 1 января	1.2244

Продолжение таблицы 3

50	Экспорт, 1000 евро	Стоимость экспортных (импортных) операций по субъектам Российской Федерации	1.2450
51	Импорт, 1000 евро	Стоимость экспортных (импортных) операций по субъектам Российской Федерации	1.2631
52	Занятые в возрасте 15–64 лет (1000 человек)	Численность занятых по полу и возрастным группам	1.2577
5	Время (часы), потраченное на работу по	Доля времени, затрачиваемого на	1.2589

53	дому помимо работы	неоплачиваемые работу по дому и труд по уходу за членами домохозяйства и семьи	
54	Работа и безработица	численность безработных граждан, зарегистрированных в государственных учреждениях службы занятости и получающих пособие по безработице на конец отчетного периода	1.2193
55	Дети в детских садах	численность детей, посещающих дошкольные образовательные учреждений	1.3638
56	Средний заработок в PPS/евро	Средний почасовой заработок женщин и мужчин в разбивке по группам занятий и возрасту	1.2305
57	Возобновляемые источники энергии	Мощность генерирующих объектов, функционирующих на основе использования возобновляемых источников энергии (без учета гидроэлектростанций установленной мощностью свыше 25 мвт)	1.1842
58	Птицы в сельском хозяйстве	поголовье скота и птицы в хозяйствах всех категорий	1.1845
59	Энергоемкость	Энергоемкость ввп	0.7824
60	Занятость	Уровень занятости	1.1004
61	Расходы на исследования и разработки	Внутренние текущие затраты на научные исследования и разработки	1.0579
62	Индекс цифровой экономики и общества (DESI)	Внутренние затраты на развитие цифровой экономики за счет всех источников	1.1957
63	Продолжительность жизни	Ожидаемая продолжительность жизни при рождении (прогноз)	0.9977
64	Самостоятельная оценка здоровья	Оценка респондентами состояния своего здоровья	1.1569
65	Экономическое неравенство (коэффициент Джини)	Коэффициент джини (индекс концентрации доходов)	1.0998
66	Риск бедности и социальной изоляции - уровень бедности	Уровень бедности	1.2666
67	Национальные индексы потребительских цен, индекс (2015 г. = 100)	индексы потребительских цен к среднегодовому значению 2010 г.	0.9812
68	Индексы цен на жилье (2015 г. = 100)	Индексы цен на рынке жилья	1.0053
69	Продажа алкогольных напитков (литры чистого спирта на человека)	- продажа алкогольных напитков в абсолютном алкоголе на душу населения	1.0318
70	Профицит, дефицит и долг сектора гос. управления	Дефицит/профицит консолидированного бюджета субъекта РФ и территориального государственного внебюджетного фонда	1.2305

Окончание таблицы 3

71	Количество патентных заявок	коэффициент изобретательской активности (число отечественных патентных заявок на изобретения, поданных в россии в расчете на 10 тыс. человек населения)	1.1935
72	Количество выпускников докторантуры или эквивалентного уровня (МСКО 2011)	Выпуск из докторантуры в отчетном году	1.1381
73	Расходы на исследования и разработки	Внутренние текущие затраты на научные исследования и разработки	1.0631

74	Интернет и широкополосный доступ в домохозяйствах	Число домохозяйств, имеющих широкополосный доступ к сети «интернет	0.8509
75	Телефонные подписки	Количество абонентских станций (устройств), подключённых к сетям подвижной радиотелефонной связи в стандарте GSM	1.2899
76	Эквивалентный средний доход в PPS/евро	Медианный эквивалентный располагаемый денежный доход	1.1692
77	коэффициент Джини	Коэффициент джини (индекс концентрации доходов)	0.9807
78	Риск бедности	Уровень бедности	1.1736
79	Общее количество пенсионеров	Общая численность пенсионеров	1.1016
80	Численность лиц, получающих финансовую социальную помощь	Численность лиц, имеющих право на получение государственной социальной помощи в виде набора социальных услуг, состоящих на учете в системе пенсионного фонда российской федерации	1.2064

4.4 Работа программы по алгоритму манхэттенского расстояния.

Хуже всего показала себя метрика манхэттенского расстояния. С ее помощью было найдено только 55 пар строк, что составило 57% от необходимого объема (см. таблица 4).

Таблица 4 – Верно сопоставленные пары строк по манхэттенскому расстоянию

№	Строки из файла EMISS	Строки из файла Nordic_stat	манхэттенское расстояние
1	Урожайность с убранной площади	Урожайность сельскохозяйственных культур (в расчете на убранную площадь)	0.1076
2	Органическое сельское хозяйств	Внесено сельскохозяйственными организациями органических удобрений на 1 га посева	0.1080
3	Аквакультура, производство	Производство (выращивание) товарной рыбы и других объектов товарного рыбоводства (аквакультуры)	0.1010
4	Кровати в коллективных средствах размещения	Число мест в коллективных средствах размещения	0.2180
5	Число ночевок пребывания гостей	Число ночевок в коллективных средствах размещения	0.2036
6	Зарегистрированные автомобильные транспорты	Наличие автомобильного транспорта	0.1342

Продолжение таблицы 4

7	Библиотеки и тома	Выдано экземпляров из библиотечного фонда библиотек Минкультуры России	0.1100
8	Театры	Количество театров Минкультуры России	0.1236
9	Музей	Число музеев Минкультуры России	0.1648
10	Книги	Количество вновь вышедших книжных изданий	0.1291
11	Население на 1 января по стране	Численность постоянного населения	0.1109

		на 1 января	
12	Количество искусственных прерываний беременности	Число прерываний беременности	0.1448
13	Прибывшие	Число прибывших	0.2307
14	Выбывшие	Число выбывших	0.2209
15	Чистая миграция	Миграционный прирост	0.1524
16	Умершие	Число зарегистрированных умерших	0.1372
17	Смертность	Смертность от самоубийств	0.1216
18	Средняя численность населения	Среднегодовая численность населения	0.2094
19	Доля населения	Доля городского населения в общей численности населения на 1 января	0.1241
20	Прямые иностранные инвестиции в процентах от ВВП	Доля прямых иностранных инвестиций в российской экономике к ввп	0.1073
21	Валовой внутренний продукт в PPS/евро	Валовой внутренний продукт в рыночных ценах в соответствии с методологией снс 1993 с частичными отклонениями	0.0871
22	Валовая добавленная стоимость (ESA2010)	Валовая добавленная стоимость в основных ценах в соответствии с методологией снс 1993 с частичными отклонениями	0.0909
23	Расходы на конечное потребление домохозяйств в миллионах евро	Расходы на конечное потребление домашних хозяйств	0.1627
24	Количество мест, занимаемых женщинами в национальных парламентах	Доля мест, занимаемых женщинами в национальных парламентах	0.2628
25	Валовое внутреннее потребление энергии	Производство и потребление электроэнергии в Российской Федерации	0.0972
26	Доля энергии из возобновляемых источников	доля производства электрической энергии генерирующими объектами, функционирующими на основе использования возобновляемых источников энергии, в совокупном объеме производства электрической энергии (без учета гидроэлектростанций установленной мощностью свыше 25 мвт)	0.0674
27	Количество смертей на 100 000 человек в результате несчастных случаев в разбивке по стране, времени, полу и возрасту	Количество пострадавших со смертельным исходом в результате зарегистрированных несчастных случаев на производстве	0.0811

Продолжение таблицы 4

28	Расходы на здравоохранение	расходы средств обязательного медицинского страхования в расчете на 1 жителя вместо «расходы консолидированного бюджета субъекта российской федерации на здравоохранение (средства обязательного медицинского страхования) в расчете на 1 жителя»	0.0709
29	Рождаемая продолжительность жизни в	Ожидаемая продолжительность	0.1397

	разбивке по стране, возрасту, полу и времени жизни	жизни при рождении (прогноз)	
30	Подтвержденное число новых случаев инфекционных заболеваний ВИЧ/СПИДа	Число зарегистрированных случаев инфекционных заболеваний (человек)	0.1166
31	Доля населения, которое не работает и не учится (в процентах)	Доля молодежи (в возрасте от 15 до 24 лет), которая не учится, не работает и не приобретает профессиональных навыков	0.0827
32	Занятость	Уровень занятости	0.2035
33	Экспорт, 1000 евро	Стоимость экспортных (импортных) операций по субъектам Российской Федерации	0.0991
34	Импорт, 1000 евро	Стоимость экспортных (импортных) операций по субъектам Российской Федерации	0.0991
35	Работа и безработица	численность безработных граждан, зарегистрированных в государственных учреждениях службы занятости и получающих пособие по безработице на конец отчетного периода	0.0716
36	Средний заработок в PPS/евро	Средний почасовой заработок женщин и мужчин в разбивке по группам занятий и возрасту	0.0864
37	Возобновляемые источники энергии	Мощность генерирующих объектов, функционирующих на основе использования возобновляемых источников энергии (без учета гидроэлектростанций установленной мощностью свыше 25 мвт)	0.0719
38	Птицы в сельском хозяйстве	поголовье скота и птицы в хозяйствах всех категорий	0.1101
39	Энергоемкость	Энергоемкость ввп	0.3144
40	Занятость	Уровень занятости	0.2034
41	Расходы на исследования и разработки	Внутренние текущие затраты на научные исследования и разработки	0.1171
42	Индекс цифровой экономики и общества (DESI)	Внутренние затраты на развитие цифровой экономики за счет всех источников	0.1397
43	Продолжительность жизни	Ожидаемая продолжительность жизни при рождении (прогноз)	0.08984
44	Самостоятельная оценка здоровья	Оценка респондентами состояния своего здоровья	0.1124
45	Экономическое неравенство (коэффициент Джини)	Коэффициент джини (индекс концентрации доходов)	0.1506
46	Риск бедности и социальной изоляции - уровень бедности	Уровень бедности	0.1747
47	Национальные индексы потребительских цен, индекс (2015 г. = 100)	индексы потребительских цен к среднегодовому значению 2010 г.	0.13211

Окончание таблицы 4

48	Индексы цен на жилье (2015 г. = 100)	Индексы цен на рынке жилья	0.1989
49	Продажа алкогольных напитков (литры чистого спирта на человека)	- продажа алкогольных напитков в абсолютном алкоголе на душу населения	0.1083
50	Количество патентных заявок	коэффициент изобретательской активности (число отечественных патентных заявок на изобретения, поданных в россии в расчете на 10 тыс. человек населения)	0.0789

51	Расходы на исследования и разработки	Внутренние текущие затраты на научные исследования и разработки	0.1171
52	Интернет и широкополосный доступ в домохозяйствах	Число домохозяйств, имеющих широкополосный доступ к сети «интернет	0.1682
53	коэффициент Джини	Коэффициент джини (индекс концентрации доходов)	0.1506
54	Риск бедности	Уровень бедности	0.1747
55	Общее количество пенсионеров	Общая численность пенсионеров	0.1372

5 ПОЛЬЗОВАТЕЛЬСКОЕ ПРИЛОЖЕНИЕ

Для создания приложения были выбраны косинусное и Евклидовое расстояния, поскольку они показали наилучшие результаты. При запуске пользователю будет предложено выбрать два исходных файла (рисунок 3). После их загрузки приложение предложит пользователю вписать название

для новых файлов с результатами (сперва для косинусного, а затем для евклидового расстояния) и выбрать папку для их сохранения (рисунок 4).

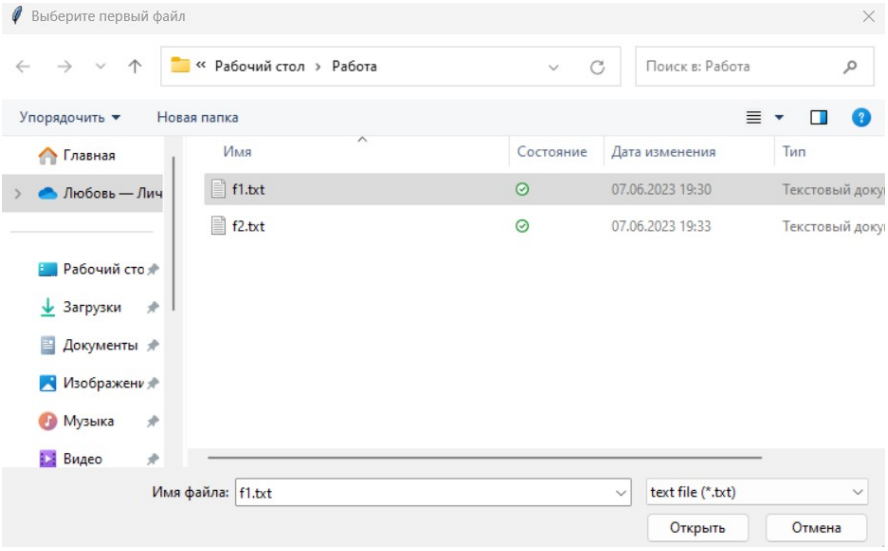

Рисунок 3 – Выбор исходных файлов

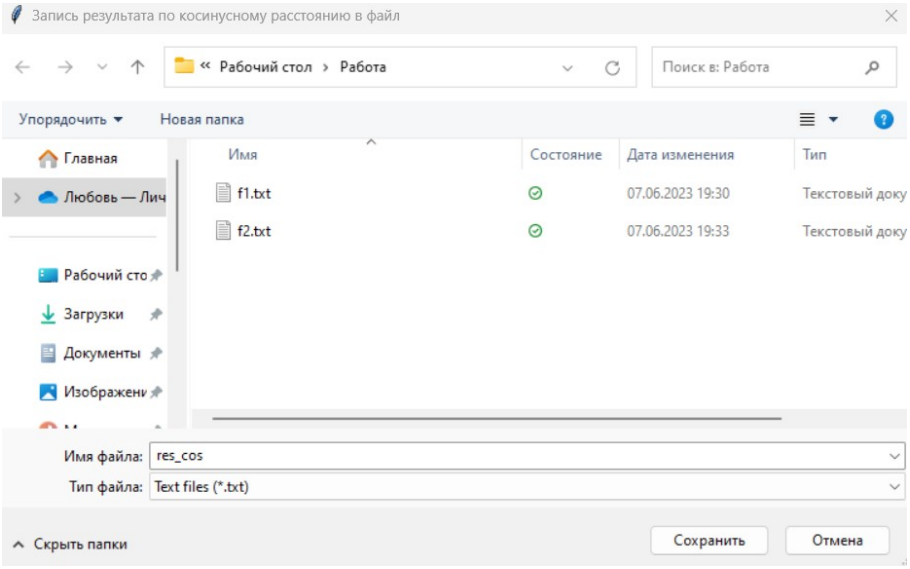

Рисунок 4 – Выбор файлов для записи результатов

Работа приложения очень быстрая. Так, например, файлы с размерность 100*100 строк сопоставятся быстрее 5-ти секунд.

6. ПРОГРАММНЫЕ СРЕДСТВА

6.1. Выбор программных средств

В своей дипломной работе я реализую алгоритмы на языке Python 3.11 с использованием библиотеки Scikit-learn.

6.2.Описание программных средств

Python - это интерпретируемый высокоуровневый язык программирования, который широко используется программистами, благодаря своей простоте использования и большому количеству библиотек, которые помогают в решении различных задач.

Scikit-learn - один из наиболее широко используемых пакетов Python для аналитики данных. Благодаря этой библиотеке легко реализуется подсчет косинусного, евклидового и манхэттенского расстояния, а также векторизация TF*IDF.

7 ПРОГРАММИРОВАНИЕ АЛГОРИТМОВ

7.1 Программирование алгоритма Жаккара

Основные функции программы для нахождения похожих строк по алгоритму Жаккара показаны в таблицы 5.

Таблица 5 – Описание функция для алгоритма Жаккара

Функция	Описание
`def clean_text(text):` `  return ''.join([c.lower() for c in text if c.isalpha()])`	Функция очистки текста от знаков пунктуации, цифр и небуквенных символов
`def get_trigrams(text):` `  return [text[i:i+3] for i in range(len(text)-2)]`	Функция разбиения слов на триграммы

`ef get_jaccard_similarity(trigrams1, trigrams2):`
`  set1 = set(trigrams1)`
`  set2 = set(trigrams2)`
`  intersection = set1.intersection(set2)`
`  union = set1.union(set2)`
`  return len(intersection) / len(union)`
	Функция сравнения двух списков триграмм на основе сходства Жаккара

7.2 Программирование алгоритма косинусного, евклидового расстояния и манхэттенского расстояния

Основные функции программы для нахождения похожих строк по алгоритму косинусного, евклидового и манхэттенского расстояния показаны в таблице 6.

Таблица 6 – Описание функция для нахождения косинусного, евклидового и манхэттенского расстояния

Функция	Описание

`def str_to_trigram(s):`
`  s = s.rstrip()`
`  s = s.lower()`
`  s = ''.join([letter for letter in s if letter.isalpha()])`
`  s = ["".join(j) for j in zip(*[s[i:] for i in range(3)])]`
`  s = ' '.join(s)`
`  return s`
	Функция очистки текста от знаков пунктуации, цифр и небуквенных символов Разбиение слов на триграммы

`tfidf_vectorizer = TfidfVectorizer`	Нахождение TF* IDF(Встроенная функция библиотеки sklearn)
`cosines = cosine_similarity(vectors[indx, :], vectors[len(text1):, :])`	Вычисление косинусного расстояния
`distances = euclidean_distances(vectors[indx, :], vectors[len(text1):, :])`	Вычисление Евклидового расстояния
`distances_m = manhattan_distances(vectors[indx, :], vectors[len(text1):, :])`	Вычисление манхэттенского расстояния

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы были изучены различные алгоритмы вычисления строковых метрик, реализовано пользовательское приложение для нахождения похожих строк из двух баз данных. Было проведено сравнение эффективности алгоритмов, их описание, выделение плюсов и минусов.

Лучше всего на практике себя показали алгоритмы косинусного и евклидового расстояния. С их помощью удалось верно сопоставить 82% пар строк от необходимого объема. Если объединить один из этих алгоритмов с алгоритмом Жаккара, выходит найти 85% необходимого объема.

В целом, данные алгоритмы является удобным инструментом для сравнения множеств и может быть полезным в различных областях, но его использование не всегда проводится безгранично и требует осторожности в интерпретации результатов.

СПИСОК ЛИТЕРАТРЫ

1. Косинусное сходство // Русские Блоги [Электронный ресурс] URL: https://russianblogs.com/article/7696266567/

2. Евклидовое расстояние// Русские Блоги [Электронный ресурс] URL: https://russianblogs.com/article/7696266567/

3. Сходство Жаккара // Хабр [Электронный ресурс] URL: https://habr.com/ru/companies/skillfactory/articles/566414/

4. Векторизация TF*IDF // Хабр [Электронный ресурс] URL: https://habr.com/ru/companies/skillfactory/articles/566414/

5. Ричмонд Элэйк. Косинусное расстояние: Понимание косинусного подобия и его применения в машинном обучении. - 2023 г.

6. Ирзум Джафри. Манхэттенское расстояние: Что такое Манхэттенское расстояние в машинном обучении? – 2023 г.

7. Манхэттенское расстояние // Кодкамп [Электронный ресурс] URL: https://www.codecamp.ru/blog/manhattan-distance-python

ПРИЛОЖЕНИЕ А. ТЕКСТ ПРОГРАММЫ ДЛЯ ВЫЧИСЛЕНИЯ РАССТОЯНИЯ ЖАККАРА

```
import string

def clean_text(text):
    return ''.join([c.lower() for c in text if c.isalpha()])

def get_trigrams(text):
#Функция разбиения слова на триграммы'''
    return [text[i:i+3] for i in range(len(text)-2)]

def get_jaccard_similarity(trigrams1, trigrams2):
    set1 = set(trigrams1)
    set2 = set(trigrams2)
    intersection = set1.intersection(set2)
    union = set1.union(set2)
    return len(intersection) / len(union)

with open('file2.txt', 'r', encoding='utf-8') as f1, open('file1.txt',
'r', encoding='utf-8') as f2:
    lines1 = f1.readlines()
    lines2 = f2.readlines()
with open('result.txt', 'w', encoding='utf-8') as f:
    for line1 in lines1:
        text1 = clean_text(line1)
        trigrams1 = get_trigrams(text1)
        max_similarity = 0
        max_line = ''
        for line2 in lines2:
            text2 = clean_text(line2)
            trigrams2 = get_trigrams(text2)
            similarity = get_jaccard_similarity(trigrams1, trigrams2)
            if similarity > max_similarity:
                max_similarity = similarity
                max_line = line2
        f.write(f'{line1.strip()} - {max_line.strip()}
({max_similarity:.4f})\n')
```

ПРИЛОЖЕНИЕ Б. ТЕКСТ ПРОГРАММЫ ДЛЯ АЛГОРИТМА ВЫЧИСЛЕНИЯ КОСИНУСНОГО РАССТОЯНИЯ И ЕВКЛИДОВОГО РАССТОЯНИЯ (ПОЛЬЗОВАТЕЛЬСКОЕ ПРИЛОЖЕНИЯ)

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.metrics.pairwise import euclidean_distances
from tkinter.filedialog import asksaveasfilename, askopenfilename
import tkinter as tk

def str_to_trigram(s):
    s = s.rstrip()
    s = s.lower()
    s = ''.join([letter for letter in s if letter.isalpha()])
    s = ["".join(j) for j in zip(*[s[i:] for i in range(3)])]
    s = ' '.join(s)
    return s

root = tk.Tk()
root.withdraw()
file_first = askopenfilename(title="Выберите первый файл",
filetypes=(("text file", "*.txt"),))
with open(file_first, encoding='utf-8') as f:
    text1 = f.readlines()
file_second = askopenfilename(title="Выберите второй файл",
filetypes=(("text file", "*.txt"),))
with open(file_second, encoding='utf-8') as f:
    text2 = f.readlines()

tfidf_vectorizer = TfidfVectorizer(analyzer='word')
vectors = tfidf_vectorizer.fit_transform([str_to_trigram(i) for i in
text1] +
                                         [str_to_trigram(i) for i in
text2])
result = []
print('Обработка данных')
for indx, s in enumerate(text1):
    cosines = cosine_similarity(vectors[indx, :], vectors[len(text1):,
:])
    result.append(f'{text1[indx].rstrip()}
{text2[cosines.argmax()].rstrip()} {cosines.max()}')
file_save = asksaveasfilename(title="Запись результата по косинусному
расстоянию в файл", filetypes=(("Text files", "*.txt"),))
if file_save:
    with open(file_save + '.txt', 'w', encoding='utf-8') as f:
        for line in result:
            f.write(f"{line}\n")

print('Обработка данных')
for indx, s in enumerate(text1):
```

```
    distances = euclidean_distances(vectors[indx, :],
vectors[len(text1):, :])
    result.append(f'{text1[indx].rstrip()}
{text2[distances.argmin()].rstrip()} {distances.min()}')
file_save = asksaveasfilename(title="Запись результата по евклидовому
расстоянию в файл", filetypes=(("Text files", "*.txt"),))
if file_save:
    with open(file_save + '.txt', 'w', encoding='utf-8') as f:
        for line in result:
            f.write(f"{line}\n")
root.destroy()
```