**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего образования**

**«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**

ИНСТИТУТ МАТЕМАТИКИ И МЕХАНИКИ ИМ. Н. И. ЛОБАЧЕВСКОГО  
КАФЕДРА КОМПЬЮТЕРНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Направление подготовки: 02.03.01 – «Математика и компьютерные науки»

Профиль: Наука о данных

КУРСОВАЯ РАБОТА

СЕМАНТИЧЕСКОЕ РАССТОЯНИЕ МЕЖДУ ДОКУМЕНТАМИ   
И МЕТОДЫ ЕГО ВЫЧИСЛЕНИЯ

Студент 3 курса

группы 05-804

«\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2021 г \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.Р. Имамутдинов

Научный руководитель:

к. п. н, доцент каф. ИМиМ

«\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2021 г. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е. К. Липачев

Казань – 2021

# Оглавление

**Аннотация3**

**Введение4**

**Раздел 1. Постановка задачи6**

**Раздел 2. Предварительные сведения7**

**Раздел 3. Методы и их оценки13**

**Раздел 4. Реализация16**

**Заключение 23**

**Литература 24**

# Аннотация

В работе представлен обзор по представлению и анализу документов и слов компьютером. Описаны некоторые подходы и методы для вычисления семантического расстояния между словами и документами, приведена их оценка. Показан пример расчета близости загружаемого документа в базу.

# **Abstract**

This work presents an overview of presentation and analysis of documents and words by a computer. Described some approaches and methods for calculating semantic distance between documents and words, and given their evaluation. Shown an example of calculating semantic distance between uploaded documents at base.

Ключевые слова: семантическое расстояние, семантическая близость документов, автоматическая обработка текстов.

Keywords: semantic distance, semantic proximity of documents, similarity between documents, automatic text processing.



# Введение

Появление интернета не только сделало информацию доступной человеку, но и добавило немало проблем программистам, которые занимаются облегчением поиска информации. А именно, уже возникают трудности представления и поиска интересующей нас информации без применения систем автоматической обработки текста (АОТ). АОТ – это преобразование неструктурированного текста на искусственном или естественном языке компьютером. Эти системы широко применяются в различных сферах и решают конкретные задачи (см. например, [1],[2]). Среди них можно выделить несколько типов этой системы: издательские, направленные на лексикографическую подготовку и использование словарей, информационно-поисковые, а также существуют системы поддержки диалога с пользователем, распознавания речи пользователя и машинного перевода. В каждом из типов, рассматриваются разные подходы к обнаружению семантических схожих данных (документов) по вычислению их семантического расстояния исходя из текстового смысла документа.

Приведем пример: Пусть у нас есть некоторый не корректный запрос от пользователя, где есть ошибка в орфографии некоторого слова, семантический анализ запроса позволяет увидеть насколько данное слово, отделяется от всего запроса, что подталкивает на мысль о том, что это слово может быть написано неправильно. Тогда, если заменить ошибочное слово на семантически близкое к нему слово, относительно всего запроса, вероятнее всего, обработка запроса станет более точным, по сравнению с обработкой запроса с ошибкой. Это можно объяснить тем, что замена не отразится на результате слишком сильно, потому что результат будет иметь общую семантическую близость с запросом, даже если результат и запрос представляют собой разные термины для описания своей сути (см. например, [3]).

Замена семантически близких слов не столь значительно отражается во всем тексте, чем замена на различные семантические слова. Аналогично можно заменять не только слова, но и предложения или даже куски текста, не меняя основной сути документа (см. например [4],[5]).

Все эти системы пока что не могут справиться самостоятельно, потому что обеспечить анализ, синтез и оценку введенного языкового знака достаточно тяжело, тем более с учетом уже существующих языков. Именно поэтому задачи эффективного поиска семантически подобных документов пока еще остаются актуальными, к тому же значительное увеличение объема информации во всем мире способствует постоянной необходимости совершенствования существующих методов и подходов. А так же хочется отметить, что подтверждением актуальности являются современные исследования на российских и международных уровнях, с целью разработки новых систем АОТ и оптимизации существующих средств обработки информации (см. например,[5],[6],[7]).

В первой части курсовой работы представлена модель хранения информации компьютером, дано определение семантического расстояния, изложены основные подходы к его вычислению.

Во второй части курсовой работы описаны некоторые методы по вычислению семантического расстояния на основе подходов, даны оценки их эффективности, представлен пример вычисления. Опираясь на изложенный материал, дано заключение.

Работа состоит из 25 страницы, содержит 19 литературных источников.

При описании методов по вычислению семантического расстояния между документами, будем предполагать, что уже существует некоторая база, состоящая из сгруппированных по своему смыслу документов. А именно, мы рассмотрим такие методы как: метод шинглов, метод латентно-семантического анализа, метод основанный на векторной модели (VSM – vector space model) представления документов для вычисления семантического расстояния между документами, и приведем оценку описанных методов.

# Раздел 1. Постановка задачи

Объектом исследования является автоматическая обработка текста, а предметом исследования – документы, которым в соответствии с их текстом, можно сопоставить другие документы по разным критериям.

Задачами данного исследования будет:

1. Дать определение семантического расстояния и описать его, то есть что оно из себя представляет.
2. Представить основные подходы и методы его вычисления.
3. Оценка полученного семантического расстояния между документами для данных методов.

Целью данной работы является исследование известных методов определения семантической близости документов и возможность их применения к анализу математического контента.

Результатом работы получим представление о том, как устанавливаются соответствия между различными документами и как современные системы их разделяют.

# Раздел 2. Предварительные сведенья

Задачу поиска и определения семантических расстояний пытаются решить на многих российских и международных конференциях. Этой задаче уделяются большое внимание (см. например [6],[7]). Почему с этой задачей люди справляются точнее, чем многие современные средства и технологии обработки текстов, хоть и за большее время? Дело в том, что человек, с начала своей жизни принимает и осознает огромную массу информации и на протяжении всей своей жизни ее структурирует, что позволяет ориентироваться как в усвоенных знаниях, так и в новых, а также человек может не стандартно мыслить. Машины же, не обладают таким сильным разумом, не могут думать как человек, особенно не стандартно (не по алгоритму), для них задачи осмысления текста даются не просто… Чтобы разобраться в том, что такое семантическое расстояние и как его считать между документами, нужно иметь представление о том, как компьютер представляет, обрабатывает и анализирует информацию из документов, после чего можно будет говорить о сравнении двух документов.

Существует векторное представление документов (согласно статье [8],[9]), рассмотрим его суть: для каждого загружаемого документа в базу проводим анализ его содержимого, а именно путем стемминга и лемматизации его слов выделяем термы (подробнее про стемминг, лемматизацию [1],[10]). Все термы записываем в одну таблицу - в строки, а документы в столбцы. На пересечении строки и столбца ставим некоторое число , которое высчитывается на анализе частоты термов среди документов, оно отражает важность данного терма t для данного документа doc , а именно его вес (см. пример в главе 4). В данной таблице термов-документов, определим близость документов, а именно семантическое расстояние по различным формулам, например:

Скалярное расстояние -

,

где n - количество термов, а результат не нормирован;

Евклидово расстояние –

*,*

где n - количество термов, а результат не нормирован;

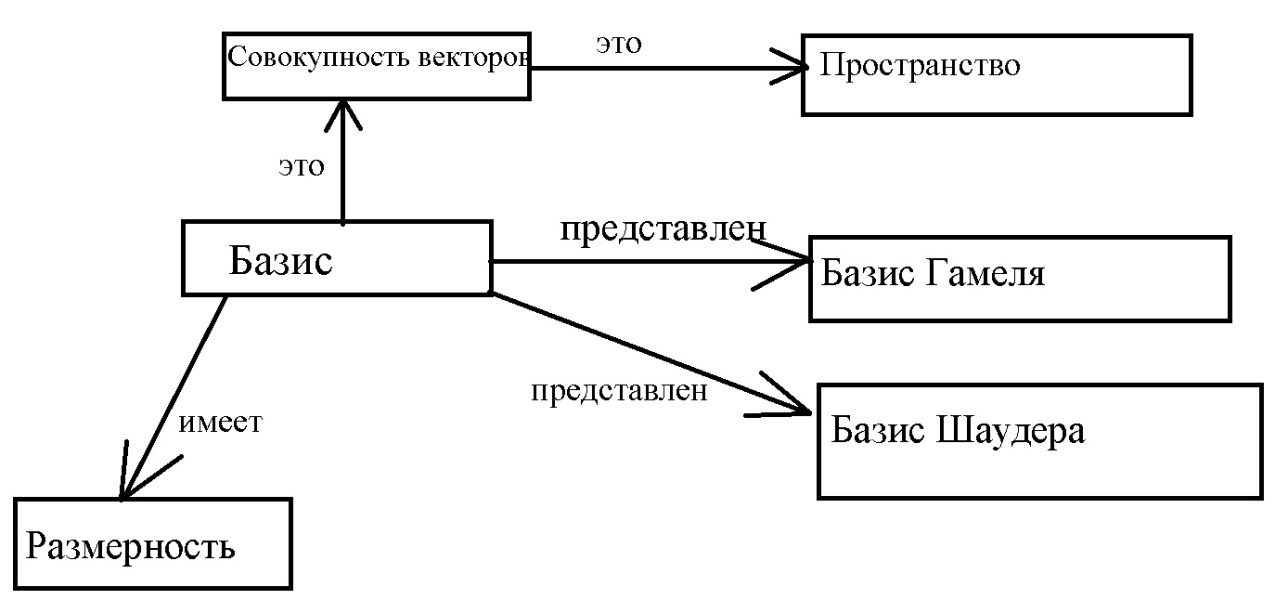
Косинусное расстояние –

где n – количество термов, и результат нормирован;

и другие – для векторов – столбцов данной таблицы. Само число, которое получается в результате измерения подобия и будет называться семантическим расстоянием от одного объекта до другого, то есть насколько они похожи (см. например [9]).

Однако, полученный результат может отражать далеко не верную оценку близости документов, потому что мы не обращались к семантике языка, а именно смотрели на термы как на отдельные единицы документа, то есть не рассматривали их в совокупности (в предложении, абзаце и т.д.). Чтобы получить более точное расстояние между документами нужно владеть хорошим знанием терминологии, которая включает в себя хорошее умение сопоставления слов их значению, то есть семантики языка. Сама семантика - это раздел логики и семиотики, в котором рассматриваются отношения между словами и предложениями к их смыслу, значению (см. например [3]).

Для этого необходимо как то хранить, различать, задать отношения между сущностями (термами). Одно из таких компьютерных представлений данных – называется семантической сетью (как указано в источнике [11],[12],[13]) . Семантическая сеть – это модель некоторой предметной области, составленная из понятий данной модели, записанная в виде графа, где вершины – это понятия, а ребра это отношение между понятиями. В качестве понятий можно рассматривать термы – субъекты, тогда отношениями будет их смысловая квалификация – предикаты между другими термами – объектами, семантически близкими с субъектом (на основе вычисления семантического расстояния). Например, с помощью sparsql запроса на ресурсе https://www.dbpedia.org/ была составлена rdf таблица и представлена в виде графа для субъекта “Базис”.

Рис. 1 Семантическая сеть “Базис”

Такая модель представления семантической сети называется RDF (Resourse Description Framework) – одно из вариаций онтологии предметной области (в этом случае “Базис”), где сущность (терм) представляется в виде триплета (набор из объекта, предиката, субъекта). Также каждый субъект в семантической сети обладает своим uri, это некий идентификатор этого ресурса, чтобы не было “дубликатов”. Базис – это субъект, ‘это’ – это предикат, совокупность векторов – объект (см. например [4],[13],[14]).

С учетом этого, компьютер будет понимать, какие узлы семантической сети (термы – субъекты) к какому классу (теме) относятся, что будет говорить нам о том, что совокупность термов идущих рядом друг с другом в тексте документа будет принадлежать какой – то одной теме (семантической сети). Распределение термов, расположенных рядом друг с другом по темам лежит в основе метода Латентно – семантического анализа, о котором подробнее написано в главе 3. Описанный процесс является Семантическим анализом документа – это выявление его тематики, смысла, сопоставление с существующими сущностями в базе сгруппированных документов (см. например [5],[6],[10],[15]).

В интернете существует множество семантических сетей, семантическую сеть объединяют в графы знаний. Графы знаний это некоторые сборники семантических сетей по их онтологиям. Онтология – это разделы графов знаний по классам (темам) с их иерархией. Например, семантическую сеть “Базис” - можно отнести в класс (тему) “Математика”, а “Математика” можно отнести к классу (теме) “Наука” и так далее (Здесь иерархия: Наука – Математика - Базис). Так же графов знаний в интернете существует некоторое множество, но они между собой связанны, то есть один граф может ссылаться на другой (так описано в [4],[14]).

Также существуют критерии оценивания классификатора, то есть алгоритма, который лежит в основе семантического анализа текста, а именно поговорим о F – мере (согласно статье [16]). Для определения F – меры введем два понятия. Понятие точности (precision) – показывают долю документов, которые точно относятся к данному классу, относительно тех документов, которые система определила к данному классу. А полнотой (recall) – будет доля отнесенных системой документов к данному классу относительно всех документов в системе. Тогда можно полноту и точность представить в виде:

1. Количество тех документов, которые точно относятся данному классу
2. Количество тех документов, которых система отнесла к данному классу
3. Количество всех документов в системе

, , ,

Тогда F – мера это отношение: , чем это число больше, тем точнее работает алгоритм.

Как вы уже заметили, в современных исследованиях по нахождению и определению семантических расстояний стараются использовать не только текстуальное сходство документов, но и пытаются задействовать лексикон доступной базы. Лексикон – это объединение всех термов в доступной базе документов. В связи с этим используют два подхода (согласно статье [15]): парадигматический и синтагматический.

Опишем парадигматический подход. Этот подход к измерению семантического расстояния основан на лексиконе, который составляется по значениям лексем (объединение термов) хранящихся внутри языкового знака, то есть внутри данного документа, а не за его пределами. А именно составления значения лексемы использованием других лексем в документе. Таким образом, данный подход определяет лексикон языка множествами лексем, которые между собой имеют отношения семантической близости. Описав значения лексем с помощью набора семантических признаков, можно создать словарь лексем. Полученный словарь лексем можно представлять в виде векторной n – мерной модели или в семантической сети, описание модели которой, приведено выше (см. например [8],[11],[12]).

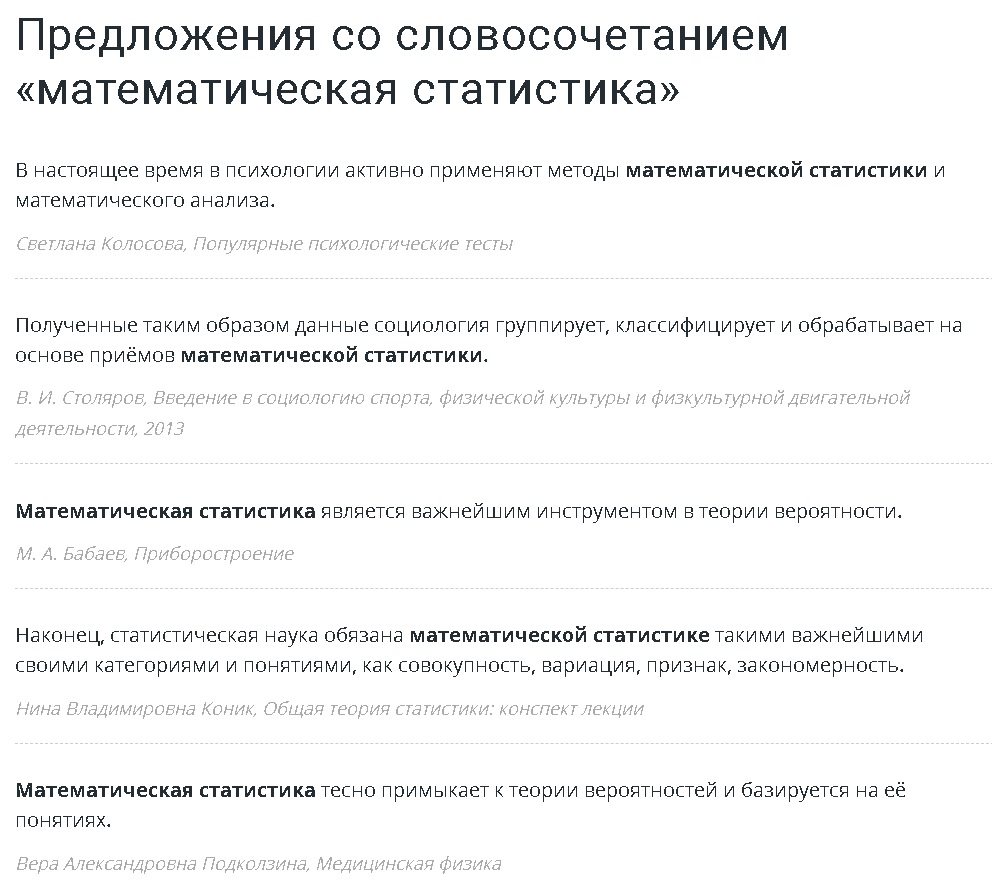
Также рассмотрим Синтагматический подход. Он основывается на вычислении и анализе семантических расстояний в данном документе и между другими, то есть внутри доступной базы документов. Использование значений лексем из других документов, позволяет определить или доопределить содержательную близость лексем, сопоставляя их синтагматические отношения (сочетаемость терма с другим термом). Например, в нескольких документах были употреблены следующие предложения (см. рис. 2).

Рис. 2 Предложения со словом “математическая статистика”.

Эти предложения, помогают доопределить значения и употребление данного терма “математическая статистика”, тем самым алгоритмы, в основе которых лежит синтагматический подход, могут отнести термину “математическая статистика” термин “наука”, ”метод”, ”теория вероятности”. Данный подход является основой метода Латентно-семантического анализа (см. статью [17]), в котором обращается внимание на отдельные слова и их связь между обобщенным контекстом (предложениями, кусками текста) в которых они встречаются. Это все помогает сформировать и определить группы близких по значению лексем (семантические сети) на основе их сочетаемости в контексте.  
 Совмещение этих двух подходов дает лучший результат, чем использование каждого в отдельности. Данные подходы породили современные методы определения семантических расстояний, такие как: метод шинглов, методы основанные на векторной модели представления документов, метод латентно-семантического анализа и другие. В следующей главе рассмотрим принципы работы каждого метода (как показано в [1],[2],[4],[5],[9],[14]).

# Раздел 3. Методы и их оценки

Начнем с раcсмотрения метода шинглов. Шинглы – это специальным образом выделенные фрагменты текста, которые записываются в виде последовательности термов, то есть обычные слова приводятся к своим нормальным формам (например, слова подвергаются стеммингу и лемматизации, удаляются союзы, предлоги и различные знаки препинания, которые не участвуют дальнейшем сравнении) и записываются в последовательности. Данный метод разбивает весь текст на группы шинглов одинаковой длины, причем количество шинглов стремится к количеству слов в документе. После этого у каждого шингла вычисляется его хеш – код. Хеш-код это строка из символов фиксированной длины полученная в результате алгоритма хеширования входных данных. Далее на основе хеш – кода для каждого шингла строится его контрольная сумма. В результате случайных выборок этих контрольных сумм между двумя документами сравнивается их значение, где в зависимости от значений контрольных сумм можно говорить о проведении лексического анализа и расчете семантического расстояния между двумя документами на предмет их частичного или даже полного совпадения.  
 Основной проблемой данного метода заключается его низкая скорость работы, так как в зависимости от увеличения размера данных документов дольше выполняется составление шинглов и их вычисление контрольных сумм, а следовательно производительность данного метода уменьшается от больших входных данных. Так же результат схожести двух документов в виде их семантического расстояния не является полностью правдивым, так как требует значительных совпадений контрольных сумм. Однако данный метод хорош для систем антиплагиата, а именно поиска дубликатов и копий текста, либо строгого поиска документов исходя из жестко поставленного запроса.

Следующий метод, который мы рассмотрим, это метод векторного представления документов и их термов. Он основан на частоте употребления конкретных терм в документе, а не последовательностей фрагментов, как описано выше в методе шинглов, и не зависит от их порядка. То есть принцип работы данного метода состоит из привидения всех слов, встречаемых в документе к терму по средствам стемминга и лемматизации. Удобно записать это в виде матрицы большого размера, где каждая строка это терм, который был получен из слова в каких либо документах. Столбцы же это некоторые различные документы. Тогда в ячейках матрицы можно записать некое число, которое будет показывать частоту его вхождений в документ. А каждую тему онтологии, можно охарактеризовать определенным набором термов с их частотами, тогда принадлежность документа теме можно высчитать. Сама матрица для большого количества документов будет разряженной, так как большое количество различных термов не будет присуще одному документу. Чтобы определить подобие двух конкретных документов нужно рассчитать семантическое расстояние между двумя столбцами матрицами - векторами. Это расстояние можно считать по разным формулам таким как: скалярное произведение векторов, евклидово расстояние, косинусная мера сходства и другие. На основе полученных расстояний можно будет определить число подобия друг другу, чем оно больше, тем более сходны документы.

Данный метод дает лучшие результаты на основе встречающихся терм в документах, нежели первый, который находит плагиаты двух документов. Его же недостаток заключается в большом количестве хранения информации, так как чем больше документов в базе, тем больший размер занимает полученная матрица.

Метод латентно-семантического анализа (см. например [17]) , расширяет предыдущий метод, так как вместо использования матриц терм-документ, подразумевает использование матриц терм – тема и тема – документ. Таким образом, повышается качество полученного далее семантического расстояния. Делается это примерно аналогичным способом, который описали выше. Разобьем алгоритм на этапы:

1. На первом этапе помимо приведения слов к термам, удаляются цитаты и вставки текстов.
2. На втором этапе составляется матрица из термов и документов(tf – idf) и удаляются термы, которые встречаются лишь в одних документах.
3. Делается разложение построенной матрицы методом SVD ( A= U\*S\*) (см, подробнее [18])
4. Выделяются строки матрицы U и столбцы матрицы V, которые по значению соответствуют наибольшим сингулярным числам в матрице S.
5. По полученным выборкам используем наилучшее приближение данной матрицы термов-документов, которое будет показывать основную структуру зависимостей, которое определялось через количество вхождений термов в данный документ (веса).

Тогда термы и документы будут представляться в виде векторов размерности q, где q – определяется предполагаемым количеством семантических тем. Расстояние же между документами можно вычислять по скалярному произведению векторов или другими способами, представленными в методе выше.

Достоинства данного метода заключаются в том, что мы уменьшаем размер хранимой информации, а так же получаем разбиение матрицы терм –документов по темам. Что приводит нас к более качественному результату при вычислении семантического расстояния. Однако данный метод влечет за собой большую сложность вычисления, чем в предыдущие, а так же менее качественные вычисления расстояния на малых количествах исходных данных в базе (документов и их текстов).

Для обобщенного понимания методов сравнения текста в источнике [1], а для рассмотрения оценок между семантически близкими термами в источнике [19].

# Раздел 4. Реализация

1) Вычисление семантического расстояния между документами в векторной модели представления данных терм – документ.

Пусть дан эталон документа q (запрос), нужно сравнить его с документами в базе (doc1, doc2, doc3) (см, рис. 3).

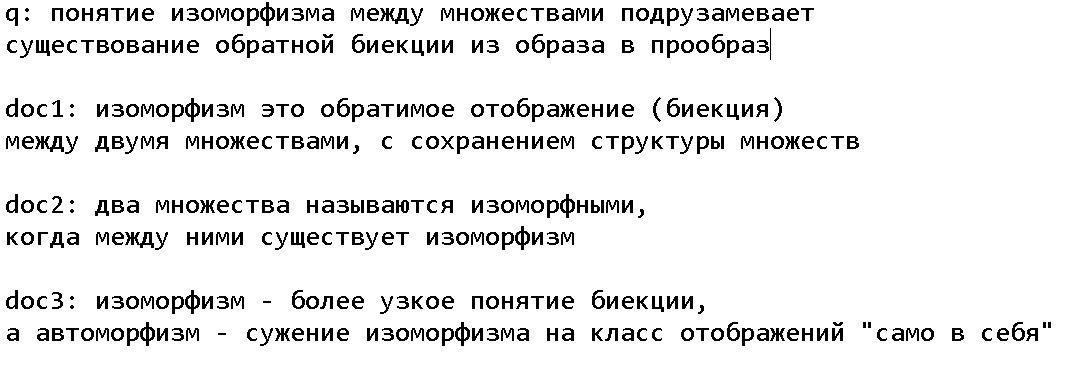


Рис. 3 Исходные данные для исследования.

С самого начала приводим все слова в документах к термам, по средствам стемминга и лемматизации, после чего подсчитываем сколько раз документ содержит термин (term frequency – tf ) (cм, рис. 4)

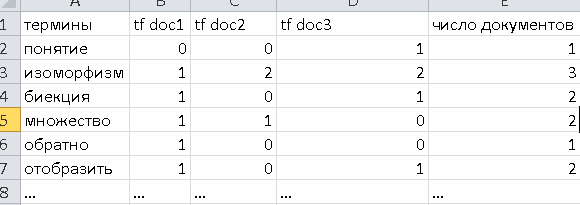


Рис. 4 Показатели числа документов (tf)

После чего рассчитаем idf – обратная документальная частота, фукнция от величины, обратной к количеству документов, в которой встречается терм

,

где n – количество документов,

а df – частота в коллекции или число документов, в которые входит терм.

Посчитаем вес конкретного терма t в документе doc, используя его частоту tf и idf по следующей формуле: (см рис. 5)

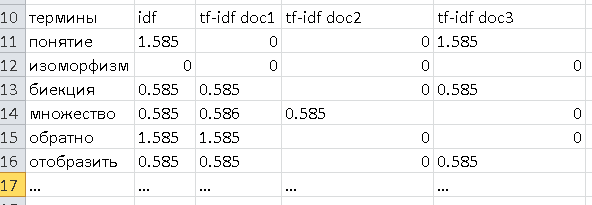


Рис.5 Векторы документов

Сам запрос (документ) q тоже приведем к векторному представлению таким же алгоритмом (см, рис. 6)

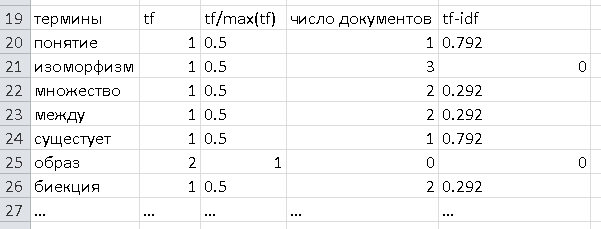


Рис. 6 Векторное представление документа q.

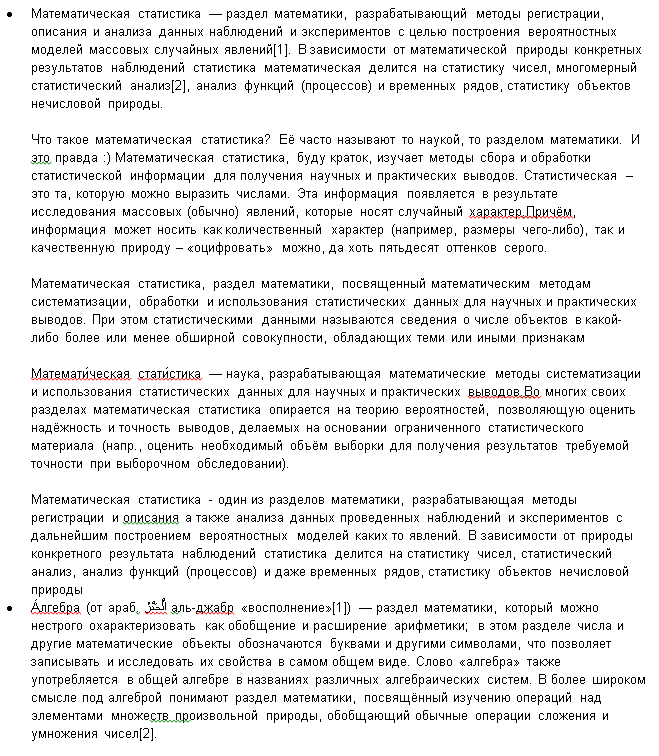
Посчитаем расстояние по косинусной мере сходства документов по следующей формуле:

Результаты:

На основе этого можно сказать, что документы q и doc1 семантически ближайшие документы.

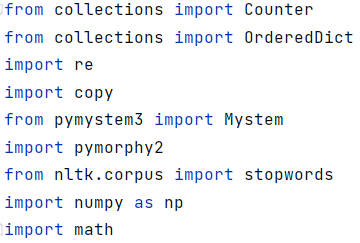
2) Код программа на python для вычисления расстояния между документами

В качестве документов были взяты определения математической статистики в разных источниках – Wikipedia, booksite, mathprofi, dic.academic; а алгебры – Wikipedia.

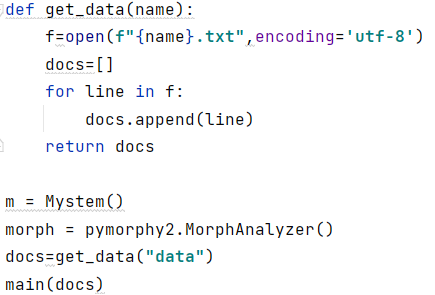


Код программы : (также можно посмотреть/скачать <https://github.com/taerd/SemanticTest>)

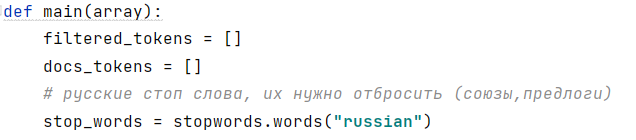
Заимствованные библиотеки:

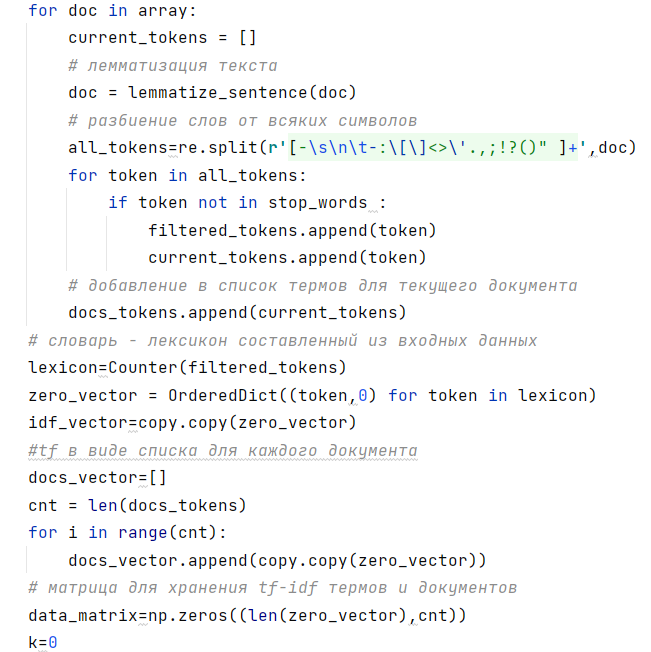


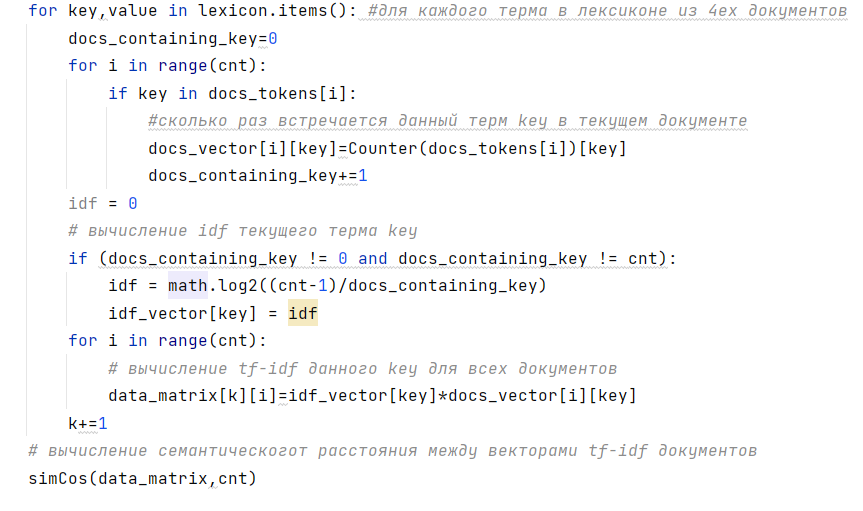
Начало программы, считывание текста из документа:



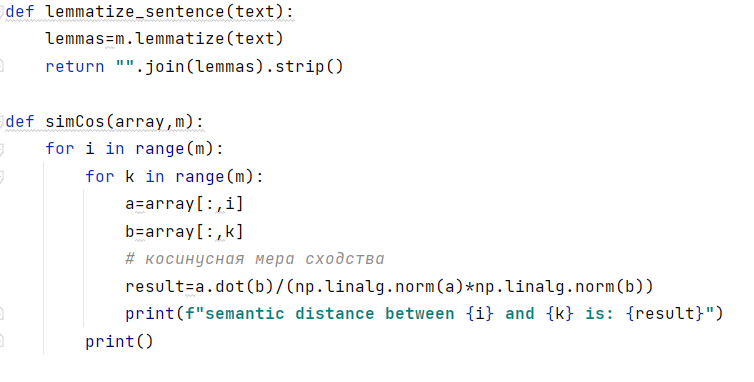
Выполнение основной части лемматизация, стемминг, создание словаря вычисление векторов tf-idf (веса терма относительно документа):



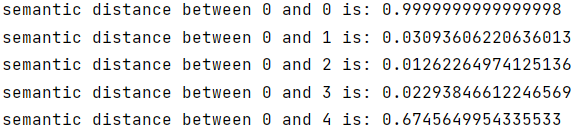




Функции лемматизации и вычисления косинусной меры сходства

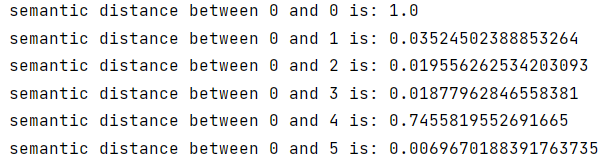


В первый раз запустил код, оставив текст только про математическую статистику. Будем учитывать результат сравнения первого текста с остальными.



Получили результат, что 1 текст схож с 5 (нумерация в python с 0)

Во второй раз запустил код, включая текст про алгебру.



Получили результат, что 1 текст схож с 5, причем результаты сходства с остальными текстами про математическую статистику улучшились.

# Заключение

В данной работе была поднята проблема автоматической обработки текста, потому что все, что человек когда-либо писал и загружал в интернет нужно как то осмыслять существующим программам и индексировать в своей базе, чтобы потом человеку можно легко было находить интересующую его информацию. Материалом для исследования послужили источники [1],[2],[4],[5],[6],[9],[10],[13],[14],[15]. На основе которых, были представлены основные задачи автоматической обработки текста, которые состоят в различии загруженных в базу данных документов и текстов, путем их неявного осмысления, за счет их контекстно смысловой загруженности, о чем как раз таки и изложено в этой работе. Для решения этой задачи были определены понятия семантического расстояния и семантического анализа, а также была представлена RDF модель, описывающая то, как хранится информация в виде слов, используя теорию графов, за счет вычисления семантического расстояния между словами. На основе предполагаемой хранимой информации в базе, описали существующие два подхода, которые сформировались после российских и мировых исследований, а именно парадигматический подход и синтагматический подход по вычислению семантического расстояния между документами. Эти два подхода породили современные методы вычисления семантического расстояния между документами, где принципы работы этих методов были описаны в главе 3, там же были указаны их особенности и приведена оценка. На основе некоторых методов реализовал примеры с вычислениями семантического расстояния.

Подытоживая, можно сказать: Дальнейшее изучение данной задачи и совершенствование методов выявления семантической близости документов улучшит результаты поиска необходимой человеку информации.

# Литература

1. Ингерсолл Г.С. Обработка неструктурированных текстов / Г.С. Ингерсолл, Т.С. Мортон, Э.Л. Фэррис. – ДМК-Пресс 2015 – 416с.
2. Xie I. Discover digital libraries: Theory and practice / I.Xie, K.Matusiak– Elsevier, 2016.
3. Беларусское агенство интернет-маркетинга CROPAS:  
   Семантический анализ.[<https://cropas.by/seo-slovar/semanticheskij-analiz/>].(дата обращения 18.12.2020)
4. Маннинг. К.Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце – М: ООО «И.Д. Вильямс», 2011. – 528 с.
5. Бермудес С. Х. Г. Метод измерения семантического сходства текстовых документов / С. Х. Г. Бермудес //Известия ЮФУ. Технические науки. – 2017. – №. 3. – С. 17-29.
6. Mohammad.S. Measuring Semantic Distance [<https://saifmohammad.com/WebDocs/Mohammad_Saif_Thesis-slides.pdf>] -Department of Computer Science University of Toronto .(дата обращения 18.12.2020)
7. Hirst.G. Semantic distance [<https://www.cs.toronto.edu/~gh/research-pages/research-semantic-distance.html>] - Professor of Computational Linguistics University of Toronto, Department of Computer Science. (дата обращения 18.12.2020)
8. Machinelearning.ru : Профессиональный информационно-аналитический ресурс.  
   Векторная модель. [<http://www.machinelearning.ru/wiki/index.php?title=%D0%92%D0%B5%D0%BA%D1%82%D0%BE%D1%80%D0%BD%D0%B0%D1%8F_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8C>]. (дата обращения 18.12.2020)
9. Сторожук Н. О. Анализ методов определения текстовой близости документов / Н. О.Сторожук, И. А. Коломойцева // Материалы студенческой секции IX Международной научно-технической конференции «Информатика, управляющие системы, математическое и компьютерное моделирование»(ИУСМКМ-2018) - Донецк: ДонНТУ. – 2018. – С. 43-47.
10. Kohlhase M. Co-Representing Structure and Meaning of Mathematical Documents / M.Kohlhase, M.Iancu //Sprache und Datenverarbeitung, International Journal for Language Data Processing. – 2014. – Т. 38.
11. Sysblock.ru : Системный Блок{Ъ} .  
    Семантические сети: как представить значения слов в виде графа.  
    [<https://sysblok.ru/nlp/semanticheskie-seti-kak-predstavit-znachenija-slov-v-vide-grafa/>].(дата обращения 18.12.2020)
12. Wikipedia.org : Википедия свободная энциклопедия.  
    Семантическая сеть.  
    [<https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BC%D0%B0%D0%BD%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D1%81%D0%B5%D1%82%D1%8C>].(дата обращения 18.12.20202)
13. Андреичев М. Д. Разработка программного комплекса генерации вопросов по заданным субъектам при помощи семантической сети / М. Д .Андреичев, А. А.Ференец //Электронные библиотеки. – 2019. – Т. 22. – №. 2. – С. 68-94.
14. Skiena. S.S. The Algorithm Design Manual / S.S.Skiena// Springer-Verlag London 2008.
15. Studbooks.net : Студенческая библиотека онлайн.  
    Семантическое расстояние. [<https://studbooks.net/2250838/informatika/issledovanie_osnovnyh_podhodov_k_avtomaticheskomu_otobrazheniyu_ontologiy>]. (дата обращения 18.12.2020).
16. Баженов.Д. Оценка классификатора (точность, полнота, F-мера)  
    [ <http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html> ]. (дата обращения 08.05.2021).
17. Wikipedia.org : Википедия свободная энциклопедия.  
    Латентно-семантический анализ. [<https://ru.wikipedia.org/wiki/%D0%9B%D0%B0%D1%82%D0%B5%D0%BD%D1%82%D0%BD%D0%BE-%D1%81%D0%B5%D0%BC%D0%B0%D0%BD%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7>].(дата обращения 18.12.2020)
18. Wikipedia.org : Википедия свободная энциклопедия.  
    Сингулярное разложение.  
    [<https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D0%B3%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D0%B7%D0%BB%D0%BE%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5>].(дата обращения 18.12.2020)
19. Личаргин Д.В. Оценка расстояния в семантическом и грамматическом пространстве между отдельными языковыми единицами и фрагментами текстов / Д.В. Личаргин, О.С. Полякова, А.О.Подлесный, М.В. Кравченко // Современные проблемы науки и образования. – 2014. – № 6. [<http://science-education.ru/ru/article/view?id=15802>].(дата обращения: 18.12.2020).