**Построение математической модели текстовых документов**

Работу выполнила:

Васильева Алина, ученица 10В класса

ГБОУ Академическая Гимназия №56

г.Санкт-Петербург

 (812) 346-00-87,  school56.spb@mail.ru

Научный руководитель:

Черкай Михаил Васильевич,

Учитель математики

Санкт-Петербург

2021г.

**Оглавление**

1. Введение

2. Постановка задачи

3. Математическая модель

4. Численные эксперименты

5. Выводы

***1.Введение***

XXI век - век прогресса, век бескрайних технологических возможностей и век информации, в первую очередь. Информация - это не просто сведения, в наше время, информация - это ключевой способ взаимосвязи и деятельности, ведь недаром современное общество так и называется - «информационное». В конце 20 века, с помощью разработок американских военных ученых из простой локальной сети между компьютерами появляется всемирная глобальная паутина WWW. В связи с этим, начинает развиваться и наука, появляется такое направление в информационной области, как Информационный поиск (Information retrieval)[5].

На идею этой работы нас натолкнуло стремительное развитие такой интегрированной дисциплины, как математическая лингвистика. Конкретнее, нас заинтересовал вопрос о вычислении релевантности информации, а именно, как ответить на вопрос какой документ ближе соответствует исходному документу шаблону. Естественно, сейчас, существуют формулы и алгоритмы, по которым подбирается информация, наиболее близкая исходному документу тексту запросу: например, ВМ25 (от. англ. *best match - лучшее соответствие*), TF-IDF [4] (term frequency - частота термина; inverse document frequency - обратная частота документа) и другие коэффициенты сходства[6]. Мы решили рассмотреть один из таких методов: сузить проблему до бытового применения и создать математическую модель, опираясь на опыт Клода Шеннона, Карен Спарк Джонс, Стивена Робертсона и других выдающихся ученых, создавших наиболее распространенные меры близости текстов [2].

***2. Постановка задачи***

Пусть существует некоторая коллекция текстовых документов, условно близких по содержанию (например, сочинения по литературе учеников одного класса на заданную тему, причем это могут быть документы как на русском языке, так и на английском языке. Например, можно рассмотреть одно из заданий экзамена ЕГЭ по английскому языку, где пишут сочинения), и шаблон - тот документ, с которым нужно сравнить остальные (например, сочинение отличника или исходное сочинение некоторого эксперта). В данной ситуации целью будет сравнить работы всего класса с работой (эталоном) отличника, а в общем виде **цель проекта**: создание алгоритма определения релевантности текста к образцу с помощью математической модели. Для выполнения данной цели, мы сформулируем следующие **задачи**:

1. Построить математическую модель коллекции предложенных документов
2. Оцифровать коллекцию реальных текстовых документов (30 сочинений 11 класса на английском языке)
3. Частично реализовать данную модель на языке Python

 Актуальность работы обусловлена тем, как можно экономить силы и время на выполнении однотипной ручной работы с помощью программы, и тем, с какой точностью «математически» можно проанализировать текстовые документы. Итак, перейдем к математической модели.

1. ***Математическая модель***

Пусть наши оцифрованные документы представляют данную коллекцию, также есть некоторое “идеальное” сочинение (например, его мог составить эксперт). Далее будем его называть шаблоном. Итак, пусть множество $A=\{A\_{1,}A\_{2},…,A\_{n}\}$, множество документов, каждому из которых соответствует свой номер n. Документ шаблон – $B$.

Для реализации математической модели, мы выбрали уже известную модель – VSM (Vector Space Model) – модель векторного пространства[1]. Основной идеей в ней является, то что, каждому документу сопоставляется вектор, состоящий из 0 и 1, где размерность всех векторов – это количество слов в словаре коллекции, а маркер 0 или 1 означает отсутствие или наличие данного слова в данном документе. Построение математической модели состоит из следующих шагов:

1. Создание словаря коллекции – “мешка слов”
2. Построение векторов каждого документа и вектора шаблона
3. Вычисление расстояния (численной характеристики) между каждым вектором и шаблоном
4. Ранжирования данных и определение наиболее релевантного документа заранее заданному шаблону

Таким образом, пусть a1, a2, …, am – все слова из словаря коллекции (“мешок слов”), тогда всей коллекции документов будет ставиться в соответствие матрица следующих векторов:

Здесь числа Sij= {0,1} – обозначают, присутствует ли данное слово из словаря в коллекции в соответствующем документе или нет. Очевидно, что вектор документа – шаблона В также будет содержать 0 и 1 на соответствующих местах термов.

Следующим шагом модели является введение численной характеристики сравнения двух векторов (документов), для этого мы остановились на простом способе – вычисления косинуса угла между векторами с помощью скалярного произведения. Естественно, что чем больше косинус угла между векторами, тем меньше угол между ними в пространстве, а значит, они более соответствуют друг другу.



$$\frac{\sum\_{i=1}^{m}x\_{i}y\_{i}}{\sqrt{\sum\_{i=1}^{m}x\_{i}^{2}}\sqrt{\sum\_{i=1}^{m}y\_{i}^{2}}}$$

На следующем шаге, каждой паре векторов (Ai,B) ставится в соответствие $ω$ – косинус угла между ними:

…

ω1

ω2

ωn

…

Далее ранжируем по убыванию и выявляем наибольший косинус, ему и соответствует наиболее релевантный документ.

1. ωmax

…

n. ωmin

-наиболее релевантный

-наименее релевантный

***4.Численные эксперименты***

Реализацию численных экспериментов мы разбили на несколько этапов:

1. Поиск коллекции англоязычных документов (сочинений учеников 11 класса)
2. Построения словаря коллекции
3. Подсчет расстояния между парами и ранжирование данных документов

Понятно, что данные задачи требуют хорошего знания одного или нескольких языков программирования, с помощью которых можно реализовать скрипты, которые разбирают коллекцию на словарь, сравнивают документы. Поэтому на данном этапе мы смогли реализовать первый этап – то есть программу, которая получает на входе файл (англоязычный документ), на выходе же дает мешок слов данного документа. Можно сказать, что данная работа является началом нашего исследования в этой области и целью ее было построение математической модели[3], то с заданной задачей мы справились обозначив алгоритм, который реализует выявление наиболее релевантного документа данному шаблону. Ниже предлагается листинг данной программы на языке Python:

import re

import string

frequency = {}

document\_text = open('test.txt', 'r')

text\_string = document\_text.read().lower()

match\_pattern = re.findall(r'\b[a-z]{3,15}\b', text\_string)

 for word in match\_pattern:

 count = frequency.get(word,0)

 frequency[word] = count + 1

 frequency\_list = frequency.keys()

for words in frequency\_list:

 print (words, frequency[words])

***5. Выводы***

Нам удалось реализовать поставленные задачи по теме в рамках школьной исследовательской работы. Однако на этом «глобальная» цель не достигнута, что может стать толчком к развитию и продолжению данного проекта на будущий год. Мы верим, что подобные алгоритмы могут быть полезными в различных бытовых или полубытовых ситуациях, как например наша - проверка классной/домашней/любой другой письменной работы. Пусть вопрос автоматизации труда стоит достаточно остро в наши дни, невозможно отрицать то, что алгоритмы, какие бы они ни были, экономят ресурсы (время, человеческие силы) и минимизруют вероятность ошибки (человеческий фактор).

**Список используемой литературы:**

[1]Автор неизвестен. Векторная модель [Электроннный ресурс]:

http://icybcluster.org.ua:34145/technology-documents/vector\_model\_rus.pdf

[2]Дж Солтон. Динамические библиотечно-поисковые системы. М.:— Мир, 1979.

[3]Ландэ Д. В., Снарский А. А., Безсуднов И. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. — M.: Либроком (Editorial URSS), 2009. — 264 с. — ISBN 978-5-397-00497-8.

[4]Оськина К.А. Оптимизация метода классификации текстов, основанного на tf-idf, за счет введения дополнительных коэффициентов// Вестник МГЛУ. — 2016.

[5]Перевод: Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011. — ISBN 978-5-8459-1623-5.

[6]Федоровский А.Н, Костин М. Ю. Mail.ru на РОМИП-2005 // в сб. «Труды РОМИП’2005» Труды третьего российского семинара по оценке методов информационного поиска. Под ред. И. С. Некрестьянова, стр. 106—124, Санкт-Петербург: НИИ Химии СПбГУ, 2005