

Московский государственный технический университет
имени Н.Э. Баумана

Факультет «Информатика и системы управления»
Кафедра «Компьютерные системы и сети»

Е.В. Смирнова, К.Н. Ле, Т.Х. Нгуен

**ПРОВЕДЕНИЕ
СЕМИНАРСКИХ ЗАНЯТИЙ ПО
ДИСЦИПЛИНЕ «АЛГОРИТМИЧЕСКАЯ
ТЕОРИЯ ИНФОРМАЦИИ В
БИОМЕДИЦИНСКИХ СИСТЕМАХ»**

Учебно-методическое пособие

Москва

Издательство МГТУ им. Н.Э. Баумана

2025

УДК 303.732
ББК 32.972.11
Б

Издание доступно в электронном виде по адресу

Факультет «Информатика и системы управления»
Кафедра «Компьютерные системы и сети»

*Рекомендовано Научно-методическим советом
МГТУ им. Н.Э. Баумана в качестве учебно-методического пособия*

Авторы:

Е.В. Смирнова, К.Н. Ле, Т.Х. Нгуен

Проведение семинарских занятий по дисциплине «Алгоритмическая теория информации в биомедицинских системах»: учебно-методическое пособие / Е.В. Смирнова, К.Н. Ле, Т.Х. Нгуен. – Москва: Издательство МГТУ им. Н.Э. Баумана, 2025. – 85 с.

Учебно-методическое пособие является руководством для проведения семинарских занятий по дисциплине «Алгоритмическая теория информации в биомедицинских системах». Семинары охватывают разделы, связанные с теоретическими аспектами и практическими применениями алгоритмической теории информации в биомедицинских системах, включая методы анализа, кодирования и сжатия данных.

Издание предназначено для студентов МГТУ имени Н.Э. Баумана, обучающихся по направлению подготовки 12.03.04 «Биотехнические системы и технологии». Может быть также полезно студентам и аспирантам других специальностей и направлений подготовки, которые интересуются вопросами алгоритмической теории информации в биомедицинских системах.

УДК 004.9
ББК 32.972.11

СОДЕРЖАНИЕ

Предисловие.....	4
Введение.....	6
Семинарское занятие №1 Мешок слов	8
Семинарское занятие №2 Вычисление меры схожести на основе колмогоровской сложности.....	14
Семинарское занятие №3 Статистические меры: частота термина и обратная частота	25
Семинарское занятие №4 Передача информации в биосистемах: теория информации Шеннона	35
Заключение	48
Литература	49
Приложение к семинарскому занятию № 1.....	50
Приложение к семинарскому занятию № 3.....	80
Приложение к семинарскому занятию № 4.....	83

Предисловие

Учебно-методическое пособие предназначено для студентов, обучающихся по направлению подготовки 12.03.04 «Биотехнические системы и технологии», в учебных планах которых предусмотрено изучение дисциплины «Алгоритмическая теория информации в биомедицинских системах». Может быть также полезно для студентов и магистрантов других специальностей и направлений подготовки, изучающих вопросы, связанные с теорией информации.

Общей целью проведения семинарских занятий является углубленное изучение и практическое освоение студентами ключевых концепций алгоритмической теории информации А.Н. Колмогорова в контексте биомедицинских систем. В рамках семинаров рассмотрены методы «мешка слов», вычисления мер схожести на основе колмогоровской сложности NGD (Normalized Google Distance) статистические меры, включая частоту термина и обратную частоту документа TF-IDF (Term Frequency – Inverse Document Frequency). Дополнительно семинары посвящены передаче информации, включая пропускную способность биологических каналов связи (зрительный, аудио-речевой, тактильный), классическому блочному кодированию, разработанному «отцом теории информации» Клодом Шенноном, а также исследование о кратковременной человеческой памяти психолога Дж. Миллера.

Представленный учебно-методический материал соответствует тематике семинаров, приведенной в программе дисциплины «Алгоритмическая теория информации в биомедицинских системах», которая также приведена в Приложении к этому пособию.

Учебно-методическое пособие содержит описание четырех семинарских занятий, посвященных следующим темам: метод «Мешок слов»; вычисление меры схожести на основе колмогоровской сложности; статистические меры:

частота термина и обратная частота; передача информации в биосистемах – описание на основе теории информации К. Шеннона.

Описание каждого семинара включает теоретическую и практическую части. Теоретический материал дает необходимые знания для понимания темы семинара. Практическая часть включает задание на исследование и требования к выполнению задания студентом. После каждой темы приведены вопросы для самоконтроля студентом полученных знаний.

Для проведения практических занятий необходимо наличие компьютерного класса, оборудованного современной вычислительной техникой, из расчёта одного рабочего места на одного студента. При выполнении заданий предусмотрено освоение графического пакета (например, Edraw Max) и системы моделирования (например, Anaconda, дистрибутив Python).

Введение

В современных биомедицинских системах важное место занимает обработка и анализ информации, связанной с функционированием биологических объектов, диагностикой заболеваний, прогнозированием состояния пациентов и управлением здравоохранительными процессами. Эффективность таких биомедицинских систем во многом определяется способностью извлекать, интерпретировать и использовать информацию, содержащуюся в разнообразных данных — от текстовых описаний клинических случаев до сигналов биомедицинских датчиков и даже до обработки геномных последовательностей.

Алгоритмическая теория информации предоставляет мощный математический аппарат для анализа сложности и структуры информации, что находит широкое применение в задачах биомедицинской информатики. Основные понятия этой теории позволяют формализовать представление о содержательности и мере схожести между различными объектами, что особенно важно при анализе неструктурированных данных, таких как медицинские тексты, изображения или временные ряды физиологических параметров.

Одним из ключевых методов обработки текстовой информации в биомедицине является «Мешок слов», который позволяет преобразовать текстовые данные в числовые векторы и использовать их в алгоритмах машинного обучения (тема первого семинара). Для оценки информативности терминов применяются статистические меры, в пособии они рассмотрены на примере частот терминов TF (Term Frequency) и обратной документальной частоты IDF (Inverse Document Frequency), которые учитывают значимость слова как внутри конкретного документа, так и в общем корпусе текстов (тема второго семинара). Еще более глубокий уровень анализа обеспечивает использование алгоритмической меры схожести, основанной на концепции

колмогоровской сложности. Этот подход позволяет сравнивать объекты по степени их алгоритмической близости, минуя необходимость явного задания признакового пространства. Использование алгоритмической теории открывает возможности для выявления скрытых связей между медицинскими понятиями, симптомами и диагнозами (тема третьего семинара).

Теория передачи информации, начиная с классических работ Клода Шеннона и продолжая современными исследованиями Дж. Миллера, также находит применение в моделировании информационных процессов в биосистемах, поэтому в этом пособии уделено внимание и этим работам (тема четвертого семинара). Классические подходы зарубежных ученых Шеннона и Миллера, а также российских ученых Колмогорова А.Н., Маркова А.А., и других обеспечивают студентов базовыми знаниями об обработке, анализе, передаче информации на разных уровнях — от молекулярного до системного.

Алгоритмическая теория информации в настоящее время становится неотъемлемой частью современных биомедицинских исследований, обеспечивая строгую основу для анализа сложных биомедицинских данных и построения эффективных систем поддержки принятия решений.

Авторы отмечают, что это учебно-методическое пособие не охватывает полной картины состояния алгоритмической теории информации на момент его издания, оно лишь описывает методику проведения семинарских занятий.

Семинарское занятие №1 «Мешок слов»

Цели занятия:

- познакомиться с понятием «мешок букв» (Bag of letters) и его применением в анализе текста [6];
- изучить основные вероятностные параметры букв в тексте, такие как частота, энтропия, и их роль в измерении количества информации в тексте;
- освоить практические навыки определения вероятностных параметров букв в тексте с помощью инструментов обработки данных программных библиотек (например: регулярное выражение, Pandas...).

План занятия

1. Краткое повторение роли вероятностных параметров букв для измерения содержащейся в тексте информации, мысли Колмогорова А.Н. и цепи Маркова А.А. – 20 мин, для этого использовать слайд-презентации [1, 2].
2. Задача 1. Нулевое приближение (подробнее см. ниже) – 10 мин.
3. Задача 2. Первое приближение (подробнее см. ниже) – 10 мин.
4. Задача 3. Второе приближение (подробнее см. ниже) – 10 мин.
5. Задача 4. Третье приближение (подробнее см. ниже) – 10 мин.
6. Задача 5. Четвёртое приближение (подробнее см. ниже) – 10 мин.
7. Сделать вывод о роли вероятностных параметров слов для измерения содержащейся в тексте информации – 20 мин.

Разбор решения задачи 1.1

Преподаватель объясняет студентам протокол выполнения задачи 1.1 «Вспомним инвентарь из 32 букв русской письменной речи, включая пробел. Представим, что мы составили разрезную азбуку из этих 32 букв и поместили её в ящик (математики сказали бы «в урну»), тщательно перемешав. Будем теперь составлять из этой азбуки случайный текст, применяя следующую процедуру: мы *вынимаем букву из ящика, записываем её, затем возвращаем в ящик, перемешиваем буквы, снова вынимаем букву, снова записываем* (приписывая её к уже имеющемуся тексту), *снова возвращаем, снова перемешиваем, снова вынимаем и т. д.* В результате мы получим что-нибудь вроде:

**СУХЕРРОВАДЦАХВЩИЮАЙЖТЛФВНЗАГФОВЕНВШ
ТЦРПХГБКУЧ ТЖЮРЯПЧЬКЙХРЫС** » взято из [2].

Примечания:

- Ящик / мешок букв уже должен быть готов заранее с помощью компьютерной программы на любом языке. Но можно провести этот семинар с настоящим ящиком букв, студенты с удовольствием играют в эту наглядную игру.
- Можно предложить студентам самостоятельно обсудить получившийся текст: насколько он похож на текст, написанный человеком? Что можно сказать о наличии или отсутствии пробелов между словами? Как это влияет на читаемость и восприятие текста?

Про полученный текст можно сказать лишь, что он составлен из русских букв, но на русскую письменную речь он не похож: мы говорим, конечно, не об осмысленности (где уж!), а лишь о внешней похожести. Дело в том, что в нашем эксперименте все **буквы** были **равновероятны**, и поэтому в полученном тексте встречались примерно с одинаковой частотой ($1/32$). В реальных же русских письменных текстах пробел и различные буквы

встречаются с различными частотами и потому ожидаются с различными вероятностями. Это, конечно, всем известный факт. Но на этом семинаре мы со студентами этот факт проверяем.

Менее известен (хотя и очевиден) и потому будет далее на семинаре воспроизведён следующий эффект: при учёте всё более и более глубоких статистических закономерностей, имеющих в реальных текстах, экспериментальный искусственный текст делается всё более и более похожим на настоящий. Тот искусственный текст, который мы получили выше, можно назвать приближением нулевого порядка к реальному тексту: здесь учитывается лишь состав алфавита и не учитываются статистические характеристики. Далее на этом семинаре переходим к их учёту.

Разбор решения задачи 1.2

При приближении первого порядка учитываются частоты каждой из букв; иными словами, теперь требуется, чтобы в экспериментально построенном искусственном тексте буквы встречались с такими же (в идеале) частотами, как и в реальных текстах.

Параметры в Приложениях 3 - 6 рассчитаны с использованием программ, аналогичных примерной программе в Приложении 1, по образцу текста в Приложении 2 к семинару №1. Таблица Б1 частоты встречаемости отдельных букв приведена в Приложении 3 к семинарскому занятию № 1.

Тогда мы получаем, например, такой текст:

ЕЫНТ ЦИЯЪА ОЕРВ ОДНГ БУЕМЛОЛИЙК ЗБЯ ЕНВТША

Он уже чуть-чуть похож на настоящий: и длина слов нормальная, и нет того чудовищного преобладания согласных, как в тексте нулевого приближения. Можно спросить у студентов «Что еще вы видите в этом

искусственно полученном тексте (может ли русское слово начинаться с мягкого знака?)».

Слова «**в идеале**» отражают философское различие между частотой появления букв в текст и теоретической вероятностью появления каждой буквы. Частоты букв, встречающихся в реальных текстах (особенно в длинных), формируют у исследователя представление о вероятностях появления этих букв. Эти эмпирические частоты и принимаются за теоретические вероятности. Затем данные вероятности используются в модели для генерации искусственного текста. После этого можно с высокой степенью уверенности ожидать, что частоты букв в созданном искусственном тексте будут близки к исходным вероятностям, взятым из реальных текстов.

Разбор решения задачи 1.3

Приближение первого порядка не учитывает частот диграмм, то есть, сочетаний двух последовательно идущих букв. В приведённом тексте, например, встречаются диграммы **ЯЬ**, **ЬА** и **ЬУ**, частота которых в реальных текстах равна нулю. Учёт частот диграмм приводит к *приближению второго порядка*:

УМАРОНО КАЧ ВСВАННЫЙ РОСЯ НЫХ КОВКРОВ НЕДАРЕ

Частоты встречаемости отдельных цепочек из 2-х букв приведены в Приложении 4 к семинарскому занятию № 1.

Разбор решения задачи 1.4

Приближение третьего порядка учитывает частоты триграмм, то есть, здесь трёхбуквенные сочетания должны встречаться примерно с теми же частотами, что и в реальных текстах:

ПОКАК ПОТ ДУРНОСКАКА НАКОНЕПНО ЗНЕ СТВОЛОВИЛ СЕ ТВОЙ ОБНИЛЬ

Частоты встречаемости отдельных цепочек из 3-х букв приведены в Приложении 5 к семинарскому занятию № 1.

Разбор решения задачи 1.5

Приближение четвёртого порядка, учитывает частоты тетраграмм:

ВЕСЕЛ ВРАТЬСЯ НЕ СУХОМ И НЕПО И КОРКО

Частоты встречаемости отдельных цепочек из 4-х букв приведена в Приложении 6 к семинарскому занятию № 1.

Замечание: если нет возможности подготовить полные материалы для семинара, можно использовать примерную программу в приложении 7 к семинару №1, чтобы студен могли выполнить упражнения самостоятельно.

Заключение по результатам проведения семинара

В заключение семинара преподаватель повторяет выводы о роли вероятностных параметров слов для измерения содержащейся в тексте информации:

- понять, как вероятностные параметры букв могут использоваться для решения задач классификации, кластеризации, а также для измерения сложности текста;
- развить навыки анализа и интерпретации результатов, полученных при анализе вероятностных параметров букв;

- познакомиться с основными принципами теории информации и ее применением в обработке текста;
- развивать навыки работы с текстовыми данными и их анализа;
- стимулировать интерес к изучению естественного языка и его обработки.

Вопросы для самоконтроля полученных знаний у студентов

1. Объясните, как «мешок букв» отличается от «мешка слов» и какие преимущества и недостатки есть у каждого подхода?

Цель вопроса – Это вопрос на понимание концепции «мешка букв» и ее сравнение с «мешком слов». Студенты должны показать, что они понимают, как эти модели работают, и какие задачи они решают.

2. Представьте, что вы анализируете два текста: один - научная статья, другой - популярная песня. Как вероятностные параметры букв могут отличаться в этих текстах, и что говорит об их «информативности»? Какие характеристики текста могут отличать один текст от другого?

Цель вопроса – Этот вопрос проверяет способность студентов связать «мешок букв» с реальными текстами и анализировать вероятностные параметры в контексте. Студенты должны сформулировать гипотезы о том, какие параметры будут отличаться и какой смысл это имеет.

3. Как можно использовать «мешок букв» для классификации текстов по жанрам? Приведите пример конкретного алгоритма и опишите, как вероятностные параметры букв будут использоваться в этом алгоритме.

Цель вопроса – Этот вопрос требует от студентов применения знаний о «мешке букв» для решения конкретной задачи. Студенты должны продемонстрировать понимание алгоритмов классификации и способность применить вероятностные параметры букв в их рамках.

Семинарское занятие № 2

«Вычисление меры схожести на основе Колмогоровской сложности»

Цели занятия:

- ознакомить студентов с концепцией колмогоровской сложности и её применением в вычислении меры схожести;
- изучить нормализованное информационное расстояние (Normalized Google Distance, NGD) и его использование в анализе данных;
- развить навыки решения практических задач с использованием NGD в контексте биомедицинских систем;
- способствовать критическому мышлению и обсуждению среди студентов по вопросам применения алгоритмической теории информации.

План занятия

1. Введение в тему – 10 мин.
2. Теоретическая часть – 20 мин.
3. Практическая часть – 30 мин.
4. Обсуждение результатов – 20 мин.
5. Заключение семинара – 10 мин.

Теоретический минимум

«Колмогоровская сложность» строки x (обозначается как $K(x)$) — это длина самой короткой программы p , которая выводит x на универсальной машине Тьюринга [6]. Если строка может быть сжата до более короткой программы, это указывает на то, что она может содержать более структурированную информацию. В противном случае, если строка не может

быть сжата, она считается случайной. Колмогоровская сложность связана с понятием энтропии в теории информации. Строки с высокой энтропией имеют высокую сложность, а строки с низкой энтропией могут быть описаны короче. Колмогоровская сложность не является вычислимой функцией, но при этом, это асимптотически приближающаяся к определенному пределу функция. Это означает, что нет общего алгоритма (метода, формулы, и т.д.), который мог бы определить сложность любого объекта. Колмогоровская сложность отличается от других мер сложности, таких как длина строки или количество уникальных символов, поскольку она учитывает возможности сжатия и структурирования данных. Эти аспекты делают колмогоровскую сложность важным инструментом для анализа информации и понимания структур данных. Отметим здесь, что Нобелевскую премию по физике 2024 года получили ученые Джон Хопфилд и Джеффри Хинтон, которые применили нейросетевые методы, основанные на принципах статистической физики и теории информации, включая концепции, связанные с мерами сложности и энтропией, для понимания работы мозга и разработки машинного обучения. Хотя они не использовали напрямую формулу Колмогорова-Арнольда их работы тесно связаны с фундаментальными вопросами о сложности данных и информации. В 2024 году Нобелевская премия по физике была частично присуждена за развитие нейросетевых технологий, которые теоретически связаны с теоремой Колмогорова-Арнольда о суперпозиции функций, лежащей в основе универсальных аппроксиматоров.

Меры близости, использующие Google-семантику. Мотивация подхода к измерению семантической близости слов и текстов достаточно проста: неструктурированное множество веб-документов есть практически универсальная выборка текстов, а потому она может быть использована для поиска скрытых смысловых связей и семантической близости так же, как и другие множества документов, на которые опираются примеры решения таких задач на первом семинаре. По мнению исследователей в области

вычислительной лингвистики и информационного поиска, слова, близкие по смыслу, должны приводить к похожим результатам поиска. Поэтому даже такой параметр, как «число документов, содержащих оба слова и которые найдены поисковой машиной (Google-машиной) для двух слов», может использоваться в качестве аргумента функции, описывающей **меру их семантической или смысловой близости**. Поскольку, чем больше относительное число таких общих страниц для пары слов и/или для пары текстов, тем они ближе по смыслу. Итоговое выражение для метрики, которую называют нормализованным Google-расстоянием имеет следующий вид:

$$NGD(x, y) = \frac{\max \{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min \{\log f(x), \log f(y)\}},$$

где x и y – это исследуемые слова или термины, для которых вычисляется степень семантической близости; $f(x)$ и $f(y)$ – число страниц, возвращаемых Google по запросам x и y , которые содержат x и y , соответственно; $f(x, y)$ – это число страниц, возвращаемых Google по запросу xy , в которых содержатся обе цепочки; $NGD(x, y)$ – семантическая близость.

Нормализованное расстояние Google (NGD) и колмогоровская сложность (КС) связаны через концепцию измерения информации и схожести. Параметр NGD оценивает семантическую близость между терминами, в то время как колмогоровская сложность определяет минимальную длину программы, генерирующей данные, что также отражает их структурную сложность.

Поясним принцип работы сначала на примере. Хотя теория, используемая в этом методе более сложная, метод достаточно прост.

Пример 2.1. В момент описываемого эксперимента поисковик Google имел число индексированных веб-страниц, равное $N = 8,058,044,651$. Поиск слова «horse» вернул 46,700,000 ссылок на найденные страницы. Число

ссылок по слову «rider» - 12,200,000. Поиск страниц, на которых есть оба эти слова «horse» и «rider» дал 2,630,000 ссылок. Используя эти числа в основной формуле NGD может быть получено нормированное расстояние Google между этими понятиями «horse» и «rider» следующее: $NGD(horse, rider) \approx 0.443$. Параметр NGD – это нормированное семантическое расстояние между понятиями, которое обычно заключено (но не всегда, как увидим ниже) между 0 (подобные) и 1 (не связанные), в понятийной сфере используемых терминов, отфильтрованных поисковиком Google.

Разбор решения задачи 2.1

Сравнение двух биомедицинских терминов с помощью параметра NGD

Целью практической части семинара является нахождение меры семантической близости между двумя заданными биомедицинскими терминами с использованием нормализованного Google-расстояния (NGD), как это показано в примере 1 выше, варианты задачи 2.1 для практической работы даны в таблице 2.1.

Таблица 2.1 – Варианты задачи 2.1 для проведения практической части семинара

№ варианта	х (первый термин)	у (второй термин)
1	инсулин	диабет
2	антибиотик	инфекция
3	вакцина	иммунитет
4	Рак	опухоль
5	Вирус	заболевание
6	Сердце	кардиология
7	Печень	гепатит
8	антиген	антитело
9	хирургия	операция
10	генетика	наследственность
11	витамин	дефицит
12	аллергия	иммунная система
13	Нейрон	мозг
14	Кровь	гемоглобин
15	бактерия	микробиология

Сбор данных. Используем API Google или другой поисковый движок для получения частоты упоминания каждого термина и их совместного упоминания.

Обозначим:

$f(\text{диабет})$ – количество документов, содержащих термин «диабет»;

$f(\text{инсулин})$ – количество документов, содержащих термин «инсулин»;

$f(\text{диабет, инсулин})$ – количество документов, содержащих оба термина;

N – общее количество проиндексированных страниц Google.

На момент написания этого учебного пособия точное общее количество проиндексированных Google страниц – тайна, тщательно охраняемая компанией OpenAI. Это число постоянно меняется, так как динамичен: страницы появляются, исчезают, обновляются и меняют свое содержание. В прошлом Google иногда предоставлял ориентировочную оценку, но перестал это делать. Последняя известная оценка (очень приблизительная) была в 2013 году - 30 триллионов страниц. Таким образом, используя NGD, лучше вообще не учитывать N , если $f(x)$, $f(y)$ и $f(x, y)$ существенно меньше триллиона (например, несколько десятков или сотен тысяч). Либо использовать 10^{12} , как верхнюю границу. Это даст более адекватные результаты.

Студенты сами находят значения $f(\text{диабет})$, $f(\text{инсулин})$ и $f(\text{диабет, инсулин})$ и заполняют пустующие столбцы таблицы 2.2.

Вычисление NGD. Каждый студент подставляет собранные данные в свою, по примеру таблицы 2.2.

Таблица 2.2 – Вычисление NGD для заданной пары терминов $(x, y) = (\text{инсулин}, \text{диабет})$.

$f(x)$	$19,1 \cdot 10^6$
$f(y)$	$106 \cdot 10^6$
$f(x, y)$	$11,7 \cdot 10^6$
N	$\approx 10^{12}$
$NGD(x, y)$	0,203

В результате для пар терминов, показанных в таблице 2.2, получаем меру схожести $NGD(\text{диабет, инсулин}) = 0,203$.

Интерпретация результатов. Если значение NGD близко к 0, это указывает на то, что термины «диабет» и «инсулин» часто встречаются вместе, что говорит о высокой степени их семантической схожести. Если значение NGD близко к 1, это указывает на то, что термины редко встречаются вместе, что говорит о низкой степени их схожести.

В данном примере значение NGD составляет примерно 0,203, что указывает на то, что термины «диабет» и «инсулин» имеют значительную степень схожести, так как они часто встречаются вместе в биомедицинских текстах. Это может быть полезно для дальнейшего анализа, например, в контексте исследований, связанных с диабетом и его лечением.

Использование NGD для сравнения биомедицинских терминов позволяет исследователям и практикам лучше понимать взаимосвязи между различными концепциями в области медицины и биологии.

Разбор решения задачи 2.2 «Классификация белков»

Условие задачи. Дан белок X . Определите, к какой из двух групп белков (группа A или группа B) он наиболее близок, используя NGD . Даны названия характерных белков для каждой группы. Группа A включает белок A_1 и белок A_2 . Группа B включает белок B_1 и белок B_2 . Варианты задачи 2.3 приведены в таблице 2.3.

Таблица 2.3 – Варианты задачи 2.2 для проведения практической части семинара

№ Варианта	Белок X	Группа A , белок A_1 , белок A_2	Группа B , белок B_1 , белок B_2
1	Инсулин	гормональные белки, глюкагон, лептин	ферменты, амилаза, липаза

Продолжение таблицы 2.3

№ Варианта	Белок X	Группа A , белок A_1 , белок A_2	Группа B , белок B_1 , белок B_2
2	Трипсин	ферменты пищеварения, пепсин, химотрипсин	транспортные белки, гемоглобин, альбумин
3	Гемоглобин	транспортные белки, миоглобин, трансферрин	иммунные белки, интерферон, интерлейкин
4	Интерлейкин	иммунные белки, интерферон, комплемент	структурные белки, коллаген, кератин
5	Актин	структурные белки, тубулин, кератин	ферменты, каталаза, лактатдегидрогеназа
6	Каталаза	ферменты, пероксидаза, амилаза	гормональные белки, инсулин, глюкагон
7	Альбумин	транспортные белки, трансферрин, гемоглобин	иммунные белки, IgG, IgM
8	Миозин	структурные белки, актин, коллаген	ферменты, пепсин, трипсин
9	Лизоцим	ферменты, амилаза, липаза	иммунные белки, антитела, комплемент
10	IgG	иммунные белки, IgM, IgA	транспортные белки, гемоглобин, альбумин
11	Коллаген	структурные белки, эластин, кератин	гормональные белки, инсулин, адреналин
12	Адреналин	гормональные белки, норадреналин, инсулин	ферменты, амилаза, каталаза
13	Фибриноген	транспортные белки, альбумин, трансферрин	структурные белки, актин, тубулин
14	Пепсин	ферменты пищеварения, трипсин, химотрипсин	иммунные белки, лизоцим, интерферон
15	Миоглобин	транспортные белки, гемоглобин, альбумин	структурные белки, актин, миозин

Вычисляем NGD для белка X и каждого белка из группы A

$$NGD(\text{белок } X, \text{белок } A_1),$$

$$NGD(\text{белок } X, \text{белок } A_2).$$

Вычисляем NGD для белка X и каждого белка из группы B

$$NGD(\text{белок } X, \text{белок } B_1),$$

$$NGD(\text{белок } X, \text{белок } B_2).$$

Вычисляем среднее NGD для белка X и каждой группы

$$AVG_NGD(\text{Белок } X, \text{Группа } A) = \frac{NGD(\text{белок } X, \text{белок } A_1) + NGD(\text{белок } X, \text{белок } A_2)}{2}$$

$$AVG_NGD(\text{Белок } X, \text{Группа } B) = \frac{NGD(\text{белок } X, \text{белок } B_1) + NGD(\text{белок } X, \text{белок } B_2)}{2}$$

Сравниваем средние значения NGD.

Если

$$AVG_NGD(\text{Белок } X, \text{Группа } A) < AVG_NGD(\text{Белок } X, \text{Группа } B),$$

то делаем вывод, что белок X ближе к группе A . Иначе, белок X ближе к группе B .

Представляется пример с числами (вариант №1), указанный в таблице 2.4.

Таблица 2.4 –Вычисление NGD для белка X и каждого белка из группы A и B

Пара	Оценка NGD (условная)
$NGD(\text{инсулин, глюкагон})$	0,25
$NGD(\text{инсулин, лептин})$	0,35
$NGD(\text{инсулин, амилаза})$	0,7
$NGD(\text{инсулин, липаза})$	0,65

Вычисляем среднее NGD для белка *X* и каждой группы.

$$\text{AVG_NGD}(\text{Инсулин, Группа A}) = 0,5 * (0,25 + 0,35) = 0,30;$$

$$\text{AVG_NGD}(\text{Инсулин, Группа B}) = 0,5 * (0,7 + 0,65) = 0,675;$$

Сравниваем средние значения NGD.

$$\text{AVG_NGD}(\text{инсулин, A}) = 0.30 < 0.675 = \text{AVG_NGD}(\text{инсулин, B}),$$

Инсулин ближе к группе А (гормональные белки), чем к группе В (ферменты), что соответствует биологической реальности.

Ответ: Белок «инсулин» наиболее близок к группе А (гормональные белки).

Обсуждение с студентами. Для получения наиболее релевантных результатов из Google важно формулировать запросы с учетом специфики искомой информации, используя точные ключевые слова и фразы. Следует помнить, что нормализованное расстояние Google (NGD) является приближенной оценкой колмогоровской сложности, и его точность зависит от контента, индексируемого Google, что может привести к смещению результатов в зависимости от популярности терминов. Использование NGD имеет свои ограничения, особенно в случаях, когда информация о сравниваемых объектах недостаточно представлена в Интернете. Важно также сравнивать NGD с другими мерами схожести, применяемыми в *биоинформатике*, такими как *методы выравнивания последовательностей* [7], которые могут предоставить более точные и надежные результаты при анализе геномных данных. Таким образом, выбор метода анализа должен основываться на конкретных задачах и доступной информации.

Заключение по результатам проведения семинара №2

В заключении семинара необходимо еще раз подчеркнуть ключевые моменты:

- колмогоровская сложность – фундаментальное понятие в теории информации, *на практике невычислима*, но на больших данных – асимптотически достижима оценка;
- NGD – мера схожести, основанная на колмогоровской сложности, которую можно приближенно оценить, используя Google distance;
- параметр NGD имеет свои преимущества и ограничения. Важно понимать, когда его можно использовать, а когда нет;
- параметр NGD можно применять в биомедицинских задачах, таких как классификация генов, выявление взаимодействий белков, сравнение заболеваний.

Вопросы для самоконтроля полученных знаний у студентов

1. Опишите, как Google distance используется в контексте вычисления нормализованного информационного расстояния (NGD). Какие факторы могут влиять на точность оценки схожести, полученной с помощью Google distance?

Цель вопроса – проверить понимание студентами практической реализации NGD с использованием Google distance. Студенты должны объяснить, как количество результатов поиска для отдельных запросов и их комбинаций используется для оценки схожести, а также понимать ограничения этого подхода (например, влияние формулировки запроса, индексация веб-страниц, лингвистические особенности, культурный контекст и т.д.). Они должны продемонстрировать понимание того, что Google Distance – это приближение, а не точное вычисление колмогоровской сложности.

2. Представьте, что вам необходимо сравнить два вирусных генома для определения степени их родства. Какие проблемы могут возникнуть при использовании NGD, основанного на Google distance, в данном сценарии, и какие альтернативные подходы к вычислению схожести могли бы быть более подходящими?

Цель вопроса – проверить способность студентов применять теоретические знания о NGD к конкретной задаче и анализировать ограничения метода в реальных условиях. Студенты должны понимать, что для малоизученных или недавно появившихся вирусов информации в Google может быть недостаточно для точной оценки NGD. Они должны предложить альтернативные подходы, основанные на анализе последовательностей ДНК, такие как выравнивание геномов или сравнение частот k-меров. Это проверяет понимание области применения и ограничений NGD, а также знакомство с альтернативными методами.

3. Объясните, как можно модифицировать формулу NGD для учета разной «важности» или «информативности» различных терминов при сравнении объектов. Приведите пример, как это можно реализовать в задаче сравнения научных статей по биомедицине.

Цель вопроса – проверить способность студентов к творческому применению и модификации концепции NGD. Студенты должны предложить способы взвешивания терминов, например, на основе их частоты встречаемости в корпусе текстов или их значимости, определяемой экспертами. Они должны продемонстрировать, как это можно применить на практике, например, присваивая больший вес ключевым терминам, относящимся к конкретному заболеванию или методу лечения. Это демонстрирует глубокое понимание формулы NGD и способность ее адаптировать для решения конкретных проблем.

Семинарское занятие № 3

Статистические меры: частота термина и обратная частота документа

Цели занятия:

- дать определение трех вероятностных мер: TF, IDF и TF-IDF;
- исследовать связь между TF-IDF и колмогоровская сложностью;
- изучение применения TF-IDF к расчетам колмогоровской сложности.

План занятия

1. Введение в тему – 10 мин.
2. Теоретическая часть – 30 мин.
3. Практическая часть – 30 мин.
4. Обсуждение результатов – 10 мин.
5. Заключение семинара – 10 мин.

Теоретическая часть

В области информационной обработки данных параметр TF-IDF является популярной статистической мерой, используемой для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Величина TF-IDF представляет собой вес слов, который пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

Ниже приведены подробные определения и формулы значений TF, IDF, TF-IDF соответственно в формулах (3.1), (3.2), (3.3), которые выполнены в

контексте для термина t документа d_i в наборе документов, содержащих D документов:

- частота термина (term frequency – TF) – отношение числа вхождения некоторого термина к общему количеству термин документа. Таким образом значение TF, показанное в формуле (3.1) ниже, оценивает важность термина t пределах отдельного документа d_i :

$$TF(t, d_i) = \frac{n_t}{\sum_k n_k}, \quad (3.1)$$

где n_t – число вхождений слова t в документ d_i , а в знаменателе – общее число слов в данном документе;

- обратная частота документа (inverse document frequency – IDF) – инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт IDF уменьшает вес широкоупотребительных слов. Значение IDF представлено ниже в формуле:

$$IDF(t, D) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|}, \quad (3.2)$$

где $|D|$ – число документов в коллекции, а $|\{d_i \in D | t \in d_i\}|$ – число документов из коллекции D , в которых встречается t (когда $n_t \neq 0$).

Выбор основания логарифма в формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов;

- TF-IDF является произведением двух сомножителей, его значение вычислено по формуле:

$$TF - IDF(t, d_i, D) = tf(t, d_i) \cdot idf(t, D), \quad (3.3)$$

где $TF(t, d)$ – частота термина t в документе d_i ; $IDF(t, D)$ – обратная частота документа термина t в D -ых документах.

Большой вес в TF-IDF получают слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Параметр TF-IDF широко применяется благодаря своей простоте и эффективности при обработке документов. В частности, tf-idf можно интегрировать со многими алгоритмами обработки информации для повышения эффективности этого алгоритма, это также можно применить к колмогоровской сложности.

Колмогоровскую сложность объекта (K) можно определить как длину или объем кратчайшей программы[8], используемой для построения этого объекта. Колмогоровскую сложность объекта x можно выразить формулой (3.4) ниже:

$$K(x) = |x|, \quad (3.4)$$

где x^* – самая короткая программа для построения объекта x , $|x|$ – длина в битах программы x^* .

Однако это лишь колмогоровская сложность отдельного объекта; обычно обработка информации охватывает множество объектов, а информационная ценность каждого объекта в наборе определяется только относительно других объектов. Считаем, что колмогоровская сложность другого объекта y равна $K(y)$, однако колмогоровская сложность x относительно y относительна и не может быть выражена интуитивно только через $K(x)$ и $K(y)$, здесь эта величина рассматривается как $K(x|y)$ [8]. Если сравнение выполняется по x и y , это сравнение можно рассматривать как программу, которая выполняет две задачи: построить x и y , дифференцировать

x и y , таким образом, колмогоровская сложность этого сравнения равна длине этой программы и может быть определена по формуле (3.5):

$$K(x, y) = K(x) + K(y|x^*) = K(y) + K(x|y^*), \quad (3.5)$$

Значение колмогоровской сложности зависит от средств реализации алгоритма, генерирующего объекта, поэтому универсальная колмогоровская сложность между всеми объектами не может быть рассчитана, поскольку не существует универсального алгоритма или средств реализации этого алгоритма. Вместо этого колмогоровскую сложность конкретной группы объектов можно рассчитать с помощью некоторой вычислимой меры информации. Для документов, мера здесь может быть нормализованным расстоянием сжатия (Normalized Compressed Distance – NCD). В NCD колмогоровскую сложность считают, как размер строки сжатия информации объекта, по сути, строка сжатия информации – это кратчайшая программа, на основе которой можно реконструировать (восстановить) объект. Размер строка сжатия обозначим $C(x)$, теперь NCD можно рассчитать по формуле:

$$NCD(x, y) = \frac{C(x, y) - \min \{C(x), C(y)\}}{\max \{C(x), C(y)\}}, \quad (3.6)$$

В этом случае TF-IDF можно применить к набору документов, чтобы удалить избыточные слова, не имеющие отношения к содержанию текста, и в то же время выбрать слова с высокой информационной ценностью для оценки точности вычисленного значения колмогоровской сложности.

Исходные данные. Даны 5 текстов:

Текст 1: «Biomedical systems refer to the integration of biology, medicine, and engineering to develop technologies that improve healthcare outcomes. These systems include medical devices, diagnostic tools, and health information technologies that enhance patient care, streamline clinical workflows, and facilitate research in biomedical sciences».

Текст 2: «One of the most exciting advancements in biomedical systems is the emergence of wearable health technology. Devices like smartwatches and fitness trackers monitor vital signs such as heart rate, oxygen levels, and physical activity. This real-time data collection empowers individuals to take charge of their health and enables healthcare providers to offer personalized care based on continuous monitoring».

Текст 3: «Biomedical systems have revolutionized healthcare delivery through telemedicine. This technology allows patients to consult with healthcare professionals remotely, improving access to care, especially in rural or underserved areas. By utilizing video conferencing and mobile health applications, telemedicine enhances patient engagement and adherence to treatment plans while reducing the burden on healthcare facilities».

Текст 4: «Biomedical imaging systems, such as MRI, CT scans, and ultrasound, play a critical role in diagnosing and monitoring diseases. These technologies provide detailed visualizations of internal body structures, enabling healthcare providers to make informed decisions about treatment options. Advancements in imaging techniques continue to enhance accuracy and reduce patient exposure to radiation».

Текст 5: «The integration of artificial intelligence (AI) into biomedical systems is transforming diagnostics and treatment planning. Machine learning algorithms analyze vast datasets from medical records and imaging studies to

identify patterns and predict outcomes. This technology aids clinicians in making quicker, more accurate diagnoses and personalizing treatment strategies based on individual patient data».

Задачи

3.1. Вычислить значения TF, IDF и TF-IDF.

3.2. Вычислить сложность Колмогорова с помощью TF-IDF и без него.

Сравнить полученные результаты.

Решение

Задачи можно решить с помощью программы в Приложении к семинаре № 3.

Задача 3.1 выполняется в три этапа:

1) выполнить предварительную обработку текста: удалить знаки препинания, преобразовать слова в нижний регистр, удалить пробелы,

2) создать словарь и корпус (набор слов) с помощью библиотеки Gensim, специализированной библиотеки обработки текста в Python,

3) рассчитать значения TF, IDF и TF-IDF для любых 10 слов в тексте, TF-IDF можно рассчитать с помощью функции библиотеки Gensim.

В таблице 3.1 ниже приведены результаты для 10-ти слов из 2 документов с самыми высокими значениями TF, IDF и TF-IDF. Студенты могут выполняет задачу со самовыбранным текстом.

Таблица 3.1 – Примерный результат выполнения первой задачи

TF		IDF		TF-IDF	
Слова	Значение	Слова	Значение	Слова	Значение
and	0,0769	vast	1,6094	that	0,4124
that	0,0476	transforming	1,6094	telemedicine	0,3342
treatment	0,0385	studies	1,6094	workflows	0,2062
telemedicine	0,0385	strategies	1,6094	tools	0,2062
workflows	0,0238	records	1,6094	streamline	0,2062
tools	0,0238	quicker	1,6094	science	0,2062
streamline	0,0238	predict	1,6094	research	0,2062
science	0,0238	planning	1,6094	refer	0,2062
research	0,0238	personalizing	1,6094	medicine	0,2062
refer	0,0238	patterns	1,6094	information	0,2062

Задача 3.2 выполняется в следующие этапы:

- определить формулу для расчета колмогоровской сложности, здесь используется формула NCD (3.6), и рассчитать колмогоровскую сложность любых 4 пар документов,

- используя значение TF-IDF, рассчитанное в предыдущем разделе, выберите слова с низкими значениями tf-idf и удалите их из документа, затем рассчитайте колмогоровскую сложность с новыми документами.

В таблице 3.2 ниже приведены значения колмогоровской сложности 4-х пар документов, рассчитанные по формуле (3.6) с удалением и без удаления слов с низкими значениями TF-IDF.

Таблица 3.2 – Результат выполнения второй задачи

Без использования tf-idf		С использованием tf-idf	
Тексты	Результат	Тексты	Результат
1 и 2	0,684982	1 и 2	0,689076
2 и 3	0,692308	2 и 3	0,705882
3 и 4	0,665428	3 и 4	0,699187
4 и 5	0,681481	4 и 5	0,674603

Полученные значения в таблице 3.2 находятся в диапазоне $[0; 1]$, где 0 – два совершенно разных документа, 1 – два совершенно одинаковых документа. Сравниваемые документы имеют достаточно схожее содержание, желаемый результат получается в диапазоне $[0,66; 0,70]$. Большинство полученных результатов увеличились после удаления слов с низкими значениями TF-IDF, однако изменение было незначительным.

Обсуждение полученных результатов

Полученные значения TF, IDF, TF-IDF довольно низкие и во многом совпадают, особенно с idf. Основных причин две: корпус небольшой (всего 5 документов) и сами документы короткие по объему.

Значение колмогоровской сложности, полученное до и после удаления слов с низкими значениями TF-IDF, не имеет большой разницы. Помимо вышеперечисленных причин, есть еще два фактора, которые влияют на

полученные результаты: алгоритм сжатия данных и пороговое значение TF-IDF. Оба параметра можно настроить для изменения полученных результатов.

Заключение по результатам проведения семинара №3

В заключении семинара необходимо для усвоения повторить следующие знания:

- параметр TF-IDF: определение и математические формулы;
- колмогоровская сложность: основное определение, относительная Колмогоровская сложность между объектами;
- параметр NCD: определение, вычисление колмогоровскую сложность через NCD;
- применение математических формул в семинаре и применение TF-IDF в обработке информации в целом и в колмогоровской сложности в частности.

Вопросы для самоконтроля полученных знаний у студентов

1. Как определяются TF, IDF и TF-IDF, и какую роль играет каждая из этих величин в оценке информативности слов в текстовых документах?

Цель вопроса – Этот вопрос направлен на проверку базового понимания ключевых понятий, используемых в обработке текстов: частоты термина (TF), обратной частоты документа (IDF) и их произведения – TF-IDF. Цель состоит в том, чтобы студенты могли не только воспроизвести формулы, но и объяснить, как каждая из величин участвует в определении важности слова в контексте отдельного документа и всего корпуса. Особое внимание уделяется пониманию того, как IDF подавляет общие слова, а TF подчёркивает редкие и значимые термины. Это позволяет оценить уровень усвоения материала и

способность к теоретическому осмыслению статистических методов анализа текста.

2. Как связаны между собой TF-IDF и колмогоровская сложность, и каким образом TF-IDF может быть использован для повышения точности вычисления колмогоровской сложности текстов?

Цель вопроса – Проверить умение устанавливать междисциплинарные связи между статистическими и алгоритмическими подходами к измерению информации. Студенты должны показать понимание того, как TF-IDF может использоваться как инструмент предобработки текста, позволяющий выделять наиболее информативные слова и тем самым уменьшать влияние шума и избыточности на вычисление колмогоровской сложности. Это способствует формированию умения интегрировать разные методы анализа информации и применять их в задачах сравнения и классификации текстов.

3. Каким образом можно использовать TF-IDF для улучшения алгоритмов сравнения текстов на основе нормализованного сжатого расстояния (NCD)? Приведите примеры и объясните, почему это приводит к улучшению результатов.

Цель вопроса – Этот вопрос направлен на развитие практических навыков применения TF-IDF в реальных задачах обработки информации. Студенты должны показать, что понимают, как предварительная фильтрация текстов с помощью TF-IDF влияет на вычисление NCD – меры, основанной на колмогоровской сложности. Также важно, чтобы они могли проанализировать, как удаление слов с низким весом TF-IDF влияет на точность и информативность результатов сравнения текстов. Это развивает навыки практического применения теоретических знаний и умение интерпретировать результаты вычислений в контексте реальных данных.

4. Почему слова с низким значением TF-IDF могут считаться менее информативными, и как удаление таких слов влияет на качество анализа текста?

Цель вопроса – Данный вопрос ставит своей целью углублённое рассмотрение проблемы избыточности в текстовых данных и способов её минимизации. Цель состоит в том, чтобы студенты смогли объяснить, почему слова с низкими значениями TF-IDF (например, предлоги, союзы, часто встречающиеся глаголы и существительные) несут минимальную смысловую нагрузку, и как их удаление может улучшить качество анализа текста, включая задачи классификации, кластеризации и сравнения документов. Также важно уметь объяснить влияние таких преобразований на результаты вычисления сложности текста и сходства между документами. Это развивает навыки критического мышления и умение оценивать влияние предобработки текста на конечные результаты.

Семинарское занятие № 4

Передача информации в биосистемах: теория информации Шеннона

Цели занятия:

- ознакомление с теорией информации Клода Шеннона, включая понятие энтропии, как меры информации и случайности, изучение классической схемы информационной системы связи К. Шеннона, а также рассмотрение связи теории Шеннона с колмогоровской сложностью;
- применение теории информации Шеннона для расчета количества информации в текстах, анализа взаимной информации между источниками, сравнения с колмогоровской сложностью и исследования влияния структуры информации (например, разбиение на «мешки слов») на способность человека запоминать информацию, включая учет «магического числа семь плюс-минус два».

План занятия

1. Введение в тему – 10 мин.
2. Теоретическая часть – 20 мин.
3. Практическая часть – 30 мин.
4. Обсуждение результатов – 20 мин.
5. Заключение семинара – 10 мин.

Теоретический минимум
Теория информации Шеннона

Количество и свойства информации можно рассчитать не только с помощью колмогоровской сложности, но и с помощью теории информации Шеннона, в данном случае исходя из определения меры неопределенности (энтропии) информации.

Можно войти в электронную библиотеку Бауманки по логину и паролю почты и посмотреть наш учебник. Оттуда взять схему

Энтропия – это мера случайности объекта, другими словами, она представляет собой минимальный объем информации, который можно использовать для описания этого объекта.

Концепция информационной энтропии была введена Клодом Шенноном в его работе «Математическая теория связи» 1948 года и также известна как энтропия Шеннона. В этой работе Шеннон определил базовую систему коммуникации, состоящую из трёх основных компонентов: источника данных, канала связи и приёмника, которые в совокупности создали базовую модель коммуникации: модель Шеннона-Уивера – показана ниже на рисунке 1[1, 10].

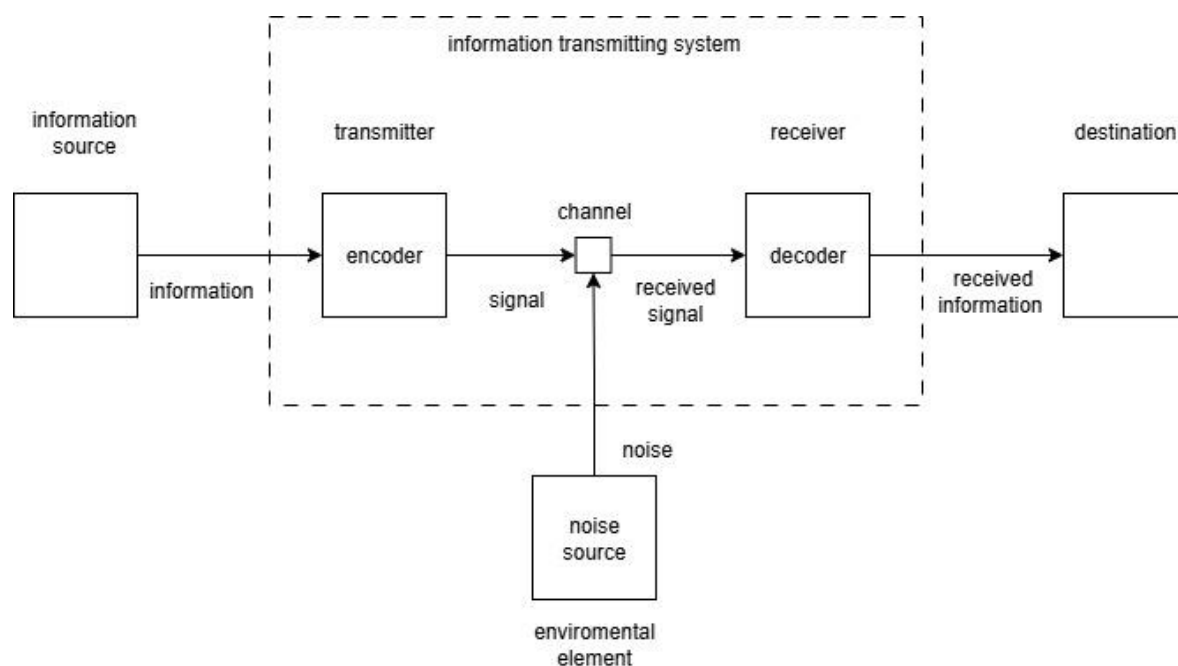


Рисунок 1 – Модель Шеннона-Уивера

Для оценки качества и количества передаваемой информации, как упоминалось выше, Шеннон использует энтропию информации [1, 9], которую можно рассчитать следующим образом: если задан источник информации, включающий в себя символы $X = \{x_1, x_2, \dots, x_n\}$ с вероятностью того, что символы являются $P = \{p_1, p_2, \dots, p_n\}$, энтропия данного источника информации - $H(X)$, рассчитывается по формуле (4.1):

$$H(X) = -\sum_{i=1}^n p_i \log_2(p_i), \quad (4.1)$$

где $\log_2(p_i)$ можно обозначить как $I(p_i)$ – неопределенность (surprisal) p_i , можно увидеть, что чем меньше вероятность p_i , тем выше неопределенность, что означает, что его возникновение дает больше информации.

Таким образом, можно также понять, что $H(X)$ – это среднее оценки неопределенности всех символов с известной вероятностью в источнике информации. Но в случае, что, информация получена из более чем одного источника, и эти источники не связаны друг с другом, то энтропия полученной информации равна сумме энтропий источников. Формула (4.2) использована для расчета энтропии информации из двух независимых источников X, Y :

$$H(X, Y) = -\sum_{x,y} p(x, y) \log(p(x, y)), \quad (4.2)$$

Однако, подобно колмогоровской сложности, информация в реальности редко встречается сама по себе, вместо этого возможен этот случай: два источника информации X, Y взаимозависимы, могут получать информацию от источника X в зависимости от некоторых условий, связанных с источником Y . В этом случае, энтропия источника X при условии Y - $H(X|Y)$ рассчитывается по формуле (4.3) [10]:

$$H(X|Y) = -\sum_y p(y) \sum_x p(x|y) \log(p(x|y)) =$$

$$= - \sum_{x,y} p(x,y) \log(p(x|y)), \quad (4.3)$$

где $p(x|y)$ — вероятность получения информации x при заданном y .

Приведённый выше случай открывает весьма интересную возможность: информация может быть получена случайным образом через другую информацию. Другими словами, если имеется любая информация из источника Y , существует вероятность получения информации из источника X , в котором описан источник Y . Количество информации, полученной таким образом, можно рассчитать по формуле (4.4):

$$I(X, Y) = - \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X|Y), \quad (4.4)$$

Теория информации Шеннона может использоваться параллельно с колмогоровской сложностью для описания информации через два независимых аспекта: случайность информации и сложность информации. Для модели коммуникации, представленной на рисунке 1, предложенной Шенноном, задача состоит в минимизации шума в процессе передачи. Согласно идее Шеннона, эта задача может быть решена с помощью алгоритмов шифрования, дешифрования и сжатия.

При передаче информации в биологической системе, включающей человеческое познание, еще один фактор оказывает большое влияние на качество передаваемой информации. Этот фактор связан с естественной способностью человека запоминать и обрабатывать информацию, а именно: кратковременная память человека позволяет ему эффективно и точно запоминать около (7 ± 2) блоков информации. Количество блоков информации, которые можно запомнить, также зависит от типа информации (звуки, буквы, цифры и т.д.) и объёма информации в каждом блоке. Эта идея, известная как «магическое число семь плюс-минус два», была предложена когнитивным психологом Джорджем А. Миллером в его статье «Магическое

число семь, плюс-минус два: некоторые ограничения нашей способности обрабатывать информацию» в 1956 году [11].

Возвращаясь к модели коммуникации на рисунке 1, в случае включения биологических систем и когнитивных способностей человека, компоненты в модели могут иметь следующие вид:

- тип информации: звук, символ, чувство, слух, ...,
- источник информации: текст, голос, изображения, ...
- передатчик: устройство или часть тела, используемые для передачи информации – рот, руки, телефон, телевизор, ...
- канал: воздух, контакт, свет, ...
- приёмник: уши, глаза, руки, ...
- место назначения: мозг
- факторы окружающей среды и шума: Звук, свет, запах, внешние источники информации, материалы, изменяющие тактильные ощущения, ...

Практическая часть

Исходные данные. Даны 4 текста:

Текст 1: «The sun shines. Birds sing in the forest. Trees whisper softly. The sun warms the leaves. Forest animals play. The sun smiles again, again.».

Текст 2: «The moon rises. The forest sleeps. Trees stand quiet. The moon watches. Stars glow above. The moon stays quiet while the forest dreams peacefully.».

Текст 3: « dog кот солнце light дерево cat свет moon дерево кот light земля dog море небо свет cat вода дерево солнце, вода glow stays свет».

Текст 4: «1, 8, 3, 7, 19, 6, 7, 27, 9, 682, 7, 3970, 2, 7, 3, 10, 1, 19, 7, 20, 21, 26, 300, 1000».

Задача

Рассматривайте каждый текст, как источник информации.

4.1. Рассчитать энтропию каждого текста, она также известна как объем информации в этом тексте, согласно теории информации Шеннона. Установить взаимосвязь между энтропией текста, длиной текста и количеством терминов в нём (сложностью).

4.2. Рассчитать объем информации, полученной из одного источника, через информацию из другого источника. Рассчитать сходство между двумя источниками информации через колмогоровскую сложность. Сравнить полученные результаты с энтропией, вычисленной в задаче 1. Анализировать связь между теорией информации Шеннона и колмогоровской сложностью.

4.3. Задача выполняется в группе по 3-4 участника: тексты разделены на мешки слов с одинаковой длиной, количество мешков слов выбирается случайным образом. Необходимо оценить способность каждого участника запоминать информацию, для этого рассчитайте объём информации каждого мешка слов согласно теории информации Шеннона, установите связь между объёмом памяти человека, объёмом информации и количеством пакетов слов.

Решение.

Задачи можно решить с помощью программы в приложении Приложении к семинаре № 4.. Все результаты в таблицах 4.1 - 4.4 рассчитаны с помощью этой программы.

Задача 1 выполняется в два этапа:

- рассчитать вероятность появления каждого слова в каждом документе. Вероятность встречи одной термины можно рассматривать как частоту термина (term frequency) – она может быть рассчитана с помощью формулы (3.1) в семинаре 3. Результаты представлены в таблице 4.1,

- рассчитать энтропию согласно теории информации Шеннона, применив формулу (4.1), результаты описаны в таблице 4.2.

Таблица 4.1 – Частота терминов в каждом тексте

Текст 1		Текст 2		Текст 3		Текст 4	
Термина	Частота	Термина	Частота	Термина	Частота	Термина	Частота
again	0,0833	above	0,0417	cat	0,0833	1	0,0833
animal	0,0417	dreams	0,0417	dog	0,0833	10	0,0417
bird	0,0417	forest	0,0833	glow	0,0417	1000	0,0417
forest	0,0833	glow	0,0417	light	0,0833	19	0,0833
in	0,0417	moon	0,125	moon	0,0417	2	0,0417
leaves	0,0417	peacefully	0,0417	stays	0,0417	20	0,0417
play	0,0417	quiet	0,0833	вода	0,0833	21	0,0417
shines	0,0417	rises	0,0417	дерево	0,125	26	0,0417
sing	0,0417	sleeps	0,0417	земля	0,0417	27	0,0417
smiles	0,0417	stand	0,0417	кот	0,0833	3	0,0833
softly	0,0417	stars	0,0417	море	0,0417	300	0,0417
sun	0,125	stays	0,0417	небо	0,0833	3970	0,0417
the	0,2083	the	0,2083	свет	0,125	6	0,0417
trees	0,0417	trees	0,0417	солнце	0,0833	682	0,0417
warms	0,0417	watches	0,0417			7	0,2083
whisper	0,0417	while	0,0417			8	0,0417
						9	0,0417
Итог	1		1		1		1

Таблица 4.2 – Энтропия и характеристика текстов

Текст	Длина текста (количество слов)	Количество символов	Энтропия текста (бит)
1	24	139	3,73644
2	24	146	3,73644
3	24	128	3,68872
4	24	87	3,85122

Из результатов на таблице 4.2, видно что, энтропия четырёх документов довольно схожа: документы 1 и 2 имеют аналогичное содержание, в то время как документы 3 и 4 представляют собой просто последовательные строки случайных слов и чисел. Энтропия документов не зависит от количества символов в документе, а пропорциональна размеру словаря документов (у документа 3 наименьший словарь, а у документа 4 наибольший). Можно сказать, что для информационных цепочек одинаковой длины, чем больше словарь, тем ниже вероятность повторения любого слова, поэтому сложнее придумать правило для цепочки. Другими словами, чем случайнее цепочка, тем выше её энтропия.

Задача 2 выполняется в четыре этапа:

- объём информации рассчитывается по формуле (4.4), приведённой выше. Здесь $p(x, y)$ рассчитывается как частота пары слов x, y , занимающих одну и ту же позицию в двух документах. В таблице 4.3 ниже примерно перечислены все пары слов в двух текстах 1 и 2,

- сходство текстов (от 0 до 1) рассчитывается с использования формулы (3.6) в семинаре 3. Колмогоровская сложность, используемая в формуле – это размер (в битах) сжатого текста, в качестве функции сжатия используется библиотеки `gzip` на Python.

- Разница в вычисленной энтропии с результатом в задаче 1 использована в качестве отличия по объёму полученной информации. Она использована для сравнения с сходством текстов, вычисляемым выше, и так установить связи между теорией информации Шеннона и колмогоровской сложностью.

Задача выполнена на парах текстов: (1 и 2), (1 и 3), (1 и 4), (3 и 4), результаты которых описаны в таблице 4.4. Студенты могут выполнять задачу на самобыбранной тексте.

Таблица 4.3 – Перечисление пар слов текстов 1 и 2

Позиция	Текст 1	Текст 2	Позиция	Текст 1	Текст 2
1	The	the	13	sun	Stars
2	sun	moon	14	warms	glow
3	shines	rises	15	the	above
4	Birds	The	16	leaves	the
5	sing	forest	17	Forest	moon
6	in	sleeps	18	animals	stays
7	the	Trees	19	play	quiet
8	forest	stand	20	The	while
9	Trees	quiet	21	sun	the
10	whisper	the	22	smiles	forest
11	softly	moon	23	again	dreams
12	The	watches	24	again	peacefully

Таблица 4.4 – Результат выполнения задачи 4.2

Тексты	Количество взаимной информации	Разлиха энтропии		Сходство с использованием колмогоровской сложности (0 – 1)
		Левый текст	Правый текст	
1 и 2	2,88792	0,84852	0,84852	0,47857
1 и 3	2,84019	0,89625	0,84853	0,65409
1 и 4	3,00271	0,73373	0,84851	0,67143
3 и 4	3,03832	0,6504	0,8129	0,72956

Как и ожидалось, объем информации, полученной косвенно, оказался ниже объема информации, полученной напрямую, как в задании 1. Разница составила 20–30 % от исходного объема информации, что является оценкой той части информации, которая не могла быть получена через косвенный источник.

Сравнение с вычисленной колмогоровской сложностью показывает, что вычисленное значение (0,47–0,73) не полностью удовлетворяет вычисленному объёму информации с использованием теории информации Шеннона. Полученные результаты также поднимают вопрос об алгоритме сжатия и формуле, используемой в случае, когда тексты 1 и 2 должны давать наибольшее сходство. Из-за случайности текстов 3 и 4 результаты сравнения, включающие эти тексты, несколько близки к ожидаемым: полученная информация составляет примерно 70 – 80% от информации, полученной напрямую, сходство между двумя текстами составляет примерно 0,66 – 0,73.

Задача 3 выполняется в пять этапов:

- выбрать любой текст (это может быть один из 4 примеров текстов), разделить выбранный текст на мешки слов (BOW) одинакового размера (по количеству слов в мешке),
- вычислить среднюю энтропию всех мешков слов,
- попросить участников группы запомнить выбранный текст двумя способами: запомнить текст напрямую, запомнить текст, запоминая каждый

мешок, и записать в отчёт оценку участников относительно того, какой способ запоминания оказался проще,

- повторять то же самое с другим текстом и количеством новых мешков слов,

- оценить взаимосвязь между следующими факторами и способностью памяти человека: количество мешков слов, тип информации (текст, буквы, цифры, ...), общий и средний объем информации на мешок слов, длина текста и количество символов в тексте.

Ниже приведено примерное выполнение для задачи 3.

Таблица 4.5 ниже используется для записи результатов работы. Эту таблицу можно использовать при составлении отчёта по задаче.

В таблице 4.5 записаны результаты четырёх наблюдений, каждый раз с разным текстом и разным размером мешка слов. Группа в данном примере состоит из 4 человек.

Таблица 4.5 – Примерная таблица отчета

Текст	Размер текста	Количество мешок слов	Энтропия текста	Средняя энтропия мешок слов	Оценка метода запоминания	
					Без мешка слов	С мешком слов
1	24	4	3,73644	2,55384	x	
2	24	3	3,73644	2,83	x	
3	24	8	3,68872	1,58496		x
4	24	6	3,85122	2,0		x

Общая оценка в таблице представляет собой среднюю оценку, согласованную участниками группы. Для записи подробной оценки участников можно создать отдельную таблицу.

В заключение, необходимо указать связь между следующими факторами: количеством мешков слов, средним объёмом информации в мешках слов и способностью участников к запоминанию.

Обсуждение полученных результатов

Общим для всех использованных текстов является их краткость и одинаковая длина (24 слова). Это, в сочетании с небольшим количеством текстов, приводит к тому, что полученные результаты не отражают чётко природу величины энтропии.

При оценке результатов в задаче 1 следует учитывать и другие характеристики: разную длину документов, типы информации в текстах (буквы, цифры, символы и т.д.), минимальные и максимальные значения частот терминов (tf).

Для результатов в задаче 2, помимо вышеописанных проблем, можно улучшить алгоритм сжатия или изменить используемую формулу Колмогорова для получения более точных результатов. Кроме того, можно изменить коэффициент, используемый для оценки общей информации, передаваемой в формуле (4) ($p(x, y)$), чтобы использовать пары документов с разными длинами.

Результаты задачи 3 сильно зависят от выбранного текста и индивидуальных способностей каждого участника группы. Выбор каждым участником группы нового текста для оставшихся участников может повысить объективность полученных результатов.

Заключение по результатам проведения семинара №4

В заключении семинара необходимо для усвоения повторить следующие знания:

- модель системы передачи информации Шеннона-Уивера,
- формулы энтропии в теории информации Шеннона,
- связь между следующими знаниями: теорией информации Шеннона, колмогоровской сложностью и естественной способностью запоминания у человека в идеальном «магическом числе семь плюс-минус 2».

Вопросы для самоконтроля полученных знаний у студентов

1. Какова связь между энтропией и количеством информации в теории Шеннона, и как можно рассчитать энтропию источника информации?

Цель вопроса – Этот вопрос направлен на проверку базового понимания ключевого понятия – энтропии Шеннона, как меры неопределенности и случайности источника информации. Студенты должны не только знать определение энтропии, но и уметь объяснить, почему именно энтропия используется для измерения информации. Также важно, чтобы они могли воспроизвести формулу Шеннона и объяснить смысл каждого элемента в ней. Это позволяет преподавателю оценить уровень усвоения теоретических основ теории информации и умение применять их на практике при анализе реальных источников данных.

2. Чем отличается подход Шеннона к измерению информации от подхода, основанного на колмогоровской сложности?

Цель вопроса – Данный вопрос направлен на развитие способности к сравнительному анализу двух ключевых подходов к измерению информации вероятностного (Шеннон) и алгоритмического (Колмогоров). Цель состоит в том, чтобы студенты могли не только сформулировать различия между этими теориями, но и понять, в каких ситуациях какой подход предпочтительнее. Это помогает формировать более глубокое представление о природе информации и её измерении, а также развивает критическое мышление и умение выбирать подходящий метод анализа информации в зависимости от задачи.

3. Как ограничения кратковременной памяти человека («магическое число семь плюс-минус два») влияют на эффективность передачи информации в биологических системах?

Цель вопроса – Этот вопрос ставит своей целью связать формальные модели передачи информации с реальными когнитивными ограничениями

человека. Студенты должны показать понимание того, как биологические и психологические особенности восприятия информации влияют на процесс коммуникации и обработки данных в биосистемах. Это позволяет развить навыки применения теоретических знаний в контексте человеческого фактора, а также научиться учитывать эти аспекты при проектировании систем передачи информации.

4. Как рассчитывается взаимная информация между двумя источниками, и что она может сказать о степени зависимости или сходства между ними?

Цель вопроса – Этот вопрос направлен на проверку понимания концепции условной и взаимной информации, а также на развитие навыков анализа взаимосвязей между источниками данных. Студенты должны уметь объяснить, как с помощью взаимной информации можно оценить степень зависимости между источниками, и как это связано с передачей информации через шумные каналы. Понимание этого вопроса помогает в дальнейшем при решении задач анализа данных, обработки сигналов, распознавания образов и других прикладных задач.

Заключение

В результате выполнения семинарских занятий, предложенных в данном учебно-методическом пособии, обучающиеся получают теоретические знания и развивают практические навыки в области алгоритмической теории информации применительно к биомедицинским системам. В ходе работы над темами семинаров студенты осваивают методы анализа и обработки информации с использованием современных информационных подходов, учатся формализовать и анализировать сложные процессы передачи и обработки данных в биосистемах.

На основе приобретённых знаний и умений у обучающихся формируется и развивается способность:

- применять современные методы теории информации и вычислительной техники для решения профессиональных задач в области биомедицинских систем;
- использовать аппарат алгоритмической теории информации для моделирования, анализа и интерпретации процессов передачи информации в сложных биологических и медицинских системах;
- анализировать научную и техническую информацию, структурировать её, выделять ключевые аспекты и представлять в виде логически связанных аналитических обзоров с обоснованными выводами;
- разрабатывать и адаптировать алгоритмические и программные решения, направленные на эффективную обработку информации в биомедицинских приложениях.

Литература

1. Павлов Ю. Н., Смирнова Е. В., Тихомирова Е. А. Теория информации для бакалавров : учеб. пособие для вузов / Павлов Ю. Н., Смирнова Е. В., Тихомирова Е. А. - М. : Изд-во МГТУ им. Н. Э. Баумана, 2016. - 173 с. : ил. - Библиогр.: с. 161. - ISBN 978-5-7038-4190-7.
2. Смирнова Е.В., Абачараева Э.Р. Алгоритмический подход к измерению количества информации А.Н. Колмогорова /Учебное пособие, 2020. – 40 с.
3. Колмогоров А.Н. Три подхода к определению понятия количество информации, Проблемы передачи информации, т. I, вып.1, 1965.
4. Успенский В.А., Вьюгин В.В. Становление алгоритмической теории информации в России / Информационные процессы, Том 10, № 2, 2010, стр. 145–158.
5. Городецкий В.И. Семантические вычисления, большие данные, кибербезопасность. Режим доступа: <http://comsec.spb.ru/imctcpa18/02.02.IM&СТСПА-2018-Gorodetsky.pdf> (дата обращения: 16.03.2025).
6. An Introduction to Kolmogorov Complexity and Its Applications. Ming Li, Paul Vitanyi. Springer, 2008.
7. Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). Eric S. Donkor, Nicholas T. K. D. Dayie, Theophilus K Adiku. Journal of Bioinformatics and Sequence Analysis. Vol. 6(1), pp. 1-6, April 2014.
8. Ming Li, Xin Chen, Xin Li, Bin Ma, Paul M.B. Vitányi. The Similarity Metric. IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. XX, NO Y. – 2004. C.1-11. (дата обращения 16.03.2025).
9. Ricky Xiaofeng Chen. A brief introduction to Shannon informational theory. Arxiv. – 2016. C.2-7. (дата обращения 16.03.2025).

10. Shannon and Weaver Model of Communication. Режим доступа: <https://www.communicationtheory.org/shannon-and-weaver-model-of-communication/> (дата обращения: 16.03.2025).

11. George Miller's Magical Number of Immediate Memory in Retrospect: Observations on the Faltering Progression of Science. Режим доступа: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4486516/> (дата обращения: 16.03.2025).

Приложение к семинарскому занятию № 1

Приложение 1: листинг кода примерной программы для расчёта частоты появления диграмм в указанном тексте

```
import re
import matplotlib.pyplot as plt

def analyze_digrams(filename):
    """
    Анализирует текст в файле и вычисляет статистические характеристики
    для диграмм,
    включая диграммы с пробелами.

    Args:
        filename: Имя файла с текстом.

    Returns:
        Словарь, где ключ - диграмм, а значение - количество его вхождений.
    """

    with open(filename, encoding='utf-8') as file:
        text = file.read()

    digrams = re.findall(r'([a-я ]{1})([a-я ]{1})', text.lower())
    digram_counts = {}
    for digram in digrams:
        digram_str = ''.join(digram) # Объединяем два символа в строку
        if digram_str in digram_counts:
            digram_counts[digram_str] += 1
        else:
            digram_counts[digram_str] = 1

    sorted_digrams = sorted(digram_counts.items())
    return sorted_digrams

if __name__ == '__main__':
    filename = 'test.txt'
```

```

sorted_digrams = analyze_digrams(filename)

print("Статистика диграммов (отсортировано по алфавиту):")
total_digrams = sum(count for _, count in sorted_digrams)

# Табличная статистика с нумерацией
print("-----")
print("№ | Диграмм | Количество | Вероятность")
print("-----")
for i, (digram, count) in enumerate(sorted_digrams, 1): # enumerate
для нумерации
    probability = count / total_digrams
    print(f"{i} | {digram} | {count} | {probability:.4f}")

# Графическая статистика
digrams, counts = zip(*sorted_digrams)
plt.figure(figsize=(12, 6))
plt.bar(digrams, counts)
plt.xlabel("Диграммы")
plt.ylabel("Количество")
plt.title("Частота появления диграмм в тексте")
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()

# Проверка суммы вероятностей
total_probability = sum(count / total_digrams for _, count in
sorted_digrams)
print(f"\nСумма вероятностей: {total_probability:.4f}")

```

Приложение 2: текст, использованный в семинаре 1

Это ранний абзац из романа «Тихий Дон». Этот текст используется в указанной выше программе или аналогичных программах для определения частоты встречаемости подобных приближений. Результаты работы программы приведены ниже:

«В предпоследнюю турецкую кампанию вернулся в хутор казак Мелехов Прокофий. Из Туретчины привел он жену – маленькую, закутанную в шаль женщину. Она прятала лицо, редко показывая тоскующие одичалые глаза. Пахла шелковая шаль далекими неведомыми запахами, радужные узоры ее питали бабью зависть. Пленная турчанка сторонилась родных Прокофия, и старик Мелехов вскоре отделил сына. В курень его не ходил до смерти, не забывая обиды.

Прокофий обстроился скоро: плотники срубили курень, сам пригородил базы для скотины и к осени увел на новое хозяйство сторбленную иноземку-жену. Шел с ней за арбой с имуществом по хутору – высыпали на улицу все, от мала до велика. Казаки сдержанно посмеивались в бороды, голосисто перекликались бабы, орда немых казачат улюлюкала Прокофию вслед, но он, распахнув чекмень, шел медленно, как по пахотной борозде, сжимал в черной ладони хрупкую кисть жениной руки, непокорно нес белесо-чубатую голову, – лишь под скулами у него пухли и катались желваки да промеж каменных, по всегдашней неподвижности, бровей проступал пот.

С той поры редко видели его в хуторе, не бывал он и на майдане. Жил в своем курене, на отшибе у Дона, бирюком. Гутарили про него по хутору чудное. Ребятишки, пасшие за прогоном телят, рассказывали, будто видели они, как Прокофий вечерами, когда вянют зори, на руках носил жену до Татарского ажик кургана. Сажал ее там на макушке кургана, спиной к источенному столетиями ноздреватому камню, садился с ней рядом, и так

подолгу глядели они в степь. Глядели до тех пор, пока истухала заря, а потом Прокофий кутал жену в зипун и на руках относил домой. Хутор терялся в догадках, подыскивая объяснение таким диковинным поступкам, бабам за разговорами поискаться некогда было. Разно гутарили и о жене Прокофия: одни утверждали, что красоты она досель невиданной, другие – наоборот. Решилось все после того, как самая отчаянная из баб, жалмерка Мавра, сбегала к Прокофию будто бы за свежей наливкой. Прокофий полез за наливкой в погреб, а за это время Мавра и разглядела, что турчанка попалась Прокофию последняя из никудышных...

Спустя время раскрасневшаяся Мавра, с платком, съехавшим набок, торочила на прогулке бабьей толпе:

– И что он, милушки, нашел в ней хорошего? Хоть бы баба была, а то так... Ни зад, ни пуза, одна страма. У нас девки глаже ее выгуливаются. В стану – перервать можно, как оса; глазюки – черные, здоровящие, стригеть ими, как Сатана, прости Бог. Должно, на сносях дохаживает, ей-бо!

– На сносях?

– дивились бабы.

– Кубыть, не махонькая, сама трех вынянчила.

– А с лица как?

– С лица-то? Желтая. Глаза тусменные, – небось не сладко на чужой стороншке. А ишо, бабоньки, ходит-то она... в Прокофьевых шароварах.

– Ну-у?... – ахали бабы испуганно и дружно.

– Сама видала – в шароварах, только без лампасин. Должно, будничные его подцепила. Длинная на ней рубаха, а из-под рубахи шаровары, в чулки вобратые. Я как разглядела, так и захолонуло во мне...

Шепотом гутарили по хутору, что Прокофьева жена ведьмачит. Сноха Астаховых (жили Астаховы от хутора крайние к Прокофию) божила, будто на второй день Троицы, перед светом, видала, как Прокофьева жена,

простоволосая и босая, доила на их базу корову. С тех пор ссохлось у коровы вымя в детский кулачок, отбила от молока и вскоре издохла.

В тот год случился небывалый падеж скота. На стойле возле Дона каждый день пятнилась песчаная коса трупами коров и молодняка. Падеж перекинулся на лошадей. Таяли конские косяки, гулявшие на станичном отводе. И вот тут-то прополз по проулкам и улицам черный слушок...

С хуторского схода пришли казаки к Прокофию.

Хозяин вышел на крыльцо, кланяясь.

– За чем добрым пожаловали, господа старики?»

Приложение 3: листинг приближения 1-го порядка

В таблице П1 приведены все буква, их количество и частота встречи в тексте

Таблица П1. Приближение 1-го порядка

Символ	Вероятность	Количество	Символ	Вероятность	Количество
Пробел	0,175	175	Я	0,018	18
О	0,090	90	Ы	0,016	16
Е	0,072	72	З	0,016	16
А	0,062	62	Ъ	0,014	14
И	0,062	62	Б	0,014	14
Н	0,053	53	Г	0,013	13
Т	0,053	53	Ч	0,012	12
С	0,045	45	Й	0,010	10
Р	0,040	40	Х	0,009	9
В	0,038	38	Ж	0,007	7
Л	0,035	35	Ю	0,006	6
К	0,028	28	Ш	0,006	6
М	0,026	26	Ц	0,004	4
Д	0,025	25	Щ	0,003	3
П	0,023	23	Э	0,003	3
У	0,021	21	Ф	0,002	2

Проверка правильности подсчетов: сумма всех вероятностей равна 1 – то есть, вычисление выполнено корректно.

Приложение 4: листинг приближения 2-го порядка

Замечание. При использовании этой таблицы: первый столбец – номер диграмм, которые сортируются по алфавиту; второй столбец - сама диграмма; третий столбец – количество появлений этой диграммы; а четвёртый столбец – вероятность появлений этой диграммы.

Проверка правильности подсчетов: сумма всех вероятностей равна 1 – то есть вычисление выполнено корректно.

1 а 10 0.0059	46 бе 3 0.0018	91 ем 2 0.0012
2 б 18 0.0107	47 бо 4 0.0024	92 ен 7 0.0041
3 в 28 0.0166	48 бр 1 0.0006	93 еп 2 0.0012
4 г 10 0.0059	49 бс 1 0.0006	94 ер 9 0.0053
5 д 21 0.0124	50 бу 1 0.0006	95 ес 3 0.0018
6 е 4 0.0024	51 бы 6 0.0036	96 ет 2 0.0012
7 ж 4 0.0024	52 бь 1 0.0006	97 ех 3 0.0018
8 з 12 0.0071	53 в 8 0.0047	98 еш 1 0.0006
9 и 17 0.0101	54 ва 9 0.0053	99 ж 1 0.0006
10 к 25 0.0148	55 ве 8 0.0047	100 жа 1 0.0006
11 л 5 0.0030	56 ви 4 0.0024	101 жд 1 0.0006
12 м 9 0.0053	57 во 3 0.0018	102 же 10 0.0059
13 н 37 0.0219	58 вр 3 0.0018	103 жи 2 0.0012
14 о 18 0.0107	59 вс 4 0.0024	104 жн 4 0.0024
15 п 49 0.0290	60 вт 1 0.0006	105 з 4 0.0024
16 р 11 0.0065	61 ву 2 0.0012	106 за 9 0.0053
17 с 35 0.0207	62 вш 2 0.0012	107 зг 2 0.0012
18 т 14 0.0083	63 вы 2 0.0012	108 зд 1 0.0006
19 у 5 0.0030	64 вю 1 0.0006	109 зе 1 0.0006
20 х 8 0.0047	65 га 4 0.0024	110 зы 1 0.0006
21 ч 9 0.0053	66 ги 1 0.0006	111 зю 1 0.0006
22 ш 6 0.0036	67 гл 2 0.0012	112 зя 1 0.0006
23 я 1 0.0006	68 го 11 0.0065	113 и 26 0.0154
24 а 37 0.0219	69 гу 3 0.0018	114 ив 3 0.0018
25 аб 7 0.0041	70 д 3 0.0018	115 иг 1 0.0006
26 ав 1 0.0006	71 да 9 0.0053	116 ид 4 0.0024
27 ад 6 0.0036	72 де 6 0.0036	117 ие 4 0.0024
28 ае 1 0.0006	73 ди 2 0.0012	118 из 1 0.0006
29 аж 2 0.0012	74 дк 1 0.0006	119 ий 1 0.0006
30 аз 8 0.0047	75 дн 4 0.0024	120 ик 4 0.0024
31 ай 1 0.0006	76 до 4 0.0024	121 ил 16 0.0095
32 ак 17 0.0101	77 др 2 0.0012	122 им 5 0.0030
33 ал 12 0.0071	78 дт 1 0.0006	123 ин 5 0.0030
34 ам 11 0.0065	79 ду 1 0.0006	124 ип 1 0.0006
35 ан 4 0.0024	80 е 13 0.0077	125 ир 1 0.0006
36 ао 1 0.0006	81 еб 3 0.0018	126 ис 7 0.0041
37 ап 1 0.0006	82 ев 5 0.0030	127 ит 3 0.0018
38 ар 10 0.0059	83 ег 4 0.0024	128 их 1 0.0006
39 ас 10 0.0059	84 ед 9 0.0053	129 иц 3 0.0018
40 ат 3 0.0018	85 ее 1 0.0006	130 иш 2 0.0012
41 ах 7 0.0041	86 еж 3 0.0018	131 ию 2 0.0012
42 ач 1 0.0006	87 еи 1 0.0006	132 ия 1 0.0006
43 аш 1 0.0006	88 ей 5 0.0030	133 й 11 0.0065
44 ая 11 0.0065	89 ек 5 0.0030	134 йд 1 0.0006
45 ба 7 0.0041	90 ел 15 0.0089	135 к 3 0.0018

136	ка	14	0.0083	181	ог	4	0.0024	226	си	2	0.0012
137	ке	2	0.0012	182	од	8	0.0047	227	ск	6	0.0036
138	ки	11	0.0065	183	ое	1	0.0006	228	сл	5	0.0030
139	ко	24	0.0142	184	ож	3	0.0018	229	см	2	0.0012
140	кр	3	0.0018	185	оз	4	0.0024	230	сн	3	0.0018
141	ку	5	0.0030	186	ои	3	0.0018	231	со	2	0.0012
142	л	8	0.0047	187	ой	8	0.0047	232	сп	1	0.0006
143	ла	21	0.0124	188	ок	8	0.0047	233	ср	1	0.0006
144	лв	1	0.0006	189	ол	9	0.0053	234	ст	12	0.0071
145	ле	7	0.0041	190	ом	8	0.0047	235	сч	1	0.0006
146	лз	1	0.0006	191	он	10	0.0059	236	сь	3	0.0018
147	ли	16	0.0095	192	ор	15	0.0089	237	ся	8	0.0047
148	лк	1	0.0006	193	ос	15	0.0089	238	т	3	0.0018
149	ло	4	0.0024	194	от	7	0.0041	239	та	13	0.0077
150	лп	1	0.0006	195	оу	1	0.0006	240	тб	1	0.0006
151	лс	1	0.0006	196	оф	3	0.0018	241	тв	2	0.0012
152	лу	1	0.0006	197	ох	2	0.0012	242	тд	1	0.0006
153	лы	1	0.0006	198	оч	1	0.0006	243	те	3	0.0018
154	ль	2	0.0012	199	ош	1	0.0006	244	ти	3	0.0018
155	ля	5	0.0030	200	па	5	0.0030	245	тк	1	0.0006
156	м	6	0.0036	201	пе	1	0.0006	246	тн	3	0.0018
157	ма	7	0.0041	202	пи	2	0.0012	247	то	23	0.0136
158	ме	6	0.0036	203	по	16	0.0095	248	тр	5	0.0030
159	ми	3	0.0018	204	пр	8	0.0047	249	тс	1	0.0006
160	мк	1	0.0006	205	пу	2	0.0012	250	ту	5	0.0030
161	мн	1	0.0006	206	пъ	1	0.0006	251	тч	2	0.0012
162	мо	2	0.0012	207	пя	1	0.0006	252	ты	3	0.0018
163	мп	1	0.0006	208	ра	7	0.0041	253	ть	2	0.0012
164	мы	2	0.0012	209	рб	1	0.0006	254	тя	1	0.0006
165	мя	1	0.0006	210	рд	1	0.0006	255	у	5	0.0030
166	н	1	0.0006	211	ре	8	0.0047	256	уб	2	0.0012
167	на	13	0.0077	212	рж	1	0.0006	257	ув	1	0.0006
168	не	11	0.0065	213	ри	5	0.0030	258	уд	5	0.0030
169	ни	7	0.0041	214	рк	1	0.0006	259	уж	3	0.0018
170	нк	2	0.0012	215	рн	3	0.0018	260	уз	1	0.0006
171	нн	11	0.0065	216	ро	27	0.0160	261	ул	8	0.0047
172	но	12	0.0071	217	рс	2	0.0012	262	ун	1	0.0006
173	ну	8	0.0047	218	ру	9	0.0053	263	уп	3	0.0018
174	ны	5	0.0030	219	ры	1	0.0006	264	ур	6	0.0036
175	нъ	3	0.0018	220	ря	1	0.0006	265	ус	2	0.0012
176	ню	1	0.0006	221	с	3	0.0018	266	ут	9	0.0053
177	ня	2	0.0012	222	са	1	0.0006	267	ух	1	0.0006
178	о	22	0.0130	223	св	1	0.0006	268	уч	1	0.0006
179	об	3	0.0018	224	сд	1	0.0006	269	уш	2	0.0012
180	ов	11	0.0065	225	се	2	0.0012	270	ущ	1	0.0006

271 ую 3 0.0018	293 ша 3 0.0018	315 ь 15 0.0089
272 фи 9 0.0053	294 ше 1 0.0006	316 ье 1 0.0006
273 фь 3 0.0018	295 ши 3 0.0018	317 ьк 3 0.0018
274 х 4 0.0024	296 шк 1 0.0006	318 ьм 1 0.0006
275 ха 5 0.0030	297 шл 1 0.0006	319 ьц 1 0.0006
276 хи 1 0.0006	298 шн 2 0.0012	320 эт 1 0.0006
277 хл 1 0.0006	299 шо 2 0.0012	321 ю 8 0.0047
278 хн 1 0.0006	300 щ 2 0.0012	322 юк 2 0.0012
279 хо 5 0.0030	301 ье 1 0.0006	323 юл 1 0.0006
280 хр 1 0.0006	302 ья 1 0.0006	324 ют 1 0.0006
281 ху 5 0.0030	303 ы 4 0.0024	325 ющ 1 0.0006
282 ца 1 0.0006	304 ыв 3 0.0018	326 я 7 0.0041
283 це 1 0.0006	305 ьи 1 0.0006	327 яв 1 0.0006
284 цк 1 0.0006	306 ый 3 0.0018	328 яд 2 0.0012
285 цу 1 0.0006	307 ыл 2 0.0012	329 яй 1 0.0006
286 цы 1 0.0006	308 ым 2 0.0012	330 як 2 0.0012
287 ча 4 0.0024	309 ын 2 0.0012	331 ял 1 0.0006
288 че 5 0.0030	310 ып 1 0.0006	332 ян 2 0.0012
289 чи 1 0.0006	311 ыс 2 0.0012	333 яс 1 0.0006
290 чн 1 0.0006	312 ыт 1 0.0006	334 ят 1 0.0006
291 чо 1 0.0006	313 ых 4 0.0024	335 ях 2 0.0012
292 чу 1 0.0006	314 ыш 2 0.0012	

Сумма вероятностей: 1.0000

Приложение 5: листинг приближения 3-го порядка

Замечание. При использовании этой таблицы: первый столбец – номер триграмм, которые сортируются по алфавиту; второй столбец - сама триграмма; третий столбец – количество появлений этой триграммы; а четвёртый столбец – вероятность появлений этой триграммы.

Проверка правильности подсчетов: сумма всех вероятностей равна 1 – то есть вычисление выполнено корректно.

1 а 6 0.0054	46 ку 3 0.0027	91 ту 4 0.0036
2 ас 1 0.0009	47 ли 2 0.0018	92 у 1 0.0009
3 ах 1 0.0009	48 ло 1 0.0009	93 ув 1 0.0009
4 ба 7 0.0063	49 ма 5 0.0045	94 уз 1 0.0009
5 би 1 0.0009	50 ме 1 0.0009	95 ул 3 0.0027
6 бо 1 0.0009	51 ми 1 0.0009	96 хо 3 0.0027
7 бр 1 0.0009	52 на 15 0.0136	97 ху 2 0.0018
8 бу 3 0.0027	53 не 7 0.0063	98 че 3 0.0027
9 бы 1 0.0009	54 ни 2 0.0018	99 чт 3 0.0027
10 в 10 0.0090	55 но 2 0.0018	100 ша 1 0.0009
11 ве 1 0.0009	56 ну 1 0.0009	101 ше 3 0.0027
12 ви 2 0.0018	57 од 2 0.0018	102 я 1 0.0009
13 вр 2 0.0018	58 он 5 0.0045	103 а а 1 0.0009
14 вс 3 0.0027	59 ор 1 0.0009	104 а б 2 0.0018
15 вт 1 0.0009	60 ос 1 0.0009	105 а д 1 0.0009
16 вы 3 0.0027	61 от 4 0.0036	106 а м 1 0.0009
17 вя 1 0.0009	62 па 3 0.0027	107 а о 1 0.0009
18 гл 4 0.0036	63 пе 3 0.0027	108 а п 5 0.0045
19 го 2 0.0018	64 пи 1 0.0009	109 а с 1 0.0009
20 гу 3 0.0027	65 пл 2 0.0018	110 або 1 0.0009
21 де 1 0.0009	66 по 12 0.0108	111 абь 1 0.0009
22 ди 1 0.0009	67 пр 9 0.0081	112 ави 1 0.0009
23 дл 1 0.0009	68 пу 2 0.0018	113 авр 1 0.0009
24 до 4 0.0036	69 пя 1 0.0009	114 аде 1 0.0009
25 др 1 0.0009	70 ра 7 0.0063	115 адо 1 0.0009
26 ег 1 0.0009	71 ре 3 0.0027	116 ает 1 0.0009
27 ее 3 0.0027	72 ру 4 0.0036	117 аже 1 0.0009
28 ей 1 0.0009	73 с 3 0.0027	118 аза 3 0.0027
29 жа 1 0.0009	74 са 5 0.0045	119 азы 1 0.0009
30 же 5 0.0045	75 сб 1 0.0009	120 азю 1 0.0009
31 жи 1 0.0009	76 св 2 0.0018	121 ак 1 0.0009
32 за 7 0.0063	77 сг 1 0.0009	122 аки 1 0.0009
33 зд 1 0.0009	78 сж 1 0.0009	123 аку 1 0.0009
34 зи 1 0.0009	79 ск 2 0.0018	124 ал 2 0.0018
35 зо 1 0.0009	80 сл 1 0.0009	125 ала 2 0.0018
36 и 10 0.0090	81 сн 3 0.0027	126 али 3 0.0027
37 из 1 0.0009	82 сп 1 0.0009	127 ало 1 0.0009
38 им 2 0.0018	83 ср 1 0.0009	128 алы 1 0.0009
39 ин 1 0.0009	84 ст 4 0.0036	129 аль 1 0.0009
40 ис 1 0.0009	85 сх 1 0.0009	130 ама 2 0.0018
41 их 1 0.0009	86 съ 1 0.0009	131 ами 1 0.0009
42 ка 11 0.0099	87 та 3 0.0027	132 ана 2 0.0018
43 кл 1 0.0009	88 те 1 0.0009	133 ани 1 0.0009
44 ко 3 0.0027	89 то 3 0.0027	134 анк 1 0.0009
45 кр 1 0.0009	90 тр 2 0.0018	135 анн 1 0.0009

136	аня	1 0.0009	181	во	1 0.0009	226	е у	1 0.0009
137	апа	1 0.0009	182	воб	1 0.0009	227	е х	1 0.0009
138	аро	1 0.0009	183	вор	1 0.0009	228	ева	2 0.0018
139	арс	1 0.0009	184	вот	1 0.0009	229	евш	1 0.0009
140	асн	1 0.0009	185	вра	2 0.0018	230	ега	1 0.0009
141	ась	1 0.0009	186	вск	1 0.0009	231	егд	1 0.0009
142	ата	1 0.0009	187	вши	1 0.0009	232	его	4 0.0036
143	ату	1 0.0009	188	вющ	1 0.0009	233	едк	1 0.0009
144	ать	1 0.0009	189	гда	1 0.0009	234	едл	1 0.0009
145	ах	1 0.0009	190	гла	1 0.0009	235	едн	1 0.0009
146	ахо	1 0.0009	191	гля	2 0.0018	236	едо	1 0.0009
147	ашн	1 0.0009	192	го	1 0.0009	237	еж	1 0.0009
148	ая	6 0.0054	193	год	1 0.0009	238	ез	1 0.0009
149	аяс	1 0.0009	194	гре	1 0.0009	239	ей	3 0.0027
150	баб	3 0.0027	195	гу	1 0.0009	240	еки	1 0.0009
151	бам	1 0.0009	196	гул	1 0.0009	241	екм	1 0.0009
152	бах	2 0.0018	197	да	2 0.0018	242	ел	2 0.0018
153	бел	1 0.0009	198	дал	2 0.0018	243	еля	1 0.0009
154	бил	1 0.0009	199	дан	1 0.0009	244	емя	2 0.0018
155	бон	1 0.0009	200	деж	1 0.0009	245	ене	1 0.0009
156	бор	1 0.0009	201	дел	5 0.0045	246	ени	3 0.0027
157	бос	1 0.0009	202	дер	1 0.0009	247	енн	4 0.0036
158	бы	2 0.0018	203	дет	1 0.0009	248	ень	5 0.0045
159	быв	2 0.0018	204	дил	2 0.0018	249	ере	1 0.0009
160	быт	1 0.0009	205	дит	1 0.0009	250	ерж	1 0.0009
161	бят	1 0.0009	206	дка	1 0.0009	251	ерн	1 0.0009
162	в д	1 0.0009	207	дко	1 0.0009	252	ес	1 0.0009
163	в п	2 0.0018	208	для	1 0.0009	253	есо	1 0.0009
164	в с	1 0.0009	209	дни	1 0.0009	254	есч	1 0.0009
165	в т	1 0.0009	210	дня	1 0.0009	255	ето	1 0.0009
166	в ч	1 0.0009	211	до	3 0.0027	256	еть	1 0.0009
167	в ш	1 0.0009	212	дол	1 0.0009	257	еха	1 0.0009
168	вак	1 0.0009	213	дре	1 0.0009	258	ехо	1 0.0009
169	вал	4 0.0036	214	дру	1 0.0009	259	ече	1 0.0009
170	вар	1 0.0009	215	дто	2 0.0018	260	ж с	1 0.0009
171	вас	1 0.0009	216	дуж	1 0.0009	261	жал	1 0.0009
172	ват	2 0.0018	217	дыс	1 0.0009	262	жан	1 0.0009
173	вая	1 0.0009	218	е в	1 0.0009	263	жды	1 0.0009
174	вед	1 0.0009	219	е д	1 0.0009	264	жел	1 0.0009
175	веж	1 0.0009	220	е е	1 0.0009	265	жен	3 0.0027
176	вел	1 0.0009	221	е з	1 0.0009	266	жив	1 0.0009
177	вид	1 0.0009	222	е к	1 0.0009	267	жил	2 0.0018
178	вил	1 0.0009	223	е п	1 0.0009	268	жни	1 0.0009
179	вин	1 0.0009	224	е с	1 0.0009	269	жно	1 0.0009
180	вки	1 0.0009	225	е т	2 0.0018	270	з п	1 0.0009

271 | за | 1 | 0.0009
 272 | зак | 1 | 0.0009
 273 | зар | 1 | 0.0009
 274 | згл | 1 | 0.0009
 275 | зго | 1 | 0.0009
 276 | zde | 1 | 0.0009
 277 | зно | 1 | 0.0009
 278 | зу | 1 | 0.0009
 279 | зы | 1 | 0.0009
 280 | и а | 1 | 0.0009
 281 | и б | 4 | 0.0036
 282 | и в | 1 | 0.0009
 283 | и д | 1 | 0.0009
 284 | и з | 1 | 0.0009
 285 | и к | 2 | 0.0018
 286 | и м | 1 | 0.0009
 287 | и о | 2 | 0.0018
 288 | и п | 1 | 0.0009
 289 | и с | 1 | 0.0009
 290 | и ш | 1 | 0.0009
 291 | ива | 1 | 0.0009
 292 | ида | 1 | 0.0009
 293 | иде | 1 | 0.0009
 294 | ие | 2 | 0.0018
 295 | ижн | 1 | 0.0009
 296 | из | 2 | 0.0018
 297 | изд | 1 | 0.0009
 298 | ий | 2 | 0.0018
 299 | ика | 1 | 0.0009
 300 | ики | 1 | 0.0009
 301 | ико | 1 | 0.0009
 302 | ил | 1 | 0.0009
 303 | ила | 4 | 0.0036
 304 | или | 2 | 0.0018
 305 | илс | 1 | 0.0009
 306 | има | 1 | 0.0009
 307 | инн | 1 | 0.0009
 308 | ино | 1 | 0.0009
 309 | ину | 1 | 0.0009
 310 | иск | 1 | 0.0009
 311 | исп | 1 | 0.0009
 312 | ист | 3 | 0.0027
 313 | ись | 3 | 0.0027
 314 | ица | 2 | 0.0018
 315 | ицу | 1 | 0.0009

316 | ишк | 1 | 0.0009
 317 | ишл | 1 | 0.0009
 318 | ишо | 1 | 0.0009
 319 | й б | 1 | 0.0009
 320 | й в | 2 | 0.0018
 321 | й д | 2 | 0.0018
 322 | й з | 1 | 0.0009
 323 | й к | 2 | 0.0018
 324 | й л | 1 | 0.0009
 325 | й п | 2 | 0.0018
 326 | й с | 1 | 0.0009
 327 | йни | 1 | 0.0009
 328 | к и | 1 | 0.0009
 329 | к к | 1 | 0.0009
 330 | к о | 1 | 0.0009
 331 | к п | 5 | 0.0045
 332 | к с | 2 | 0.0018
 333 | ка | 3 | 0.0027
 334 | каз | 2 | 0.0018
 335 | как | 1 | 0.0009
 336 | кал | 2 | 0.0018
 337 | кам | 1 | 0.0009
 338 | кат | 1 | 0.0009
 339 | ках | 1 | 0.0009
 340 | ква | 1 | 0.0009
 341 | ке | 1 | 0.0009
 342 | ки | 3 | 0.0027
 343 | кив | 1 | 0.0009
 344 | кли | 1 | 0.0009
 345 | ко | 1 | 0.0009
 346 | ког | 1 | 0.0009
 347 | кой | 1 | 0.0009
 348 | кор | 3 | 0.0027
 349 | кос | 1 | 0.0009
 350 | кот | 1 | 0.0009
 351 | коф | 3 | 0.0027
 352 | кра | 1 | 0.0009
 353 | кур | 2 | 0.0018
 354 | кут | 2 | 0.0018
 355 | кую | 1 | 0.0009
 356 | л в | 2 | 0.0018
 357 | л д | 2 | 0.0018
 358 | л м | 1 | 0.0009
 359 | л с | 1 | 0.0009
 360 | ла | 3 | 0.0027

361 | лад | 1 | 0.0009
 362 | лам | 1 | 0.0009
 363 | лас | 2 | 0.0018
 364 | лед | 1 | 0.0009
 365 | лез | 1 | 0.0009
 366 | лен | 2 | 0.0018
 367 | лех | 1 | 0.0009
 368 | лжн | 2 | 0.0018
 369 | ли | 3 | 0.0027
 370 | лиц | 1 | 0.0009
 371 | лка | 1 | 0.0009
 372 | лко | 1 | 0.0009
 373 | лме | 1 | 0.0009
 374 | ло | 1 | 0.0009
 375 | лос | 3 | 0.0027
 376 | лпе | 1 | 0.0009
 377 | лся | 3 | 0.0027
 378 | лта | 1 | 0.0009
 379 | луш | 1 | 0.0009
 380 | ль | 2 | 0.0018
 381 | льк | 1 | 0.0009
 382 | ляр | 1 | 0.0009
 383 | м г | 1 | 0.0009
 384 | м д | 2 | 0.0018
 385 | м и | 1 | 0.0009
 386 | м н | 2 | 0.0018
 387 | м о | 1 | 0.0009
 388 | м п | 1 | 0.0009
 389 | м т | 1 | 0.0009
 390 | м ч | 1 | 0.0009
 391 | ма | 2 | 0.0018
 392 | май | 1 | 0.0009
 393 | меи | 1 | 0.0009
 394 | мел | 1 | 0.0009
 395 | мер | 1 | 0.0009
 396 | ми | 4 | 0.0036
 397 | мку | 1 | 0.0009
 398 | мне | 1 | 0.0009
 399 | мню | 1 | 0.0009
 400 | мпа | 1 | 0.0009
 401 | мым | 1 | 0.0009
 402 | на | 7 | 0.0063
 403 | нак | 1 | 0.0009
 404 | нас | 1 | 0.0009
 405 | нев | 2 | 0.0018

406 | ней | 2 | 0.0018
407 | нек | 1 | 0.0009
408 | нем | 1 | 0.0009
409 | неп | 1 | 0.0009
410 | ни | 3 | 0.0027
411 | ник | 1 | 0.0009
412 | нию | 1 | 0.0009
413 | нно | 2 | 0.0018
414 | нны | 1 | 0.0009
415 | но | 1 | 0.0009
416 | нов | 1 | 0.0009
417 | ноз | 1 | 0.0009
418 | ной | 1 | 0.0009
419 | нск | 1 | 0.0009
420 | ну | 2 | 0.0018
421 | нут | 1 | 0.0009
422 | ную | 1 | 0.0009
423 | нчи | 1 | 0.0009
424 | нщи | 1 | 0.0009
425 | ны | 1 | 0.0009
426 | ные | 1 | 0.0009
427 | ным | 1 | 0.0009
428 | ных | 1 | 0.0009
429 | нюю | 1 | 0.0009
430 | о а | 1 | 0.0009
431 | о б | 1 | 0.0009
432 | о в | 1 | 0.0009
433 | о ж | 1 | 0.0009
434 | о к | 1 | 0.0009
435 | о н | 3 | 0.0027
436 | о п | 4 | 0.0036
437 | о с | 1 | 0.0009
438 | о т | 1 | 0.0009
439 | о х | 1 | 0.0009
440 | оби | 1 | 0.0009
441 | обо | 1 | 0.0009
442 | обр | 1 | 0.0009
443 | обс | 1 | 0.0009
444 | обь | 1 | 0.0009
445 | ов | 2 | 0.0018
446 | ова | 2 | 0.0018
447 | ове | 1 | 0.0009
448 | ово | 1 | 0.0009
449 | ову | 1 | 0.0009
450 | овы | 1 | 0.0009

451 | ога | 1 | 0.0009
452 | огд | 1 | 0.0009
453 | ого | 2 | 0.0018
454 | ода | 1 | 0.0009
455 | одв | 1 | 0.0009
456 | оди | 2 | 0.0018
457 | оды | 1 | 0.0009
458 | ое | 1 | 0.0009
459 | оем | 1 | 0.0009
460 | ожн | 1 | 0.0009
461 | озе | 1 | 0.0009
462 | озл | 1 | 0.0009
463 | оиц | 1 | 0.0009
464 | ой | 2 | 0.0018
465 | ойл | 1 | 0.0009
466 | око | 4 | 0.0036
467 | оле | 1 | 0.0009
468 | оло | 3 | 0.0027
469 | ом | 2 | 0.0018
470 | омо | 1 | 0.0009
471 | ому | 2 | 0.0018
472 | он | 1 | 0.0009
473 | она | 2 | 0.0018
474 | оно | 1 | 0.0009
475 | ону | 1 | 0.0009
476 | опа | 1 | 0.0009
477 | ор | 1 | 0.0009
478 | орб | 1 | 0.0009
479 | оре | 2 | 0.0018
480 | орн | 1 | 0.0009
481 | оро | 4 | 0.0036
482 | ору | 1 | 0.0009
483 | оры | 1 | 0.0009
484 | оса | 2 | 0.0018
485 | осе | 1 | 0.0009
486 | оси | 1 | 0.0009
487 | оск | 1 | 0.0009
488 | осл | 2 | 0.0018
489 | ост | 3 | 0.0027
490 | ось | 1 | 0.0009
491 | ося | 2 | 0.0018
492 | от | 1 | 0.0009
493 | оти | 1 | 0.0009
494 | отн | 2 | 0.0018
495 | ото | 1 | 0.0009

496 | оты | 1 | 0.0009
497 | оул | 1 | 0.0009
498 | офи | 6 | 0.0054
499 | офь | 2 | 0.0018
500 | оха | 2 | 0.0018
501 | охл | 1 | 0.0009
502 | пас | 1 | 0.0009
503 | пка | 1 | 0.0009
504 | пку | 1 | 0.0009
505 | пла | 1 | 0.0009
506 | по | 1 | 0.0009
507 | под | 3 | 0.0027
508 | пож | 1 | 0.0009
509 | пок | 1 | 0.0009
510 | пол | 1 | 0.0009
511 | пор | 1 | 0.0009
512 | пос | 2 | 0.0018
513 | пот | 2 | 0.0018
514 | про | 4 | 0.0036
515 | пун | 1 | 0.0009
516 | р с | 1 | 0.0009
517 | ра | 1 | 0.0009
518 | раз | 1 | 0.0009
519 | рам | 1 | 0.0009
520 | рас | 1 | 0.0009
521 | рат | 1 | 0.0009
522 | рах | 2 | 0.0018
523 | рбо | 1 | 0.0009
524 | рга | 1 | 0.0009
525 | ре | 1 | 0.0009
526 | ред | 2 | 0.0018
527 | рек | 1 | 0.0009
528 | рен | 1 | 0.0009
529 | рер | 1 | 0.0009
530 | рет | 1 | 0.0009
531 | рец | 1 | 0.0009
532 | рив | 1 | 0.0009
533 | риг | 2 | 0.0018
534 | рик | 2 | 0.0018
535 | рил | 1 | 0.0009
536 | рка | 1 | 0.0009
537 | рно | 1 | 0.0009
538 | рну | 1 | 0.0009
539 | рны | 1 | 0.0009
540 | ров | 1 | 0.0009

541	ро	1	0.0009	586	т х	1	0.0009	631	фию	1	0.0009
542	род	1	0.0009	587	так	2	0.0018	632	фия	2	0.0018
543	рок	8	0.0072	588	тал	1	0.0009	633	х в	1	0.0009
544	ром	1	0.0009	589	тар	2	0.0018	634	х д	1	0.0009
545	рон	2	0.0018	590	тат	1	0.0009	635	х к	1	0.0009
546	рос	1	0.0009	591	тах	1	0.0009	636	хам	1	0.0009
547	рот	1	0.0009	592	тво	2	0.0018	637	хла	1	0.0009
548	роу	1	0.0009	593	теп	1	0.0009	638	хли	1	0.0009
549	роч	1	0.0009	594	тех	2	0.0018	639	хну	1	0.0009
550	рош	1	0.0009	595	тия	1	0.0009	640	хов	1	0.0009
551	рск	1	0.0009	596	тко	1	0.0009	641	хоз	2	0.0018
552	рук	1	0.0009	597	тни	1	0.0009	642	хол	1	0.0009
553	рюк	1	0.0009	598	тно	1	0.0009	643	хон	1	0.0009
554	ряд	1	0.0009	599	то	3	0.0027	644	хру	1	0.0009
555	рял	1	0.0009	600	тор	2	0.0018	645	хут	3	0.0027
556	рят	1	0.0009	601	точ	1	0.0009	646	цеп	1	0.0009
557	с л	1	0.0009	602	тре	1	0.0009	647	чал	1	0.0009
558	с н	1	0.0009	603	тро	1	0.0009	648	чан	1	0.0009
559	с т	1	0.0009	604	туп	1	0.0009	649	чат	1	0.0009
560	с х	1	0.0009	605	тур	1	0.0009	650	чин	1	0.0009
561	сил	1	0.0009	606	тча	1	0.0009	651	чит	1	0.0009
562	ски	1	0.0009	607	ть	1	0.0009	652	чно	1	0.0009
563	ско	2	0.0018	608	у в	1	0.0009	653	чок	1	0.0009
564	скр	1	0.0009	609	у н	1	0.0009	654	что	1	0.0009
565	сле	1	0.0009	610	у ч	1	0.0009	655	чуб	1	0.0009
566	слу	1	0.0009	611	уби	1	0.0009	656	чуж	1	0.0009
567	сме	1	0.0009	612	уга	1	0.0009	657	чул	1	0.0009
568	сох	1	0.0009	613	уги	1	0.0009	658	шад	1	0.0009
569	спа	1	0.0009	614	удн	1	0.0009	659	шар	2	0.0018
570	спо	1	0.0009	615	удт	1	0.0009	660	шел	2	0.0018
571	спу	1	0.0009	616	уды	1	0.0009	661	шеп	1	0.0009
572	сск	1	0.0009	617	ула	2	0.0018	662	шиб	1	0.0009
573	ста	4	0.0036	618	упа	1	0.0009	663	шие	1	0.0009
574	ств	1	0.0009	619	ург	1	0.0009	664	шил	1	0.0009
575	сто	2	0.0018	620	урч	1	0.0009	665	шке	1	0.0009
576	стр	1	0.0009	621	ута	1	0.0009	666	шни	1	0.0009
577	сту	1	0.0009	622	утв	1	0.0009	667	шны	1	0.0009
578	сть	1	0.0009	623	уто	3	0.0027	668	шок	1	0.0009
579	стя	1	0.0009	624	уха	1	0.0009	669	шь	1	0.0009
580	сши	1	0.0009	625	учи	1	0.0009	670	ы в	1	0.0009
581	сын	1	0.0009	626	ушк	1	0.0009	671	ы з	1	0.0009
582	сып	1	0.0009	627	уще	1	0.0009	672	ы о	1	0.0009
583	съ	2	0.0018	628	ую	1	0.0009	673	ы п	1	0.0009
584	ся	3	0.0027	629	ующ	1	0.0009	674	ы р	1	0.0009
585	т м	1	0.0009	630	фий	1	0.0009	675	ы ва	1	0.0009

676 | ые | 1 | 0.0009
677 | ый | 1 | 0.0009
678 | ыла | 1 | 0.0009
679 | ыло | 1 | 0.0009
680 | ыль | 1 | 0.0009
681 | ым | 1 | 0.0009
682 | ымя | 1 | 0.0009
683 | ыня | 1 | 0.0009
684 | ыты | 1 | 0.0009
685 | ых | 1 | 0.0009
686 | ь ж | 1 | 0.0009
687 | ь м | 1 | 0.0009
688 | ь н | 1 | 0.0009
689 | ь п | 2 | 0.0018
690 | ь у | 1 | 0.0009

691 | ьев | 1 | 0.0009
692 | ьей | 1 | 0.0009
693 | ька | 1 | 0.0009
694 | ьки | 1 | 0.0009
695 | ьку | 1 | 0.0009
696 | ьма | 1 | 0.0009
697 | это | 1 | 0.0009
698 | ю б | 1 | 0.0009
699 | ю в | 1 | 0.0009
700 | ю г | 1 | 0.0009
701 | ю з | 1 | 0.0009
702 | ю к | 1 | 0.0009
703 | ю п | 1 | 0.0009
704 | юлю | 1 | 0.0009
705 | ютс | 1 | 0.0009

706 | я к | 1 | 0.0009
707 | я м | 1 | 0.0009
708 | я о | 1 | 0.0009
709 | я с | 1 | 0.0009
710 | я т | 1 | 0.0009
711 | яде | 2 | 0.0018
712 | яин | 1 | 0.0009
713 | яйс | 1 | 0.0009
714 | яки | 1 | 0.0009
715 | яли | 1 | 0.0009
716 | янн | 1 | 0.0009
717 | ясн | 1 | 0.0009
718 | ясь | 1 | 0.0009
719 | яя | 1 | 0.0009

Сумма вероятностей: 1.0000

Приложение 6: листинг приближения 4-го порядка

Замечание. При использовании этой таблицы: первый столбец – номер 4-грамм, которые сортируются по алфавиту; второй столбец - сама 4-грамма; третий столбец – количество появлений этой 4-граммы; а четвёртый столбец – вероятность появлений этой 4-граммы.

Проверка правильности подсчетов: сумма всех вероятностей равна 1 – то есть вычисление выполнено корректно.

0		вы		1		0.000766	45		зап		1		0.000766	90		пас		1		0.000766
1		ли		1		0.000766	46		зор		1		0.000766	91		пах		2		0.001531
2		ма		1		0.000766	47		и к		2		0.001531	92		пер		1		0.000766
3		ажн		1		0.000766	48		и н		1		0.000766	93		пит		1		0.000766
4		арб		1		0.000766	49		и с		1		0.000766	94		пле		1		0.000766
5		баб		2		0.001531	50		из		1		0.000766	95		пло		1		0.000766
6		баз		1		0.000766	51		иму		1		0.000766	96		по		4		0.003063
7		бел		1		0.000766	52		ино		1		0.000766	97		под		1		0.000766
8		бир		1		0.000766	53		к о		1		0.000766	98		пок		1		0.000766
9		бор		2		0.001531	54		каз		3		0.002297	99		пор		1		0.000766
10		бро		1		0.000766	55		как		2		0.001531	100		пос		1		0.000766
11		буд		1		0.000766	56		кам		2		0.001531	101		пот		1		0.000766
12		быв		1		0.000766	57		кат		1		0.000766	102		пре		1		0.000766
13		в б		1		0.000766	58		кис		1		0.000766	103		при		2		0.001531
14		в к		1		0.000766	59		ког		1		0.000766	104		про		8		0.006126
15		в с		1		0.000766	60		кур		4		0.003063	105		пря		1		0.000766
16		в х		2		0.001531	61		лад		1		0.000766	106		пух		1		0.000766
17		в ч		1		0.000766	62		лиц		1		0.000766	107		рад		1		0.000766
18		в ш		1		0.000766	63		лиш		1		0.000766	108		рас		2		0.001531
19		вел		1		0.000766	64		май		1		0.000766	109		реб		1		0.000766
20		вер		1		0.000766	65		мал		2		0.001531	110		ред		2		0.001531
21		веч		1		0.000766	66		мед		1		0.000766	111		род		1		0.000766
22		вид		2		0.001531	67		мел		2		0.001531	112		рук		2		0.001531
23		все		2		0.001531	68		на		5		0.003828	113		с и		1		0.000766
24		вск		1		0.000766	69		не		3		0.002297	114		с н		1		0.000766
25		всл		1		0.000766	70		нев		1		0.000766	115		сам		1		0.000766
26		выс		1		0.000766	71		нег		2		0.001531	116		сво		1		0.000766
27		вян		1		0.000766	72		ней		1		0.000766	117		сго		1		0.000766
28		гла		1		0.000766	73		нем		1		0.000766	118		сде		1		0.000766
29		гол		2		0.001531	74		неп		2		0.001531	119		сжи		1		0.000766
30		гут		1		0.000766	75		нес		1		0.000766	120		ско		2		0.001531
31		да		1		0.000766	76		но		1		0.000766	121		ску		1		0.000766
32		дал		1		0.000766	77		нов		1		0.000766	122		сме		1		0.000766
33		для		1		0.000766	78		нос		1		0.000766	123		сру		1		0.000766
34		до		3		0.002297	79		оби		1		0.000766	124		ста		1		0.000766
35		дон		1		0.000766	80		обс		1		0.000766	125		сто		1		0.000766
36		его		2		0.001531	81		оди		1		0.000766	126		сын		1		0.000766
37		ее		1		0.000766	82		он		3		0.002297	127		тат		1		0.000766
38		жел		1		0.000766	83		она		1		0.000766	128		тел		1		0.000766
39		жен		4		0.003063	84		они		1		0.000766	129		той		1		0.000766
40		жил		1		0.000766	85		орд		1		0.000766	130		тос		1		0.000766
41		за		2		0.001531	86		осе		1		0.000766	131		тур		3		0.002297
42		заб		1		0.000766	87		от		1		0.000766	132		у д		1		0.000766
43		зав		1		0.000766	88		отд		1		0.000766	133		у н		1		0.000766
44		зак		1		0.000766	89		отш		1		0.000766	134		уве		1		0.000766

135		узо		1		0.000766	180		акут		1		0.000766	225		бабь		1		0.000766
136		ули		1		0.000766	181		ал в		1		0.000766	226		базы		1		0.000766
137		улю		1		0.000766	182		ал о		1		0.000766	227		бату		1		0.000766
138		ход		1		0.000766	183		ал п		1		0.000766	228		бе у		1		0.000766
139		хоз		1		0.000766	184		ала		3		0.002297	229		беле		1		0.000766
140		хру		1		0.000766	185		алек		1		0.000766	230		биды		1		0.000766
141		хут		4		0.003063	186		ален		1		0.000766	231		били		1		0.000766
142		чек		1		0.000766	187		али		3		0.002297	232		бирю		1		0.000766
143		чер		1		0.000766	188		алис		3		0.002297	233		блен		1		0.000766
144		чуд		1		0.000766	189		алые		1		0.000766	234		бой		1		0.000766
145		шал		2		0.001531	190		аль		2		0.001531	235		боро		2		0.001531
146		шел		3		0.002297	191		ам п		1		0.000766	236		бров		1		0.000766
147		а ар		1		0.000766	192		амен		1		0.000766	237		бстр		1		0.000766
148		а би		1		0.000766	193		ами		3		0.002297	238		будт		1		0.000766
149		а в		1		0.000766	194		ампа		1		0.000766	239		бы о		1		0.000766
150		а вя		1		0.000766	195		ана		1		0.000766	240		быва		2		0.001531
151		а до		1		0.000766	196		ане		1		0.000766	241		бью		1		0.000766
152		а ка		1		0.000766	197		анию		1		0.000766	242		бяти		1		0.000766
153		а ли		1		0.000766	198		анка		1		0.000766	243		в бо		1		0.000766
154		а ма		1		0.000766	199		анно		1		0.000766	244		в вс		1		0.000766
155		а не		1		0.000766	200		анну		1		0.000766	245		в ку		1		0.000766
156		а но		1		0.000766	201		апах		1		0.000766	246		в пр		2		0.001531
157		а от		1		0.000766	202		арбо		1		0.000766	247		в св		1		0.000766
158		а па		1		0.000766	203		арик		1		0.000766	248		в ху		2		0.001531
159		а пр		4		0.003063	204		арил		1		0.000766	249		в че		2		0.001531
160		а ру		1		0.000766	205		арск		1		0.000766	250		в ша		1		0.000766
161		а ст		1		0.000766	206		аспа		1		0.000766	251		ваки		1		0.000766
162		а ул		1		0.000766	207		асск		1		0.000766	252		вал		1		0.000766
163		а ше		1		0.000766	208		асши		1		0.000766	253		вали		2		0.001531
164		абы		1		0.000766	209		ась		1		0.000766	254		вая		3		0.002297
165		абыв		1		0.000766	210		ат у		1		0.000766	255		ведо		1		0.000766
166		абыю		1		0.000766	211		атал		1		0.000766	256		вей		1		0.000766
167		авис		1		0.000766	212		атар		1		0.000766	257		вел		2		0.001531
168		адон		1		0.000766	213		атую		1		0.000766	258		вели		1		0.000766
169		адуж		1		0.000766	214		ах н		1		0.000766	259		верн		1		0.000766
170		ажни		1		0.000766	215		ахам		1		0.000766	260		вече		1		0.000766
171		аза		1		0.000766	216		ахла		1		0.000766	261		виде		2		0.001531
172		азак		2		0.001531	217		ахну		1		0.000766	262		вижн		1		0.000766
173		азач		1		0.000766	218		ахот		1		0.000766	263		вист		1		0.000766
174		азы		1		0.000766	219		ачат		1		0.000766	264		во с		1		0.000766
175		азыв		2		0.001531	220		ашне		1		0.000766	265		вое		1		0.000766
176		айда		1		0.000766	221		ая о		1		0.000766	266		воем		1		0.000766
177		ак м		1		0.000766	222		ая т		2		0.001531	267		вом		1		0.000766
178		ак п		2		0.001531	223		ая ш		1		0.000766	268		все		1		0.000766
179		аки		2		0.001531	224		бабы		1		0.000766	269		всег		1		0.000766

270		вско		1		0.000766	315		дто		1		0.000766	360		елик		1		0.000766
271		всле		1		0.000766	316		дужн		1		0.000766	361		елил		1		0.000766
272		ву		1		0.000766	317		ды г		1		0.000766	362		елко		1		0.000766
273		высы		1		0.000766	318		дыпр		1		0.000766	363		елят		1		0.000766
274		вяну		1		0.000766	319		е бы		1		0.000766	364		ем к		1		0.000766
275		гана		1		0.000766	320		е гл		1		0.000766	365		емку		1		0.000766
276		гда		1		0.000766	321		е жи		1		0.000766	366		емыт		1		0.000766
277		гдаш		1		0.000766	322		е за		2		0.001531	367		ене		1		0.000766
278		глаз		1		0.000766	323		е на		1		0.000766	368		ени		1		0.000766
279		го а		1		0.000766	324		е не		1		0.000766	369		енин		1		0.000766
280		го в		1		0.000766	325		е од		1		0.000766	370		енна		1		0.000766
281		го н		1		0.000766	326		е от		2		0.001531	371		енно		1		0.000766
282		го п		2		0.001531	327		е пи		1		0.000766	372		енну		1		0.000766
283		голо		2		0.001531	328		е ре		1		0.000766	373		енны		1		0.000766
284		гоно		1		0.000766	329		е сж		1		0.000766	374		ену		3		0.002297
285		горб		1		0.000766	330		е у		1		0.000766	375		енщи		1		0.000766
286		горо		1		0.000766	331		е уз		1		0.000766	376		ень		3		0.002297
287		гута		1		0.000766	332		е хо		2		0.001531	377		еньк		1		0.000766
288		д но		1		0.000766	333		ебят		1		0.000766	378		епод		1		0.000766
289		д ск		1		0.000766	334		евед		1		0.000766	379		епок		1		0.000766
290		да в		1		0.000766	335		егда		1		0.000766	380		ерам		1		0.000766
291		да н		1		0.000766	336		его		4		0.003063	381		ерек		1		0.000766
292		да п		1		0.000766	337		ед н		1		0.000766	382		ержа		1		0.000766
293		дале		1		0.000766	338		едко		2		0.001531	383		ерно		1		0.000766
294		дане		1		0.000766	339		едле		1		0.000766	384		ерну		1		0.000766
295		дашн		1		0.000766	340		едню		1		0.000766	385		ерти		1		0.000766
296		движ		1		0.000766	341		едом		1		0.000766	386		ес б		1		0.000766
297		де с		1		0.000766	342		едпо		1		0.000766	387		есоч		1		0.000766
298		дели		3		0.002297	343		ее п		1		0.000766	388		еств		1		0.000766
299		держ		1		0.000766	344		еж к		1		0.000766	389		етчи		1		0.000766
300		дил		2		0.001531	345		еива		1		0.000766	390		ехов		2		0.001531
301		дича		1		0.000766	346		ей з		1		0.000766	391		ецку		1		0.000766
302		дко		2		0.001531	347		ей н		1		0.000766	392		ечер		1		0.000766
303		длен		1		0.000766	348		ей п		1		0.000766	393		ж ка		1		0.000766
304		для		1		0.000766	349		еким		1		0.000766	394		жанн		1		0.000766
305		дное		1		0.000766	350		екли		1		0.000766	395		желв		1		0.000766
306		дных		1		0.000766	351		екме		1		0.000766	396		жени		1		0.000766
307		днюю		1		0.000766	352		ел м		1		0.000766	397		жену		3		0.002297
308		до в		1		0.000766	353		ел н		1		0.000766	398		женщ		1		0.000766
309		до с		1		0.000766	354		ел о		1		0.000766	399		жил		1		0.000766
310		до т		1		0.000766	355		ел с		1		0.000766	400		жима		1		0.000766
311		домы		1		0.000766	356		елва		1		0.000766	401		жник		1		0.000766
312		дона		1		0.000766	357		елес		1		0.000766	402		жнос		1		0.000766
313		дони		1		0.000766	358		елех		2		0.001531	403		жные		1		0.000766
314		дпос		1		0.000766	359		ели		2		0.001531	404		з ту		1		0.000766

405	за а 1 0.000766	450	ие з 1 0.000766	495	й об 1 0.000766
406	за п 2 0.001531	451	ие о 1 0.000766	496	й по 1 0.000766
407	за бы 1 0.000766	452	ижно 1 0.000766	497	й пр 1 0.000766
408	зави 1 0.000766	453	из т 1 0.000766	498	й ру 1 0.000766
409	зак 1 0.000766	454	ий в 1 0.000766	499	й с 1 0.000766
410	заки 1 0.000766	455	ий и 1 0.000766	500	йдан 1 0.000766
411	заку 1 0.000766	456	ий о 1 0.000766	501	йств 1 0.000766
412	запа 1 0.000766	457	ик к 1 0.000766	502	к ку 1 0.000766
413	зача 1 0.000766	458	ик м 1 0.000766	503	к ме 2 0.001531
414	zde 1 0.000766	459	ика 1 0.000766	504	к ос 1 0.000766
415	земк 1 0.000766	460	икал 1 0.000766	505	к по 1 0.000766
416	зори 1 0.000766	461	ики 1 0.000766	506	к пр 1 0.000766
417	зоры 1 0.000766	462	ил б 1 0.000766	507	ка к 1 0.000766
418	зы д 1 0.000766	463	ил в 1 0.000766	508	ка с 1 0.000766
419	зыва 2 0.001531	464	ил д 1 0.000766	509	каза 3 0.002297
420	зяйс 1 0.000766	465	ил ж 1 0.000766	510	казы 2 0.001531
421	и ба 1 0.000766	466	ил с 1 0.000766	511	как 2 0.001531
422	и бр 1 0.000766	467	илас 1 0.000766	512	кала 1 0.000766
423	и бу 1 0.000766	468	или 2 0.001531	513	кали 1 0.000766
424	и да 1 0.000766	469	ился 1 0.000766	514	каме 1 0.000766
425	и ег 1 0.000766	470	имал 1 0.000766	515	камп 1 0.000766
426	и за 1 0.000766	471	ими 1 0.000766	516	ката 1 0.000766
427	и и 1 0.000766	472	имущ 1 0.000766	517	ках 1 0.000766
428	и к 1 0.000766	473	иноз 1 0.000766	518	ки д 1 0.000766
429	и ка 2 0.001531	474	иной 1 0.000766	519	ки н 1 0.000766
430	и ко 1 0.000766	475	ину 1 0.000766	520	ки п 1 0.000766
431	и ку 1 0.000766	476	ины 2 0.001531	521	ки с 2 0.001531
432	и на 3 0.002297	477	ирюк 1 0.000766	522	кими 1 0.000766
433	и не 3 0.002297	478	исто 1 0.000766	523	кист 1 0.000766
434	и он 1 0.000766	479	истъ 2 0.001531	524	клик 1 0.000766
435	и па 1 0.000766	480	ись 3 0.002297	525	кмен 1 0.000766
436	и пр 1 0.000766	481	итал 1 0.000766	526	ко в 1 0.000766
437	и ра 1 0.000766	482	ицо 1 0.000766	527	ко п 1 0.000766
438	и сд 1 0.000766	483	ицу 1 0.000766	528	кова 1 0.000766
439	и ср 1 0.000766	484	ичал 1 0.000766	529	когд 1 0.000766
440	и ст 1 0.000766	485	ишки 1 0.000766	530	кого 1 0.000766
441	и у 1 0.000766	486	ишь 1 0.000766	531	ком 1 0.000766
442	и ув 1 0.000766	487	ию в 2 0.001531	532	коре 1 0.000766
443	и хр 1 0.000766	488	ия и 1 0.000766	533	корн 1 0.000766
444	ибе 1 0.000766	489	й бо 1 0.000766	534	коро 1 0.000766
445	ивал 1 0.000766	490	й ве 1 0.000766	535	коти 1 0.000766
446	ивел 1 0.000766	491	й за 1 0.000766	536	кофе 5 0.003828
447	игор 1 0.000766	492	й из 1 0.000766	537	ку же 1 0.000766
448	идел 2 0.001531	493	й ла 1 0.000766	538	кула 1 0.000766
449	идып 1 0.000766	494	й не 1 0.000766	539	кург 1 0.000766

540 куре 3 0.002297	584 лову 1 0.000766	628 на о 1 0.000766
541 кута 1 0.000766	585 лоси 1 0.000766	629 на п 1 0.000766
542 кую 3 0.002297	586 лотн 1 0.000766	630 на р 1 0.000766
543 кующ 1 0.000766	587 лся 2 0.001531	631 на у 1 0.000766
544 л ба 1 0.000766	588 лые 1 0.000766	632 ная 1 0.000766
545 л в 2 0.001531	589 ль д 1 0.000766	633 не б 1 0.000766
546 л до 1 0.000766	590 ль ж 1 0.000766	634 не ж 1 0.000766
547 л же 1 0.000766	591 люка 1 0.000766	635 не з 1 0.000766
548 л ме 1 0.000766	592 люлю 1 0.000766	636 не н 1 0.000766
549 л на 1 0.000766	593 ля с 1 0.000766	637 не х 1 0.000766
550 л он 2 0.001531	594 лят 1 0.000766	638 не ве 1 0.000766
551 л по 1 0.000766	595 м гу 1 0.000766	639 него 2 0.001531
552 л с 1 0.000766	596 м ку 1 0.000766	640 ней 2 0.001531
553 л сы 1 0.000766	597 м по 1 0.000766	641 немь 1 0.000766
554 ла д 1 0.000766	598 м пр 1 0.000766	642 непо 2 0.001531
555 ла л 1 0.000766	599 м те 1 0.000766	643 нес 1 0.000766
556 ла п 1 0.000766	600 майд 1 0.000766	644 ни к 1 0.000766
557 ла ш 1 0.000766	601 мал 1 0.000766	645 ни у 1 0.000766
558 ладо 1 0.000766	602 мала 1 0.000766	646 ни х 1 0.000766
559 лаза 1 0.000766	603 мале 1 0.000766	647 ник 1 0.000766
560 ламы 1 0.000766	604 медл 1 0.000766	648 ники 1 0.000766
561 лась 1 0.000766	605 меж 1 0.000766	649 нила 1 0.000766
562 лвак 1 0.000766	606 меив 1 0.000766	650 нино 1 0.000766
563 лед 1 0.000766	607 меле 2 0.001531	651 нию 1 0.000766
564 ледн 1 0.000766	608 менн 1 0.000766	652 нка 1 0.000766
565 леки 1 0.000766	609 мень 1 0.000766	653 нная 1 0.000766
566 ленн 3 0.002297	610 мерт 1 0.000766	654 нно 2 0.001531
567 лень 1 0.000766	611 ми з 1 0.000766	655 нную 2 0.001531
568 лесо 1 0.000766	612 ми к 1 0.000766	656 нных 1 0.000766
569 лехо 2 0.001531	613 ми н 1 0.000766	657 но к 1 0.000766
570 ли б 2 0.001531	614 ми р 1 0.000766	658 но н 1 0.000766
571 ли е 1 0.000766	615 ми у 1 0.000766	659 но о 1 0.000766
572 ли и 1 0.000766	616 мкуж 1 0.000766	660 но п 1 0.000766
573 ли к 1 0.000766	617 мпан 1 0.000766	661 ново 1 0.000766
574 ли н 1 0.000766	618 муще 1 0.000766	662 ное 1 0.000766
575 ли о 1 0.000766	619 мыми 1 0.000766	663 нозе 1 0.000766
576 ли п 1 0.000766	620 мыты 1 0.000766	664 ной 3 0.002297
577 лика 2 0.001531	621 н же 1 0.000766	665 ном 1 0.000766
578 лил 1 0.000766	622 н и 1 0.000766	666 носи 1 0.000766
579 лись 3 0.002297	623 н ра 1 0.000766	667 ност 1 0.000766
580 лицо 1 0.000766	624 на б 1 0.000766	668 ну 1 0.000766
581 лицу 1 0.000766	625 на в 1 0.000766	669 ну д 1 0.000766
582 лишь 1 0.000766	626 на м 1 0.000766	670 ну о 1 0.000766
583 лков 1 0.000766	627 на н 1 0.000766	671 ну ш 1 0.000766

672	нув	1	0.000766	717	одил	2	0.001531	762	осен	1	0.000766
673	нулс	1	0.000766	718	одич	1	0.000766	763	осил	1	0.000766
674	нут	1	0.000766	719	одны	1	0.000766	764	осис	1	0.000766
675	ную	2	0.001531	720	оды	1	0.000766	765	оску	1	0.000766
676	нщин	1	0.000766	721	ое р	1	0.000766	766	осле	1	0.000766
677	ны и	1	0.000766	722	ое х	1	0.000766	767	осме	1	0.000766
678	ны п	1	0.000766	723	оем	1	0.000766	768	ости	1	0.000766
679	ные	1	0.000766	724	озде	1	0.000766	769	осту	1	0.000766
680	ных	2	0.001531	725	озем	1	0.000766	770	от м	1	0.000766
681	нь е	1	0.000766	726	озяй	1	0.000766	771	отде	1	0.000766
682	нь с	1	0.000766	727	оилс	1	0.000766	772	отин	1	0.000766
683	нь ш	1	0.000766	728	ой б	1	0.000766	773	отни	1	0.000766
684	ньку	1	0.000766	729	ой л	1	0.000766	774	отно	1	0.000766
685	нюю	1	0.000766	730	ой п	1	0.000766	775	отс	1	0.000766
686	о аж	1	0.000766	731	ой р	1	0.000766	776	отши	1	0.000766
687	о в	1	0.000766	732	ой с	1	0.000766	777	офий	3	0.002297
688	о ве	1	0.000766	733	оказ	1	0.000766	778	офию	1	0.000766
689	о ви	2	0.001531	734	окор	1	0.000766	779	офия	1	0.000766
690	о вс	1	0.000766	735	окоф	5	0.003828	780	очуб	1	0.000766
691	о ка	1	0.000766	736	олов	1	0.000766	781	пал	1	0.000766
692	о не	3	0.002297	737	олос	1	0.000766	782	пали	1	0.000766
693	о он	1	0.000766	738	ом г	1	0.000766	783	пани	1	0.000766
694	о па	1	0.000766	739	ом п	1	0.000766	784	пасш	1	0.000766
695	о пе	1	0.000766	740	ом т	1	0.000766	785	паха	1	0.000766
696	о пл	1	0.000766	741	омеж	1	0.000766	786	пахл	1	0.000766
697	о по	3	0.002297	742	омым	1	0.000766	787	пахн	1	0.000766
698	о пу	1	0.000766	743	он ж	1	0.000766	788	пахо	1	0.000766
699	о ре	1	0.000766	744	он и	1	0.000766	789	пере	1	0.000766
700	о сг	1	0.000766	745	он р	1	0.000766	790	пита	1	0.000766
701	о см	1	0.000766	746	она	2	0.001531	791	пкую	1	0.000766
702	о та	1	0.000766	747	они	2	0.001531	792	плен	1	0.000766
703	о ху	2	0.001531	748	онил	1	0.000766	793	плот	1	0.000766
704	обид	1	0.000766	749	оном	1	0.000766	794	по в	1	0.000766
705	обст	1	0.000766	750	ор к	1	0.000766	795	по п	1	0.000766
706	ов в	1	0.000766	751	орбл	1	0.000766	796	по х	2	0.001531
707	ов п	1	0.000766	752	орда	1	0.000766	797	под	1	0.000766
708	овая	1	0.000766	753	оре	2	0.001531	798	подв	1	0.000766
709	овей	1	0.000766	754	ори	1	0.000766	799	пока	1	0.000766
710	овое	1	0.000766	755	орно	1	0.000766	800	поко	1	0.000766
711	ову	1	0.000766	756	оро	1	0.000766	801	поры	1	0.000766
712	огда	1	0.000766	757	ород	2	0.001531	802	посл	1	0.000766
713	ого	1	0.000766	758	ороз	1	0.000766	803	посм	1	0.000766
714	огон	1	0.000766	759	орон	1	0.000766	804	потс	1	0.000766
715	од с	1	0.000766	760	ору	2	0.001531	805	пред	1	0.000766
716	одви	1	0.000766	761	оры	2	0.001531	806	прив	1	0.000766

807	приг 1 0.000766	852	роко 5 0.003828	897	ство 2 0.001531
808	про 1 0.000766	853	роме 1 0.000766	898	сти 1 0.000766
809	прог 1 0.000766	854	рони 1 0.000766	899	сто 1 0.000766
810	прок 5 0.003828	855	рост 1 0.000766	900	стор 1 0.000766
811	пром 1 0.000766	856	рско 1 0.000766	901	стро 1 0.000766
812	прос 1 0.000766	857	рти 1 0.000766	902	ступ 1 0.000766
813	прят 1 0.000766	858	ру 1 0.000766	903	сть 2 0.001531
814	пухл 1 0.000766	859	ру ч 1 0.000766	904	сшие 1 0.000766
815	р ка 1 0.000766	860	руби 1 0.000766	905	сына 1 0.000766
816	раду 1 0.000766	861	рука 1 0.000766	906	сыпа 1 0.000766
817	рами 1 0.000766	862	руки 1 0.000766	907	съ б 1 0.000766
818	расп 1 0.000766	863	рупк 1 0.000766	908	съ в 1 0.000766
819	расс 1 0.000766	864	рчан 1 0.000766	909	съ ж 1 0.000766
820	рбле 1 0.000766	865	ры е 1 0.000766	910	съ р 1 0.000766
821	рбой 1 0.000766	866	ры р 1 0.000766	911	ся в 1 0.000766
822	рган 1 0.000766	867	рюко 1 0.000766	912	ся с 1 0.000766
823	рда 1 0.000766	868	рята 1 0.000766	913	т зо 1 0.000766
824	ре н 1 0.000766	869	с бе 1 0.000766	914	т ма 1 0.000766
825	ре о 1 0.000766	870	с им 1 0.000766	915	т ра 1 0.000766
826	ребя 1 0.000766	871	с не 1 0.000766	916	т ул 1 0.000766
827	редк 2 0.001531	872	с то 1 0.000766	917	тала 1 0.000766
828	редп 1 0.000766	873	сам 1 0.000766	918	тали 2 0.001531
829	рекл 1 0.000766	874	свое 1 0.000766	919	танн 1 0.000766
830	рене 1 0.000766	875	сгор 1 0.000766	920	тари 2 0.001531
831	рень 2 0.001531	876	сдер 1 0.000766	921	тарс 1 0.000766
832	ретч 1 0.000766	877	се о 1 0.000766	922	тата 1 0.000766
833	рецк 1 0.000766	878	сегд 1 0.000766	923	тво 1 0.000766
834	ржан 1 0.000766	879	сени 1 0.000766	924	твои 1 0.000766
835	ри н 1 0.000766	880	сжим 1 0.000766	925	тдел 1 0.000766
836	риве 1 0.000766	881	сил 1 0.000766	926	теля 1 0.000766
837	риго 1 0.000766	882	сист 1 0.000766	927	ти б 1 0.000766
838	рик 1 0.000766	883	сказ 1 0.000766	928	ти н 1 0.000766
839	рили 1 0.000766	884	ског 1 0.000766	929	тины 1 0.000766
840	рно 1 0.000766	885	скор 2 0.001531	930	тишк 1 0.000766
841	рной 1 0.000766	886	скот 1 0.000766	931	тник 1 0.000766
842	рнул 1 0.000766	887	скул 1 0.000766	932	тной 1 0.000766
843	ро н 1 0.000766	888	скую 1 0.000766	933	то в 1 0.000766
844	ро п 1 0.000766	889	след 2 0.001531	934	то п 1 0.000766
845	рове 1 0.000766	890	смеи 1 0.000766	935	той 1 0.000766
846	рого 1 0.000766	891	смер 1 0.000766	936	тор 1 0.000766
847	роди 1 0.000766	892	сочу 1 0.000766	937	торе 1 0.000766
848	родн 1 0.000766	893	спаш 1 0.000766	938	торо 1 0.000766
849	роды 1 0.000766	894	сруб 1 0.000766	939	тору 2 0.001531
850	розд 1 0.000766	895	сска 1 0.000766	940	тоск 1 0.000766
851	роил 1 0.000766	896	стар 1 0.000766	941	трои 1 0.000766

942		тс		т		1		0.000766	987		ухли		1		0.000766	1032		щест		1		0.000766				
943		тупа		1		0.000766			988		ущес		1		0.000766	1033		щие		1		0.000766				
944		туре		2		0.001531			989		ую		в		1		0.000766	1034		щину		1		0.000766		
945		турч		1		0.000766			990		ую		г		1		0.000766	1035		ы		го		1		0.000766
946		тую		1		0.000766			991		ую		з		1		0.000766	1036		ы		дл		1		0.000766
947		тчин		1		0.000766			992		ую		и		1		0.000766	1037		ы		ее		1		0.000766
948		тшиб		1		0.000766			993		ую		к		2		0.001531	1038		ы		и		1		0.000766
949		тых		1		0.000766			994		ующи		1		0.000766	1039		ы		ор		1		0.000766		
950		ть		ж		1		0.000766	995		фий		3		0.002297	1040		ы		пр		1		0.000766		
951		ть		п		1		0.000766	996		фию		1		0.000766	1041		ы		ре		1		0.000766		
952		у		в		1		0.000766	997		фия		1		0.000766	1042		ывал		2		0.001531				
953		у		л		1		0.000766	998		х		ка		1		0.000766	1043		ывая		2		0.001531		
954		у		м		1		0.000766	999		х		но		1		0.000766	1044		ые		г		1		0.000766
955		у		вс		1		0.000766	1000		х		по		1		0.000766	1045		ые		у		1		0.000766
956		у		до		2		0.001531	1001		х		пр		1		0.000766	1046		ыми		1		0.000766		
957		у		не		1		0.000766	1002		хами		1		0.000766	1047		ына		1		0.000766				
958		у		он		1		0.000766	1003		хла		1		0.000766	1048		ыпал		1		0.000766				
959		у		чу		1		0.000766	1004		хли		1		0.000766	1049		ыпро		1		0.000766				
960		у		ше		1		0.000766	1005		хнув		1		0.000766	1050		ысып		1		0.000766				
961		убат		1		0.000766			1006		хов		2		0.001531	1051		ытых		1		0.000766				
962		убил		1		0.000766			1007		ходи		1		0.000766	1052		ых		к		1		0.000766		
963		ув		ч		1		0.000766	1008		хозя		1		0.000766	1053		ых		п		2		0.001531		
964		увел		1		0.000766			1009		хотн		1		0.000766	1054		ь		ба		1		0.000766		
965		удно		1		0.000766			1010		хруп		1		0.000766	1055		ь		в		1		0.000766		
966		удто		1		0.000766			1011		хуто		4		0.003063	1056		ь		да		1		0.000766		
967		ужен		1		0.000766			1012		цкую		1		0.000766	1057		ь		ег		1		0.000766		
968		ужны		1		0.000766			1013		цо		р		1		0.000766	1058		ь		же		3		0.002297
969		узор		1		0.000766			1014		цу		в		1		0.000766	1059		ь		пл		1		0.000766
970		уках		1		0.000766			1015		чалы		1		0.000766	1060		ь		по		1		0.000766		
971		уки		1		0.000766			1016		чанк		1		0.000766	1061		ь		ро		1		0.000766		
972		улам		1		0.000766			1017		чат		1		0.000766	1062		ь		са		1		0.000766		
973		улиц		1		0.000766			1018		чекм		1		0.000766	1063		ь		ше		1		0.000766		
974		улся		1		0.000766			1019		чера		1		0.000766	1064		ькую		1		0.000766				
975		улюл		1		0.000766			1020		черн		1		0.000766	1065		ью		з		1		0.000766		
976		упал		1		0.000766			1021		чины		1		0.000766	1066		ю		в		1		0.000766		
977		упку		1		0.000766			1022		чуба		1		0.000766	1067		ю		ве		1		0.000766		
978		урга		1		0.000766			1023		чудн		1		0.000766	1068		ю		вс		1		0.000766		
979		урен		3		0.002297			1024		шаль		2		0.001531	1069		ю		го		1		0.000766		
980		урет		1		0.000766			1025		шел		2		0.001531	1070		ю		за		2		0.001531		
981		урец		1		0.000766			1026		шелк		1		0.000766	1071		ю		ин		1		0.000766		
982		урча		1		0.000766			1027		шибе		1		0.000766	1072		ю		ка		1		0.000766		
983		ут		з		1		0.000766	1028		шие		1		0.000766	1073		ю		ки		1		0.000766		
984		утан		1		0.000766			1029		шки		1		0.000766	1074		ю		ту		1		0.000766		
985		утар		1		0.000766			1030		шней		1		0.000766	1075		юкал		1		0.000766				
986		утор		4		0.003063			1031		шь		п		1		0.000766	1076		юком		1		0.000766		

1077 юлюк 1 0.000766	1086 я ша 1 0.000766	1087 яйст 1 0.000766
1078 ющие 1 0.000766	1082 я об 1 0.000766	1088 янут 1 0.000766
1079 юю т 1 0.000766	1083 я ск 2 0.001531	1089 ят р 1 0.000766
1080 я в 1 0.000766	1084 я то 1 0.000766	1090 ятал 1 0.000766
1081 я и 1 0.000766	1085 я ту 1 0.000766	1091 ятиш 1 0.000766

Сумма вероятностей: 1.0000

Приложение 7: листинг кода примерной программы для самостоятельной работы

Представленная ниже программа может быть использована студентами для проведения задачи дома или в аудитории в случае, если нет возможности подготовить материалы для семинара 1.

Программа лучше всего работает в формате Jupyter Notebook, полную версию программы можно найти по следующей ссылке: https://github.com/thaiha123/magis_summer_practise

```
import re
import pandas as pd
from collections import Counter
from random import sample
punctuations = ['.', '...', ',', ';', ':', '!', '?', '\n', '-', '-',
'_', '(', ')']

text = open('text.txt', 'r', encoding='utf8').read()
text
def preprocessing(text):
    for p in punctuations:
        text = text.replace(p, '')          #remove punctuation marks and
new lines
    text = text.lower()                    #lower registered all letters
    text = re.sub(r'[""]', '', text)      #remove quotation mark in
english encoding
    text = re.sub(r'[«»]', '', text)      #remove quotation mark in
russian encoding
    text = text.replace(' ', '_')          #replace space with underline
mark
    text = text.replace('__', '_')
    text = text + "E"                     #add end token to the text
(all letter is lower case so E won't be misrepresented)
    return text
clean_text = preprocessing(text)
```

```

clean_text
def ngram_letter_count(n, text):
    ngrams = [text[i:i+n] for i in range(len(text) - n + 1)]
    total = len(ngrams)
    counter = Counter(ngrams)
    # Convert counts to frequency
    frequencies = {ng: count / total for ng, count in counter.items()}
    return frequencies, total

def to_file(frequencies: dict, total: int,
filename='ngram_freq.xlsx'):
    rows = []
    for gram, freq in frequencies.items():
        rows.append({'n_gram': gram, 'frequency': freq,
'total_ngrams': total * freq})
    df = pd.DataFrame(rows)
    df.to_excel(filename, index=False)
    print(f"Saved to {filename}")

ngram = 3
freq, total = ngram_letter_count(ngram, clean_text)
to_file(freq, total)
print("Total n-grams:", total)
print("Sum of frequencies:", sum(freq.values()))

def create_BOW(frequencies, total):
    bow = []
    for f in frequencies.items():
        for i in range(0, int(f[1] * total)):
            bow.append(f[0])
    return bow

# Creating a list that simulate BOW of ngrams extracted from processed
text
bow = create_BOW(freq, total)
# Randomly choose a ngram from above created list, this action
simulate random selection from a BOW
# Selected ngram should be stringed together to form a sentence
print(sample(bow, 1))

```

Приложение к семинарскому занятию № 3

Листинг кода примерной программы использована в семинаре «Статистические меры: частота термина и обратная частота документа».

```
import gensim
from gensim import corpora
import numpy as np
import zlib

#insert required texts in below variables
text1 = ""
text2 = ""
text3 = ""
text4 = ""
text5 = ""

#not every punctuation but enough for this exercise
punctuations = ['.', ',', ';', ':', '!', '?']

def preprocessing(text):
    for p in punctuations:
        text = text.replace(p, '')
    text = text.lower()
    wlist = text.split(" ")
    return wlist

def compute_tf(corpus, dict):
    tf = {}
    for doc in corpus:
        doc_len = sum(count for _, count in doc)
        for word_id, count in doc:
            tf[dict[word_id]] = count/doc_len
    return tf

def compute_idf(corpus, dict):
    idf = {}
    num_docs = len(corpus)
    for doc in corpus:
```

```

        for word_id, _ in doc:
            if word_id not in idf:
                idf[word_id] = 0
            idf[word_id] += 1
    idf = {dict[word_id]: np.log(num_docs / count) for word_id, count in
idf.items()}
    return idf

def comput_tfidf(corpus, dict):
    tfidf = {}
    tfidf_model = gensim.models.TfidfModel(corpus, id2word=dict)
    for bow in corpus:
        for word_id, value in tfidf_model[bow]:
            tfidf[dict[word_id]] = value
    return tfidf

def NCD(str1, str2):
    s1 = zlib.compress(str1.encode())
    s2 = zlib.compress(str2.encode())
    s3 = zlib.compress((str1 + str2).encode())
    res2 = (s3.__sizeof__() - min(s1.__sizeof__(),
s2.__sizeof__()))/max(s1.__sizeof__(), s2.__sizeof__())
    print("by size: %.6f" % res2)

#preprocessing and merge texts
texts = []
for text in [text1, text2, text3, text4, text5]:
    processed_text = preprocessing(text)
    texts.append(processed_text)

#create dictionary and corpus (bag of word)
dict = corpora.Dictionary(texts)
corpus = [dict.doc2bow(text) for text in texts]

#calculate tf, idf and tf_idf
tf = compute_tf(corpus, dict)
idf = compute_idf(corpus, dict)
tfidf = comput_tfidf(corpus, dict)

```

```

#get words low tfidf score, for example: 0.1
low_score_words = []
for k, v in tfidf.items():
    if v < 0.1:
        low_score_words.append(k)

#sort by increasing order
tf = sorted(tf.items(), key=lambda item: item[1])
idf = sorted(idf.items(), key=lambda item: item[1])
tfidf = sorted(tfidf.items(), key=lambda item: item[1])

#calculate NCD
#without filtering low tfidf score words
new_texts = []
for text in texts:
    new_texts.append(' '.join(text))
NCD(new_texts[0], new_texts[1])
NCD(new_texts[1], new_texts[2])
NCD(new_texts[2], new_texts[3])
NCD(new_texts[3], new_texts[4])
#filter out low tfidf score words and calculate NCD again
filtered_texts = []
for text in new_texts:
    filtered_text = text
    for word in low_score_words:
        filtered_text = filtered_text.replace(word, '')
    filtered_texts.append(filtered_text)
#new NCD values
NCD(filtered_texts[0], filtered_texts[1])
NCD(filtered_texts[1], filtered_texts[2])
NCD(filtered_texts[2], filtered_texts[3])
NCD(filtered_texts[3], filtered_texts[4])
#printout tf, idf, tf-idf values
print(tf)
print(idf)
print(tfidf)

```

Приложение к семинарскому занятию № 4

Листинг кода примерной программы использована в семинаре «Передача информации в биосистемах: теория информации Шеннона».

Программа лучше всего работает в формате Jupyter Notebook, полную версию программы можно найти по следующей ссылке: https://github.com/thaiha123/magis_summer_practise

```
import zlib
from gensim import corpora
import math
from collections import Counter

#text and punctuation
text1 = "The sun shines. Birds sing in the forest. Trees whisper
softly. The sun warms the leaves. Forest animals play. The sun smiles
again, again."
text2 = "The moon rises. The forest sleeps. Trees stand quiet. The
moon watches. Stars glow above. The moon stays quiet while the forest
dreams peacefully."
text3 = "dog кот солнце light дерево cat свет moon дерево кот light
земля dog море небо свет cat вода дерево солнце, вода glow stays свет"
text4 = "1, 8, 3, 7, 19, 6, 7, 27, 9, 682, 7, 3970, 2, 7, 3, 10, 1,
19, 7, 20, 21, 26, 300, 1000"
punctuations = ['.', ',', ';', ':', '!', '?']

#preprocessing text
def preprocessing(text):
    for p in punctuations:
        text = text.replace(p, '')
    text = text.lower()
    wlist = text.split(" ")
    return wlist

#compute term frequency for each word in a text
def compute_tf(corpus, dict):
    tf = {}
```

```

    for doc in corpus:
        doc_len = sum(count for _, count in doc)
        for word_id, count in doc:
            tf[dict[word_id]] = count/doc_len
    return tf

#compute entropy of each text
def compute_entropy(tf):
    entropy = 0
    values = list(tf.values())
    for value in values:
        entropy += -value*math.log2(value)
    return entropy

#compute normalized compressed distance between 2 texts. This value
represent kolmogorov complexity of a text in relation to the other
def NCD(str1, str2):
    s1 = zlib.compress(str1.encode())
    s2 = zlib.compress(str2.encode())
    s3 = zlib.compress((str1 + str2).encode())
    res2 = (s3.__sizeof__() - min(s1.__sizeof__(),
s2.__sizeof__()))/max(s1.__sizeof__(), s2.__sizeof__())
    return res2

#compute volume of recived information from 1 source by reading from
the other source
#for the sake of simplicity, use positional pairing relation, every
texts therefor has the same length
def compute_transinfo(text1, text2):
    text1 = preprocessing(text1)
    text2 = preprocessing(text2)
    if len(text1) != len(text2):
        raise ValueError("texts must have the same number of words for
alignment")
    total = len(text1)
    joint_pairs = list(zip(text1, text2))
    joint_counts = Counter(joint_pairs)
    count1 = Counter(text1)

```

```

count2 = Counter(text2)
res = 0
for (x, y), joint_count in joint_counts.items():
    p_xy = joint_count / total
    p_x = count1[x] / total
    p_y = count2[y] / total
    res += p_xy * math.log2(p_xy / (p_x * p_y))
return res

#get term frequency for every words
tf = []
for text in [text1, text2, text3, text4]:
    text = [preprocessing(text)]
    dict = corpora.Dictionary(text)
    corpus = [dict.doc2bow(text) for text in text]
    tf.append(compute_tf(corpus, dict))

#compute entropy for each texts
for tfx in tf:
    print(tfx)
    print(compute_entropy(tfx))

#number of symbols in each text
len(text1), len(text2), len(text3), len(text4)

#calculate entropy and NCD of texts pair
print(compute_transinfo(text1, text2))
print(compute_transinfo(text1, text3))
print(compute_transinfo(text1, text4))
print(compute_transinfo(text3, text4))

print(NCD(text1, text2))
print(NCD(text1, text3))
print(NCD(text1, text4))
print(NCD(text3, text4))

#split text into BOWs and calculate their entropy
def bow_entropy(text, bow_size):

```

```

bows = []
if bow_size*2 >= len(text) or len(text)%bow_size != 0:
    return "bow size error"
else:
    for i in range(0, len(text), bow_size):
        bows.append([text[i:i+bow_size]])
    res = []
    for bow in bows:
        dic = corpora.Dictionary(bow)
        corpus = [dic.doc2bow(bow) for bow in bow]
        tf = compute_tf(corpus, dic)
        res.append(compute_entropy(tf))
    return res

#print entropy of splited BOWs
print(bow_entropy(preprocessing(text1), 6))
print(bow_entropy(preprocessing(text2), 8))
print(bow_entropy(preprocessing(text3), 3))
print(bow_entropy(preprocessing(text4), 4))

```

Смирнова Елена Валентиновна

Ле Куанг Ньюе

Нгуен Тхай Ха

**МЕТОДИКА ПРОВЕДЕНИЯ
СЕМИНАРСКИХ ЗАНЯТИЙ ПО ДИСЦИПЛИНЕ
«АЛГОРИТМИЧЕСКАЯ ТЕОРИЯ ИНФОРМАЦИИ В
БИОМЕДИЦИНСКИХ СИСТЕМАХ»:
УЧЕБНО-МЕТОДИЧЕСКОЕ ПОСОБИЕ**

© 2025 МГТУ имени Н.Э. Баумана