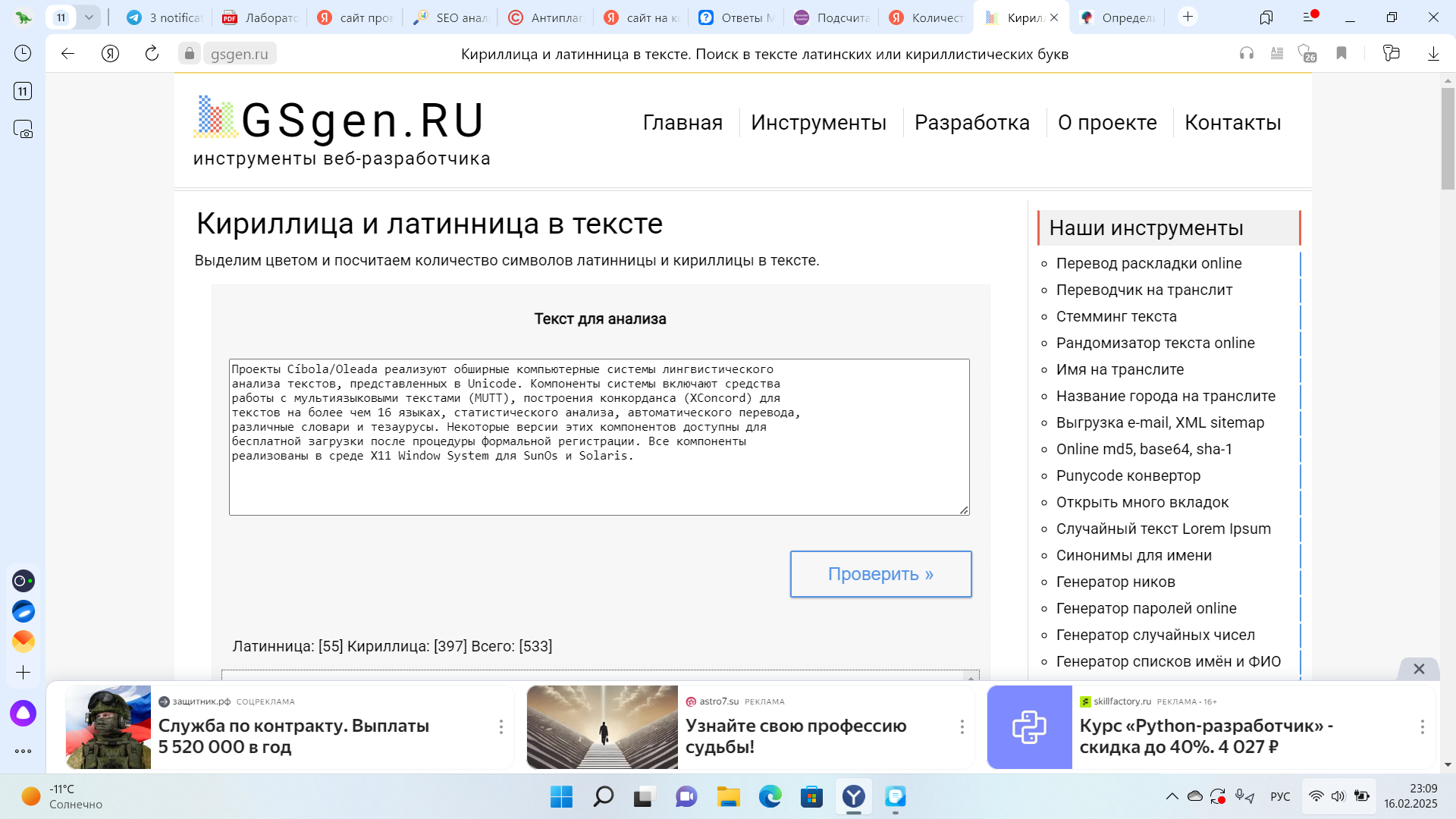
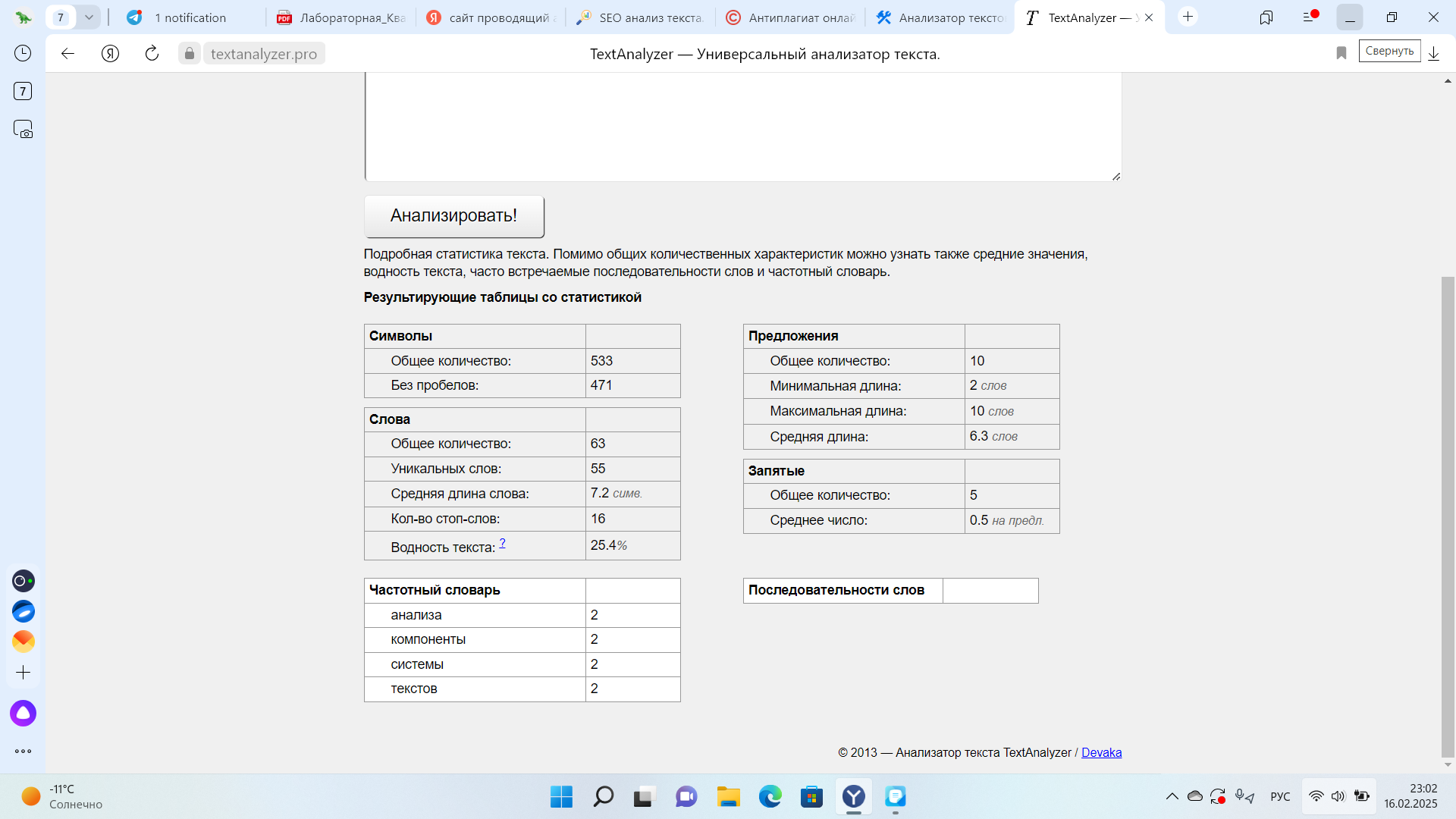
Лабораторная работа номер 1

Уровни 1, 2, 3

1.Для того чтобы провести анализ статистических характеристик текста, я воспользовалась сайтами <https://textanalyzer.pro/> и [https://gsgen.ru (скриншоты](https://gsgen.ru(скриншоты) прикрепляю ниже) . Результаты анализа занесла в таблицу:

|  |  |
| --- | --- |
| Общее количество слов | 63 |
| Общее количество символов (без учета пробелов) | 533 |
| Общее количество символов (с учетом пробелов) | 471 |
| Количество символов, написанных в латинской графике | 55 |
| Общее количество чисел | 2 |
| Средняя длина слов | 7.2 |



2.Средняя длина слов в данном тексте соответствует русскому языку. Для того, чтобы сделать такой вывод среднюю длину слова в латинской графике, я количество символов(их 55) поделила на количество слов – 8, и получилось 6,875. Тоже самое проделала со словами, написанными на русском языке: общее количество символов составляет 473 – 55(латиница) – 5(скобки и /)=413, количество слов 63-8=55, тогда средняя длина русского слова равна 7,5. Для анализа данного текста могут быть существенные такие параметры как, ключевые слова, количество самостоятельных частей речи.

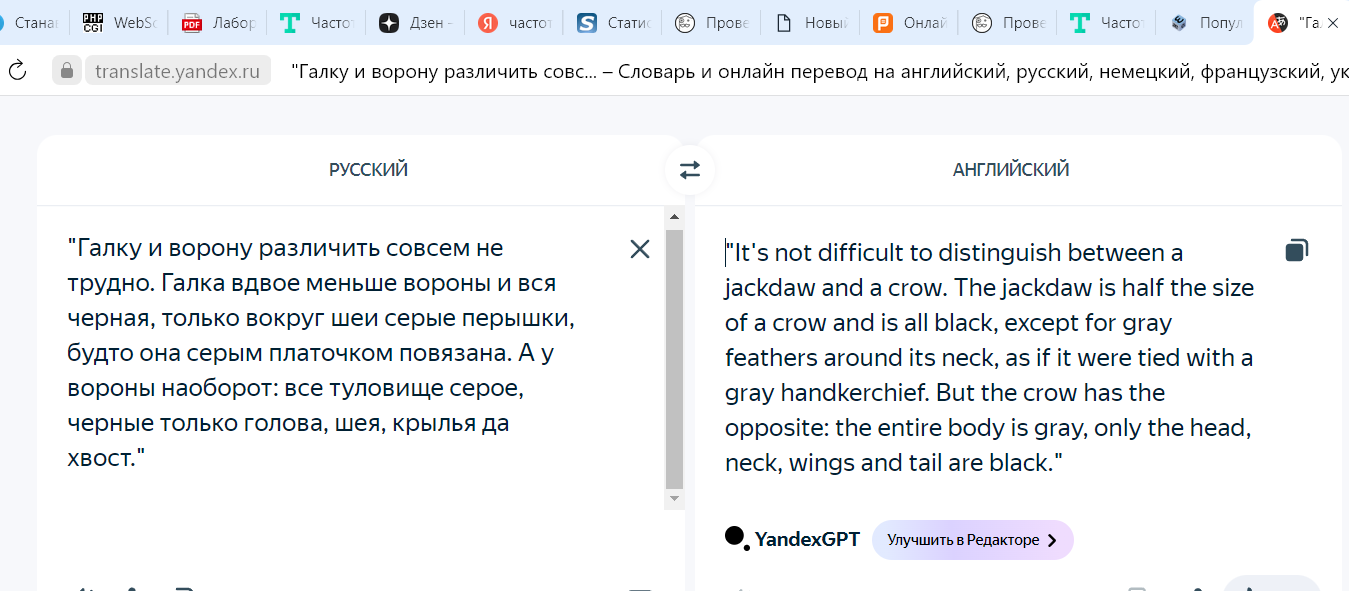
Уровни 1, 2, 3

"Галку и ворону различить совсем не трудно. Галка вдвое меньше вороны и вся черная, только вокруг шеи серые перышки, будто она серым платочком повязана. А у вороны наоборот: все туловище серое, черные только голова, шея, крылья да хвост."

* Для того чтобы составить частотный словарь, я использовала сайт <http://webscript.ru/cgi-bin/text/text.cgi> .



* Чтобы перевести данный текст на английский язык воспользовалась яндекс переводчиком, и по переведенному тексту сделала частотный словарь.





* Я нашла абсолютную частоту слов в тексте, посчитав сколько раз они встречаются в тексте, и относительную частоту, поделив значение частоты данного слова на общее число всех слов в тексте и умножить на 100%. Результаты занесла в таблицу:

|  |  |  |
| --- | --- | --- |
| слова | абсолютная частота | относительная частота |
| вороны | 2 | 4.26 |
| только | 2 | 4.26 |
| хвост | 1 | 3.125 |
| шея | 1 | 3.125 |
| совсем | 1 | 3.125 |
| различить | 1 | 3.125 |
| повязана | 1 | 3.125 |
| перышки | 1 | 3.125 |
| вокруг | 1 | 3.125 |
| наоборот | 1 | 3.125 |
| шеи | 1 | 3.125 |
| все | 1 | 3.125 |
| черные | 1 | 3.125 |
| серые | 1 | 3.125 |
| голова | 1 |  |
| серое | 1 | 3.125 |
| она | 1 | 3.125 |
| вдвое | 1 | 3.125 |
| меньше | 1 | 3.125 |
| платочком | 1 | 3.125 |
| серым | 1 | 3.125 |
| трудно | 1 | 3.125 |
| ворону | 1 | 3.125 |
| Галку | 1 | 3.125 |
| будто | 1 | 3.125 |
| туловище | 1 | 3.125 |
| Галка | 1 | 3.125 |
| крылья | 1 | 3.125 |
| вся | 1 | 3.125 |
| черная | 1 | 3.125 |

|  |  |  |
| --- | --- | --- |
| слова | абсолютная частота | относительная частота |
| the | 5 | 10.64 |
| crow | 3 | 6.38 |
| gray | 3 | 6.38 |
| and | 3 | 6.38 |
| neck | 2 | 4.26 |
| black | 2 | 4.26 |
| jackdaw | 2 | 4.26 |
| around | 1 | 2.13 |
| wings | 1 | 2.13 |
| has | 1 | 2.13 |
| feather | 1 | 2.13 |
| not | 1 | 2.13 |
| entire | 1 | 2.13 |
| tail | 1 | 2.13 |
| its | 1 | 2.13 |
| head | 1 | 2.13 |
| only | 1 | 2.13 |
| for | 1 | 2.13 |
| opposite | 1 | 2.13 |
| with | 1 | 2.13 |
| The | 1 | 2.13 |
| half | 1 | 2.13 |
| tied | 1 | 2.13 |
| handkerchief | 1 | 2.13 |
| are | 1 | 2.13 |
| all | 1 | 2.13 |
| size | 1 | 2.13 |
| body | 1 | 2.13 |
| But | 1 | 2.13 |
| expert | 1 | 2.13 |
| distinguish | 1 | 2.13 |
| between | 1 | 2.13 |
| difficult | 1 | 2.13 |
| were | 1 | 2.13 |

* При сравнении частотных списков оригинала и перевода можно сделать несколько наблюдений:  
  Наблюдения по текстам  
  - Оригинал может содержать культурные слова, специфичные для языка, которые могут быть переведены как аналогичные, так и с потерей смысла.  
  - Перевод может использовать более универсальные слова для достижения аналогичного значения, что может изменить его стиль.  
  Общие черты оригинала и перевода  
  - Основные тематические слова, передающие главную идею, как правило, присутствуют в обоих текстах.  
  - Структурные элементы (условные связки, местоимения) могут совпадать по частоте.  
  Отличия в частотных характеристиках  
  - Некоторые слова в оригинале могут встречаться реже, но они играют важную роль в передаче чувства или настроения текста.  
  - Перевод может иметь более равномерное распределение частот, отражая упрощённые выражения.  
  Эти наблюдения помогут глубже понять, как язык формирует мысли и передает культурное значение.

Уровни 2, 3

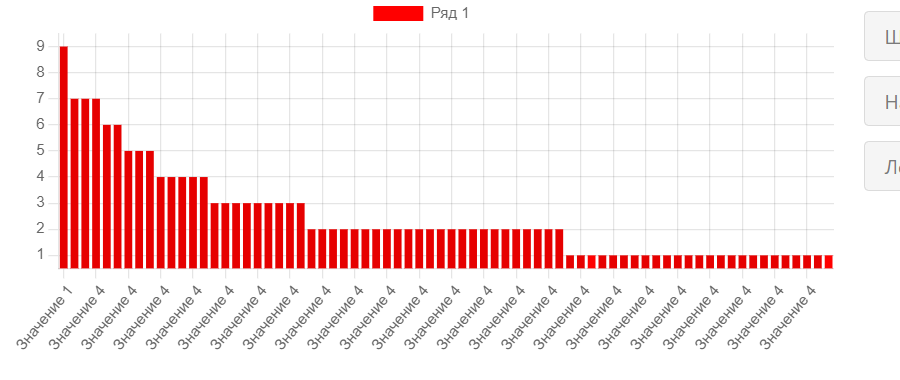
* Для выполнения задания выбрала один текст написанный человеком, а другой нейросетью.

Когда человек сознательно или интуитивно выбирает себе в жизни какую-то цель, жизненную задачу, он невольно дает себе оценку. По тому, ради чего человек живет, можно судить и о его самооценке - низкой или высокой.  
Если человек живет, чтобы приносить людям добро, облегчать их страдания, давать людям радость, то он оценивает себя на уровне этой своей человечности. Он ставит себе цель, достойную человека.  
Только такая цель позволяет человеку прожить свою жизнь с достоинством и получить настоящую радость. Да, радость! Подумайте: если человек ставит себе задачей увеличивать в жизни добро, приносить людям счастье, какие неудачи могут его постигнуть? Не тому помочь? Но много ли людей не нуждаются в помощи?  
Если жить только для себя, своими мелкими заботами о собственном благополучии, то от прожитого не останется и следа. Если же жить для других, то другие сберегут то, чему служил, чему отдавал силы.  
Можно по-разному определять цель своего существования, но цель должна быть. Надо иметь и принципы в жизни. Одно правило в жизни должно быть у каждого человека, в его цели жизни, в его принципах жизни, в его поведении: надо прожить жизнь с достоинством, чтобы не стыдно было вспоминать.  
Достоинство требует доброты, великодушия, умения не быть эгоистом, быть правдивым, хорошим другом, находить радость в помощи другим.  
Ради достоинства жизни надо уметь отказываться от мелких удовольствий и немалых тоже… Уметь извиняться, признавать перед другими ошибку - лучше, чем врать.  
Обманывая, человек прежде всего обманывает самого себя, ибо он думает, что успешно соврал, а люди поняли и из деликатности промолчали.  
Жизнь - прежде всего творчество, но это не значит, что каждый человек, чтобы жить, должен родиться художником, балериной или ученым. Можно творить просто добрую атмосферу вокруг себя. Человек может принести с собой атмосферу подозрительности, какого-то тягостного молчания, а может внести сразу радость, свет. Вот это и есть творчество.  
 (Человек)

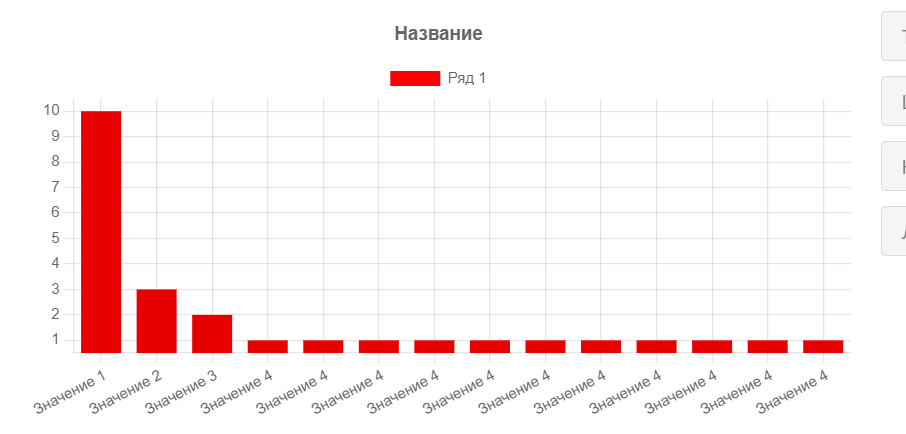
В последние десятилетия экология стала одной из самых актуальных тем на мировой арене. Изменение климата, загрязнение окружающей среды и утрата биологического разнообразия — все эти проблемы требуют незамедлительных действий. Люди начинают осознавать важность защиты природы и стремятся внести свой вклад в сохранение планеты. Различные движения и организации по всему миру призывают к сокращению использования пластика, переходу на возобновляемые источники энергии и охране редких видов животных. Индивидуальные усилия, такие как сортировка отходов, использование общественного транспорта и поддержка устойчивого сельского хозяйства, могут существенно повлиять на ситуацию. Каждый шаг имеет значение, и вместе мы можем создать более чистое и безопасное будущее для нас и будущих поколений. Важно помнить, что забота о планете — это не только обязанность, но и привилегия, позволяющая нам наслаждаться ее красотой и щедротами.(Нейросеть)

Проанализировала частотное распределение слов в обоих текстах сс использованием закона Ципфа. Для этого воспользовалась ресурсами, такими как PR-CY.

Для того чтобы построить гистограммы, я нашла частоту слов с помощью сайта <https://sanstv.ru/> и построила гистограммы с помощью <https://anytools.pro/> .



(текст создан человеком)



(текст создан нейросетью)

Если сравнивать графики частоты слов в тексте с распределением по закону Цимфа, можно обнаружить отклонения на концах графика и гистограммы.

В квантативной лингвистике, помимо закона Цимфа, существуют и другие законы, такие как закон Мартина, закон Менцерата-Альтмана. Данную информацию нашла на сайтах: <https://present5.com/lekciya-6-kvantitativnaya-lingvistika-soderzhanie-1-problematika-kvantitativnoj/> , <https://ru.ruwiki.ru/wiki/Квантитативная_лингвистика> .

Закон Мартина: Этот закон касается лексических цепочек, которые образуются при поиске дефиниций (определений) слов в словаре, затем дефиниций только что найденных дефиниций и т. д. В конце концов все эти дефиниции образуют иерархию всё более и более общих значений, при этом дефиниций становится тем меньше, чем более общее значение получается. Среди уровней данной иерархии существует целый ряд подчиняющихся закону отношений.

[Закон Менцерата](https://ru.ruwiki.ru/w/index.php?title=%D0%97%D0%B0%D0%BA%D0%BE%D0%BD_%D0%9C%D0%B5%D0%BD%D1%86%D0%B5%D1%80%D0%B0%D1%82%D0%B0&action=edit&redlink=1) (также, в особенности в лингвистике, известный как закон Менцерата-Альтмана): Данный закон гласит, что размеры составляющих конструкции уменьшаются с увеличением самой изучаемой конструкции. Чем длиннее, например, предложение (измеренное количеством входящих в его состав придаточных предложений), тем короче входящие в его состав придаточные предложения (измеренные количеством слов), или: чем длиннее слово (в слогах или морфах), тем короче слоги или их звучание.

Применяя закон Мартина к данному тексту, можно сделать выводы о его лексическом составе и акцентах. В тексте написанном человеком можно заметить, частоту ключевых слов(цель жизни, ценности).Это может указывать на то, что автор целенаправленно подчеркивает важные аспекты человеческого существования и отношений. Таким образом, анализируя частотное распределение слов, можно глубже понять, как автор передает свои идеи и какие темы являются центральными в данном тексте. В тексте написанном нейросетью можно заметить, что наиболее часто встречаются ключевые слова, связанные с экологией. Это помогает проследить главную мысль текста.

Таким образом, закон Мартина помогает анализировать текст, но не может служить методом для определения, был ли текст написан нейросетью или человеком.  
уровень 2,3

Я ознакомилась с основными принципами работы сервиса RusVectōrēs. Он имеет следующие инструменты:

1. Поиск похожих слов. Данный инструмент помогает найти синонимы. Например, если вывести слово "животное", может вывести слово "питомец".
2. Визуализация семантических связей между словами. Данный инструмент может  построить карту взаимного расположения введенных слов выбранной модели/моделях и отобразив двумерную проекцию этой карты. Например, модели FastText могут использовать для анализа слов на уровне морфем, что позволяет учитывать различные формы слов. Это особенно полезно для русского языка с его богатой морфологией, где слово может иметь множество форм.
3. Вычисление семантической близости. Данный инструмент помогает найти уровень близости пары слов. Это помогает выполнять ряд задач в компьютерной лингвистике: **Создание и функционирование информационно-поисковых систем**. Алгоритмы на основе семантической близости помогают поисковикам понимать, насколько близко связаны друг с другом слова и фразы в контенте, и делать выдачу точнее.   
   **• Создание вопросно-ответных систем**. Метрики семантической близости позволяют определять семантические отношения между словами и создавать инструменты автоматического понимания текстов.   
   **• Автоматическая обработка текстов**. Например, разрешение многозначности фразеологизмов, встреченных в тексте.
4. [Решать аналогии](https://rusvectores.org/ru/calculator) вида «найти слово **X**, которое так относится к слову **Y**, как слово **A** относится к слову **B**». Например, пицца – Италия, тогда Россия – это пельмени.
5. [Генерировать контекстно-зависимые лексические подстановки](https://rusvectores.org/ru/contextual) для контекстуализированных дистрибутивных моделей. Для этого вводится фраза или предложение(5-15 слов), затем генерируются контекстуализированные вектора для слов в нашем запросе. Затем для каждого такого вектора находят наиболее похожие слова среди 10 тысяч самых частотных слов в словаре этой же модели.



Выводы о применимости и эффективности инструментов дистрибутивной семантики, предлагаемых RusVectōrēs, следующие:

1. **Адаптация к специфике русского языка**: Инструменты RusVectōrēs учитывают морфологическую сложность и разнообразие русского языка, что позволяет более точно анализировать семантические связи. Это делает ресурс особенно полезным для лингвистов и специалистов в области обработки естественного языка.
2. **Широкий функционал**: Сервис предлагает разнообразные инструменты для поиска семантических отношений, анализа векторных моделей и кластеризации слов. Это позволяет пользователям проводить комплексный анализ текстов, выявлять синонимы, антонимы и другие связи между словами, а также исследовать контексты их употребления.
3. **Удобство использования**: Интуитивно понятный интерфейс и доступность различных инструментов обеспечивают легкость в навигации и использовании платформы. Это делает RusVectōrēs доступным не только для профессионалов, но и для студентов и исследователей, которые могут быстро начать работу с ресурсом.
4. **Поддержка научных исследований**: Инструменты RusVectōrēs могут быть успешно использованы в научных исследованиях, связанных с лексикографией, семантикой и синтаксисом, а также в проектах по машинному обучению и анализу текстов. Это делает сервис ценным ресурсом для академического сообщества.
5. **Интеграция с другими технологиями**: Возможность интеграции инструментов с другими библиотеками и системами обработки данных расширяет горизонты их применения, позволяя разработчикам создавать более сложные решения на основе полученных данных.

В целом, инструменты дистрибутивной семантики, предлагаемые RusVectōrēs, являются эффективными и применимыми в различных областях, от научных исследований до практического использования в NLP, что делает этот ресурс важным вкладом в развитие лингвистики и обработки естественного языка в русскоязычном контексте.