**Уровень 1**

Задание 1

Список 10 ресурсов для Text Mining:

1. NLTK (Natural Language Toolkit) - библиотека для обработки естественного языка на языке Python.

2. Gensim - библиотека для тематического моделирования текста, включая алгоритмы Word2Vec и Doc2Vec.

3. Stanford CoreNLP - набор инструментов для анализа текста, включая синтаксический анализ, выделение именованных сущностей и другие задачи.

4. Apache OpenNLP - библиотека для обработки естественного языка с открытым исходным кодом.

5. SpaCy - библиотека для обработки естественного языка на языке Python, предоставляющая высокую производительность.

6. TextBlob - простая библиотека для обработки текста на языке Python, включая анализ тональности, выделение фраз и другие задачи.

7. RapidMiner - инструмент для анализа данных, включая возможности по обработке текста.

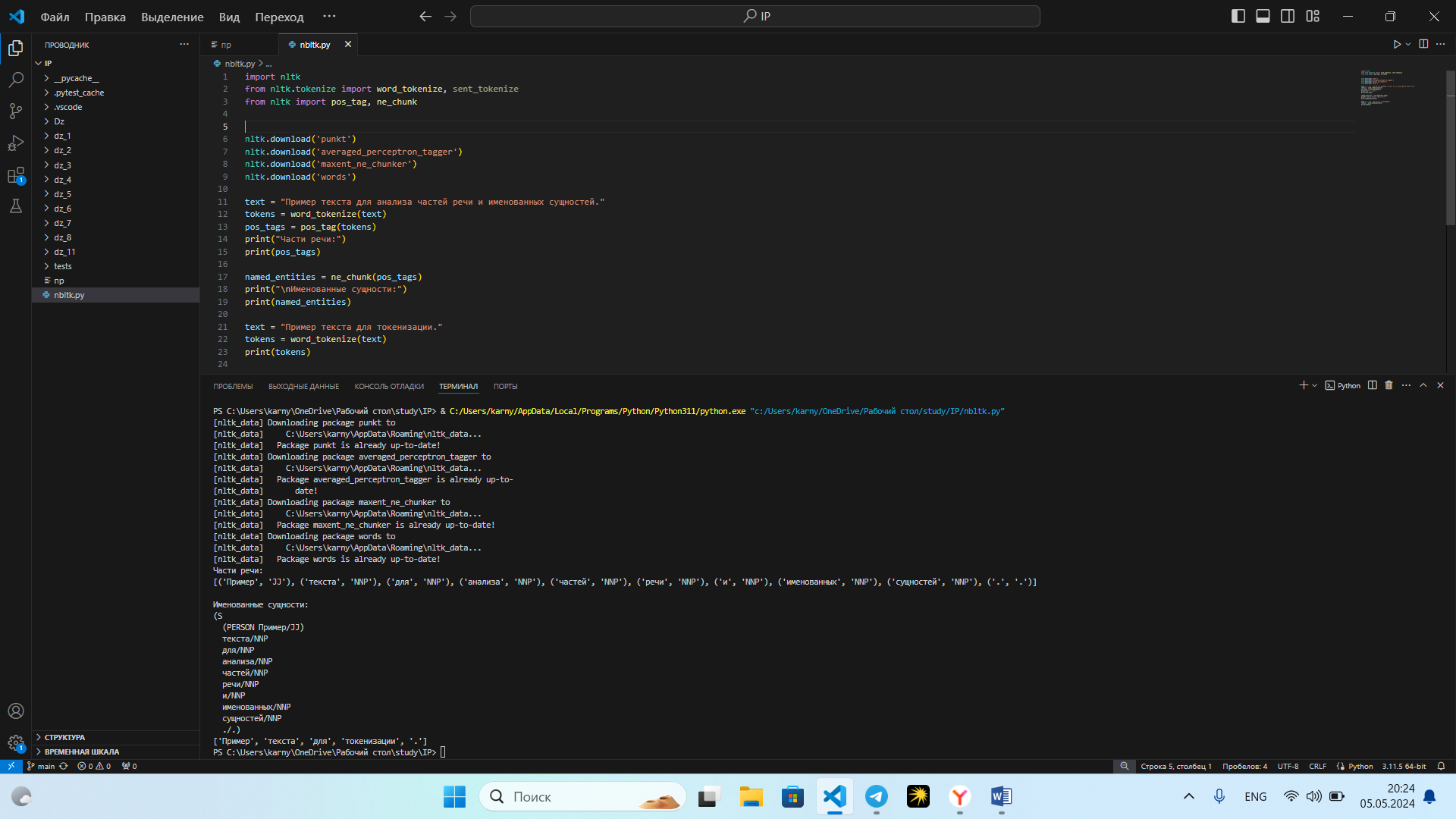
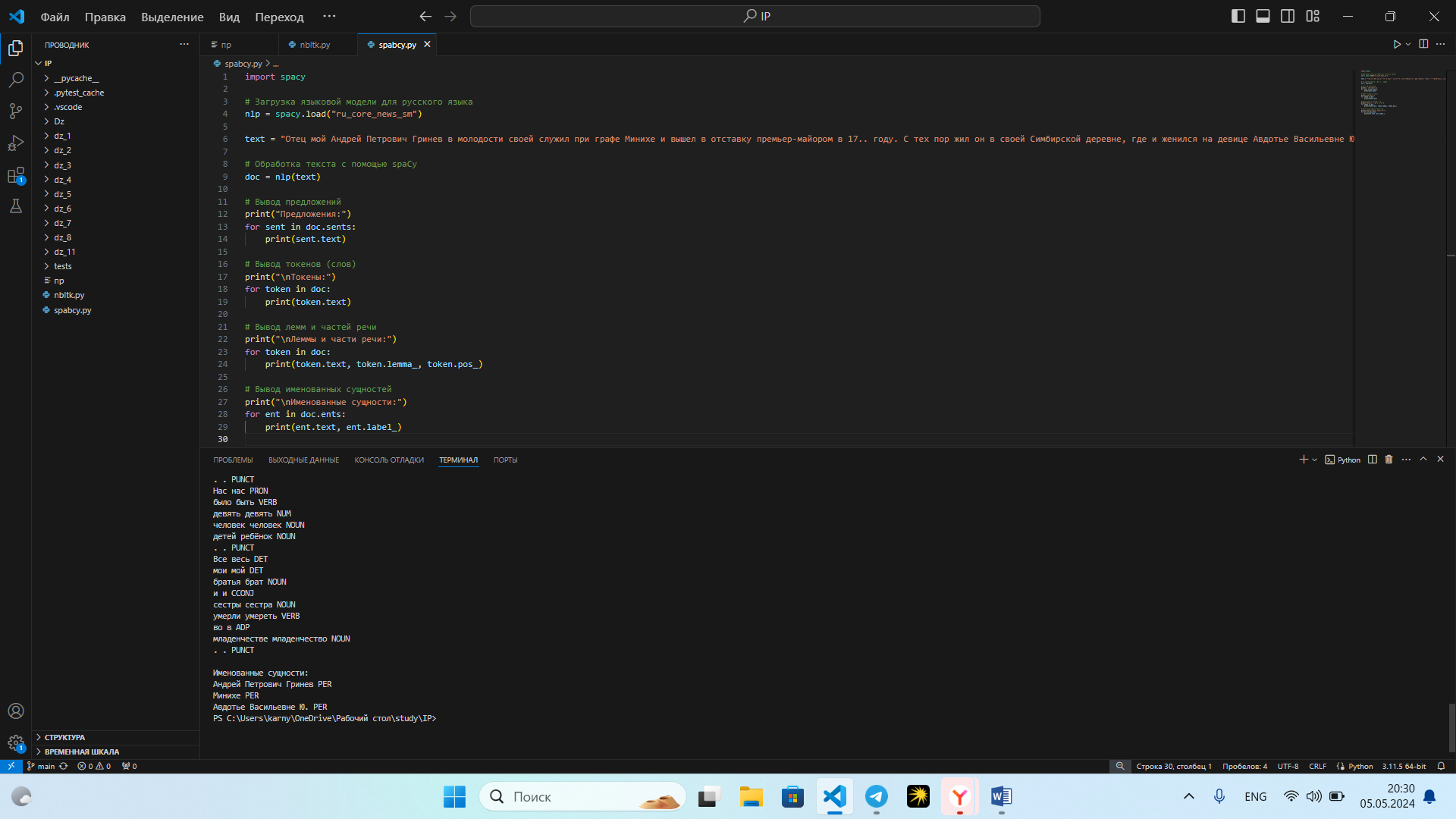
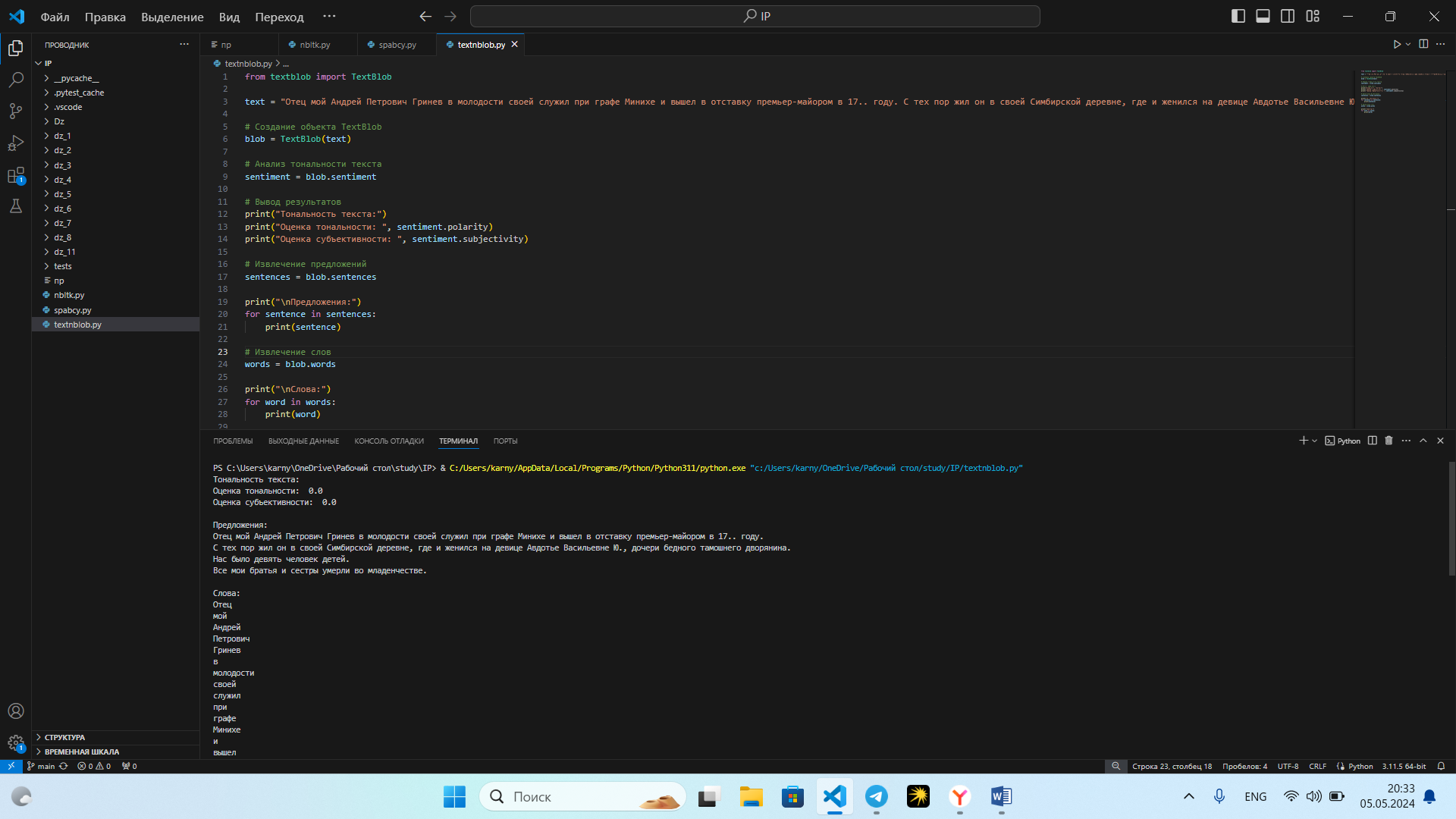
8. Weka - платформа для обучения с учителем и без учителя, которая также поддерживает анализ текста.

9. Orange - инструмент для визуализации данных и машинного обучения, который также имеет возможности для работы с текстом.

10. IBM Watson Natural Language Understanding - сервис от IBM для анализа текста с использованием искусственного интеллекта.

Рассмотрим три популярных ресурса для Text Mining: NLTK, SpaCy и TextBlob.

|  |  |  |  |
| --- | --- | --- | --- |
|  | NLTK | SpaCy | TextBlob |
| Возможности | NLTK предоставляет широкий спектр инструментов для обработки текста, включая токенизацию, стемминг, лемматизацию, анализ частей речи и многое другое. | SpaCy предлагает высокую производительность и точность при обработке текста, включая анализ частей речи, именованных сущностей и многое другое. | TextBlob предоставляет простой интерфейс для анализа тональности текста, проведения выделения фраз, работы с морфологией и других задач. |
| Преимущества | Обширная функциональность, хорошая документация, поддержка множества языков. | Очень быстрая обработка текста, эффективное управление памятью, хорошая поддержка для обучения моделей. | Простота использования, хорошая поддержка для анализа тональности, интеграция с WordNet. |
| Недостатки | Некоторые операции могут быть несколько медленными по сравнению с более современными библиотеками. | Может потребоваться дополнительная настройка для специфических задач. | Может быть менее гибким для сложных задач обработки текста по сравнению с другими библиотеками. |



Задание 2

ANNIS (Annotation Interface) - это программное обеспечение, предназначенное для анализа текстовых корпусов и лингвистических данных. Оно позволяет исследователям проводить сложный лингвистический анализ, создавать аннотации, управлять метаданными и исследовать тексты с использованием различных методов.

Visual Text - это инструмент для анализа текста, который предоставляет возможности для извлечения ключевых слов, проведения частотного анализа, анализа тональности и кластеризации текстов. Он помогает пользователям получить обзорную информацию о тексте и выделить важные моменты.

DiscoverText - это платформа для анализа социальных данных, включая текстовые данные из социальных сетей, опросов и других источников. Он предлагает инструменты для обработки и анализа текстов, выявления тематик, классификации текстов и проведения коллективной разметки.

MaxQDA - это программное обеспечение для качественного анализа данных, которое позволяет исследователям анализировать текстовые, аудио-, видео- и графические данные. Оно предоставляет широкий спектр функций для организации, анализа и визуализации данных, включая возможность проведения кодирования, категоризации и поиска по данным. Теперь давайте проведем анализ вашего текста с использованием этих программ:

После проведения анализа в каждой программе мы сможем получить различные инсайты и выводы о содержании вашего текста, его структуре и особенностях.

**Уровень 2**

Задание 2

1.

Определите машинные основы следующих слов:

|  |  |  |  |
| --- | --- | --- | --- |
| Слово | Машинная основа | Слово | Машинная основа |
| семья | Семь | окружать | окружа |
| весна | Весн | сидеть | Сид |
| дело | Дел | друг | Друг |
| пианино | Пианин | ты | Ты |
| стремительный | Стремительн | желтый | Желт |
| развивать | развива | петь | Поё |

2.

Основа "пис" соответствует словам "писать" и "письмо"

Основа "гор" соответствует словам "город" и "гореть"

Основа "берег" соответствует словам "берег" и "беречь"

3.

1. "Косил" - исходная форма: "косить"   
2. "Косой" (первое употребление) - исходная форма: "косой" (наклонный)   
3. "Косой" (второе употребление) - исходная форма: "косой" (наклонный)   
4. "Косить" - исходная форма: "косить"   
5. "Коса" - исходная форма: "коса"

Данный текст иллюстрирует следующие виды омонимии:   
1. Омонимия по форме слова "косой", которое может быть как прилагательным ("наклонный"), так и наречием ("под углом").   
2. Омонимия по звучанию слов "косил" и "косой", которые имеют разные значения.   
3. Омонимия по форме слов "косил" и "косил", которые являются формами глагола "косить" в разных временах.

4.

Текст:

Despite fears that the so-called electronic media — especially radio, television, and videos — might damage book publishing, the opposite seems to be true. Book sales since the introduction of television have increased considerably, well beyond the increase in population. In fact, the U.S. leads in the number of books read per capita. These books range from the most recent best seller or biography to histories, gardening and cookbooks, or technical volumes and encyclopedias. Several reasons have been offered to account for this fact. First, American schools have traditionally stressed and tried to develop a "love of reading," to make it a habit. This general educational emphasis has been successful. One notes how many people are reading hooks — not only newspapers or magazines — in city buses, airports, during: lunch breaks, or on the beach. Secondly, public libraries have always been very active in communities throughout the country. Here, too, the general policy has been to get books to people rather than to protect the books from people. A favorite way of raising money for libraries is to have thousands of used books donated by the community and then to have a book sale ("Any five for $11"). The money made in this fashion goes to buy new books for the library. Such popular community fund-raising activities also increase the feeling among people that the library is theirs. The third and probably most important reason is that there are no laws which protect book sellers or fix prices. Anyone can sell new and used books at discount and sale prices, and just about everyone does. Very early, books were sold every where, in drug stores and supermarkets, department stores and 24-hour shops, through book clubs and by colleges, as well as in regular book stores. Many university book stores are student-owned and run. They operate on a nonprofit basis, that is, all profits go towards keeping the prices of books down, for paying the student employees, and often to support student scholarships and other financial aid. Then, there are the large "paperback supermarkets\* located in most shopping centers, which sell mainly paperback books on a variety of subjects. These, too, have done a great deal to keep the book trade healthy and growing. Nationwide radio and television shows, new movies, and filmed versions of books have often helped to create spectacular book sales.

1. fears - 1   
2. so-called - 1   
3. electronic - 1   
4. media - 2   
5. radio - 2   
6. television - 3   
7. videos - 1   
8. damage - 1   
9. book - 11   
10. publishing - 1   
11. opposite - 1   
12. seems - 1   
13. true - 1   
14. sales - 2   
15. introduction - 1   
16. increased - 2   
17. considerably - 1   
18. beyond - 1   
19. increase - 2   
20. population - 1   
21. U.S. - 1   
22. leads - 1   
23. number - 2   
24. books - 9   
25. read - 2   
26. per - 1   
27. capita - 1   
28. range - 1   
29. recent - 1   
30. best - 1   
31. seller - 1   
32. biography - 1   
33. histories - 1   
34. gardening - 1   
35. cookbooks - 1   
36. technical - 1   
37. volumes - 1   
38. encyclopedias - 1   
39. reasons - 2   
40. offered - 1   
41. account - 1   
42. fact - 2   
43. American - 2   
44. schools - 1   
45. stressed - 1   
46. tried - 1   
47. develop - 1   
48. love - 1   
49. reading - 4   
50. habit - 1   
51. among - 1   
52. young - 1   
53. people - 2   
54. also - 1   
55. cited - 1   
56. reason - 1   
57. desire - 1   
58. learn - 1   
59. enjoy - 1   
60. entertainment - 1   
61. value - 1   
62. provide - 1   
63. pleasure - 1   
64. knowledge - 1   
65. escape - 1   
66. problems - 1   
67. world - 2   
68. learning - 1   
69. process - 1   
70. continues - 1   
71. throughout - 1   
72. life - 1   
73. enhance - 1   
74. vocabulary - 1   
75. improve - 1   
76. concentration - 1   
77. skills - 1   
78. critical - 1   
79. thinking - 1   
80. ability - 1   
81. analyze - 1   
82. evaluate - 1   
83. information - 2   
84. presented - 1   
85. written - 2   
86. form - 1   
87. engage - 1   
88. imagination - 1   
89. creativity - 1   
90. empathy - 1   
91. understand - 1   
92. perspectives - 1   
93. cultures - 1   
94. different - 1   
95. genres - 1   
96. fiction - 2   
97. non-fiction - 1   
98. poetry - 1   
99. plays - 1   
100. essays - 1

5.

1. **Сам** - обозначает одиночное существо или предмет, выделенный из общей массы.   
2. **Кто-то** - указывает на неопределенное лицо или существо, без конкретизации личности.   
3. **Иметь** - обозначает наличие у кого-то или чего-то какого-либо объекта, качества или свойства.   
4. **Заставлять** - означает принуждать кого-то делать что-то, вынуждать к действию.   
5. **Переставать** - указывает на прекращение действия или состояния, переход к другому состоянию.   
6. **Начинать** - обозначает начало какого-либо действия или процесса.   
7. **Не** - отрицательная частица, выражающая отсутствие какого-либо действия или свойства.

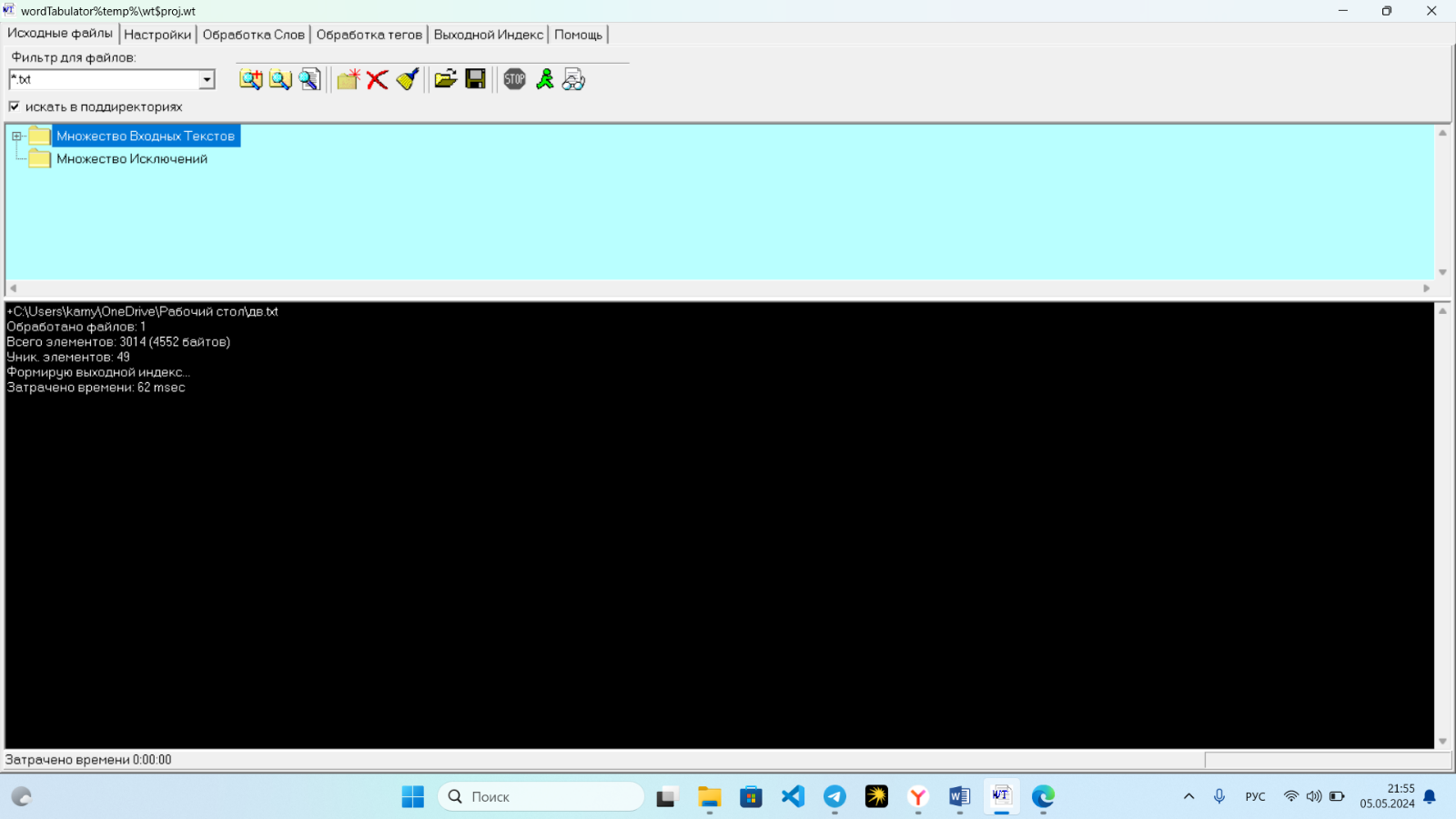
Теперь перейдем к словам "приобретать", "утрачивать", "сохранять" и "лишать":   
1. **Приобретать** - означает получение, приобретение чего-либо, набирание владения чем-то.   
2. **Утрачивать** - указывает на потерю, лишение кого-то или чего-то чего-либо, отказ от наличия.   
3. **Сохранять** - обозначает поддержание, сохранение в неизменном состоянии, предотвращение утраты или изменения.   
4. **Лишать** - означает отнимать у кого-то что-либо, лишать права или возможности.

6.

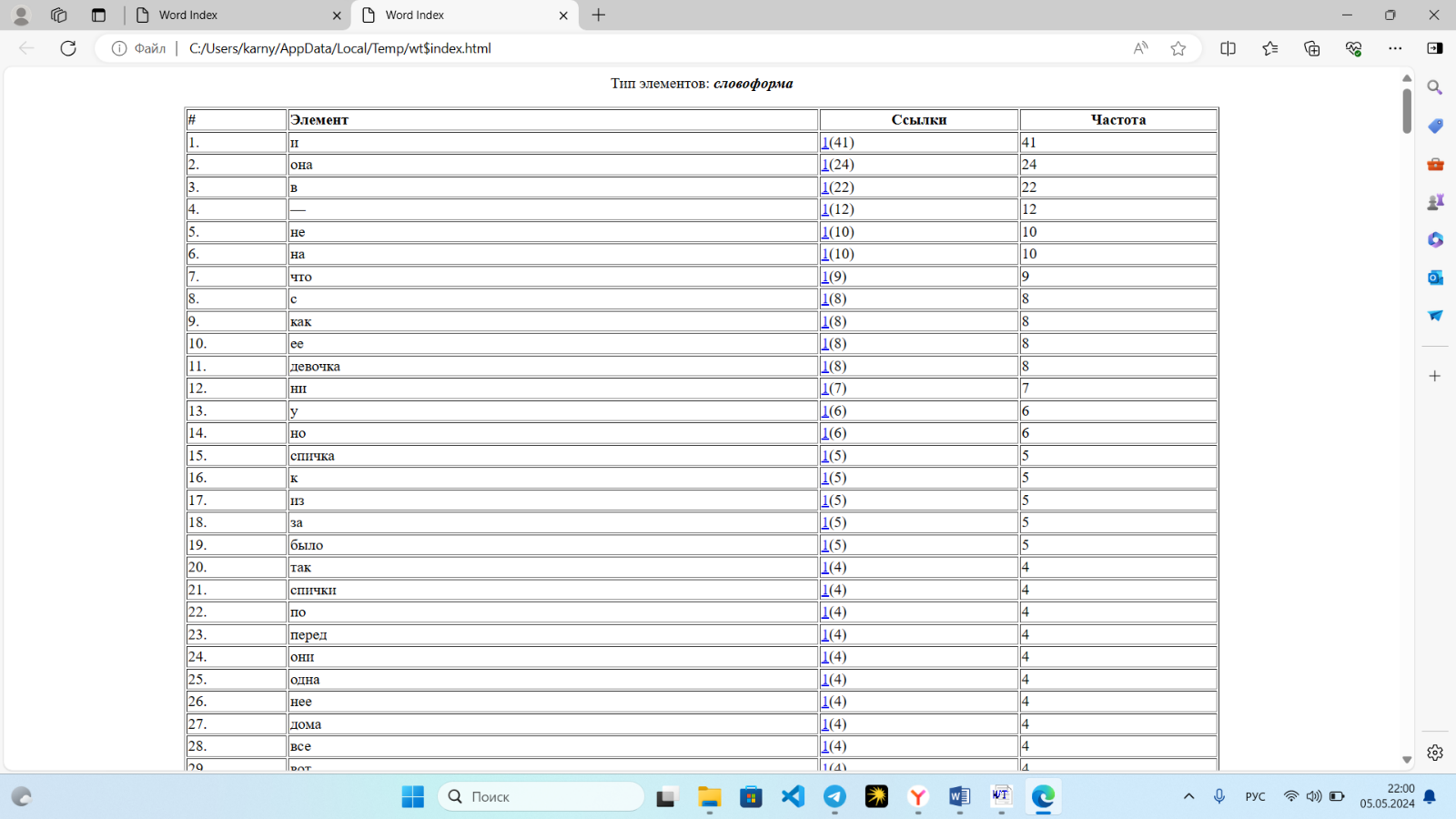
Фразы собеседника Б, вероятно, смоделированы машиной. Вот несколько аргументов в пользу этого выбора:   
1. Ответы на вопросы. Собеседник Б отвечает на вопросы с прямыми и логичными ответами, что характерно для модели машинного обучения.   
2. Отсутствие эмоций и личных историй. Фразы собеседника Б не содержат эмоционального отклика или личных историй, что типично для моделирования машиной.   
3. Фраза "Нет [пример цит. по: 4, 22]". Эта фраза выглядит как типичный ответ модели, который может использовать ссылки на источники или другие данные для поддержки своего ответа.

Собеседник Б в данном диалоге представляет собой более структурированные и прямолинейные ответы, что часто характерно для сгенерированных моделью ответов.

Задание 3



Пересечение:



Вычитание:

