

МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ, СВЯЗИ И МАССОВЫХ
КОММУНИКАЦИЙ РОССИЙСКОЙ ФЕДЕРАЦИИ
Ордена Трудового Красного Знамени федеральное государственное
бюджетное образовательное учреждение высшего образования
**«Московский Технический Университет Связи И Информатики
(MTUCI)»**

Кафедра «Математическая кибернетика и информационные технологии»

Лабораторная Работа 1

по дисциплине

«Машинное обучение»

Выполнил: студент 3 курса гр. БВТ2201
Ньяти Каелиле

Москва 2025 г

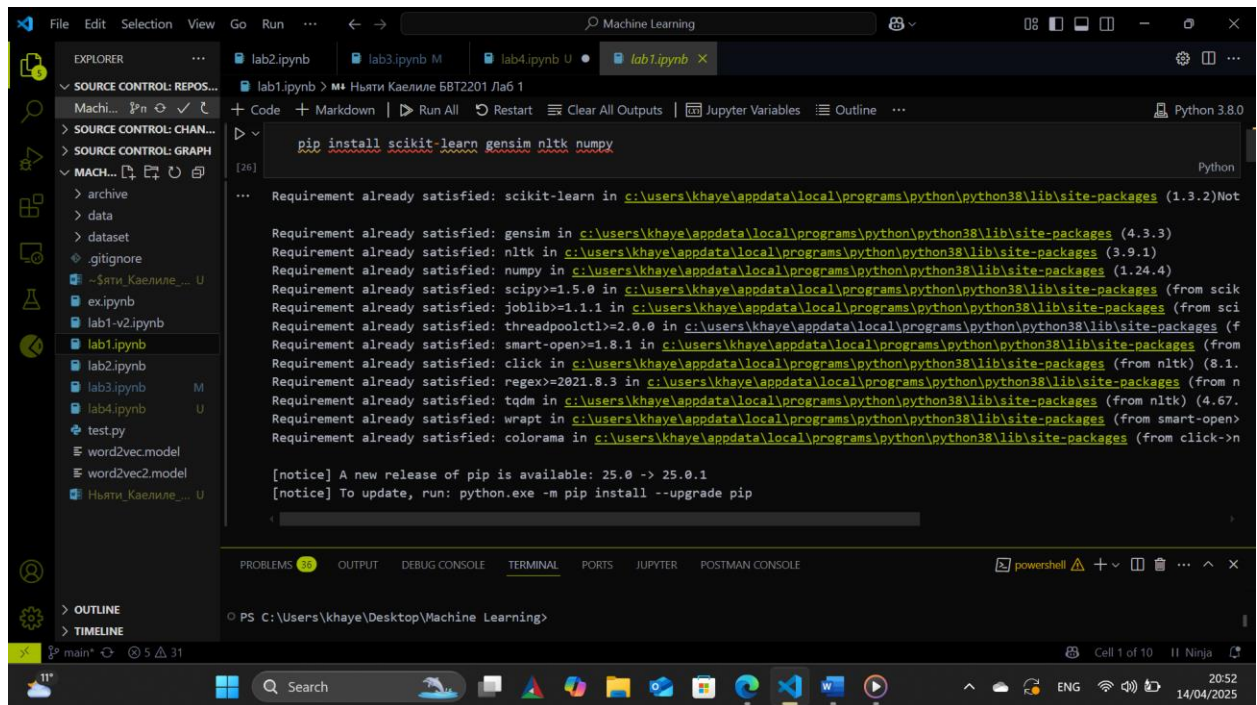
Содержание

1. Задание
2. Ход работы
3. Вывод

1. Задание

Реализовать алгоритмы TF-IDF + word2vec (чтобы можно было складывать и вычитать слова).
Также необходимо найти датасет для обучения написанных алгоритмов)

2. Ход работы



The screenshot shows a Jupyter Notebook interface with a dark theme. The left sidebar contains a file explorer with a tree view showing folders like 'archive', 'data', 'dataset', and 'gignore', and files like 'ex.ipynb', 'lab1-v2.ipynb', 'lab1.ipynb', 'lab2.ipynb', 'lab3.ipynb', 'lab4.ipynb', 'test.py', 'word2vec.model', and 'word2vec2.model'. The main area displays a code cell with the command `pip install scikit-learn gensim nltk numpy`. The output shows that all requirements are already satisfied for the specified versions in the local environment. The bottom status bar indicates the current cell is 1 of 10, the user is 'Ninja', and the date is 14/04/2025.

```
File Edit Selection View Go Run ... Machine Learning
lab1.ipynb lab3.ipynb M lab4.ipynb U lab1.ipynb X
lab1.ipynb > М+ Няти Каелиле ББТ2201 Лаб 1
+ Code + Markdown | ▶ Run All ⌂ Restart ☒ Clear All Outputs | Jupyter Variables ☰ Outline ... Python 3.8.0
[26]
pip install scikit-learn gensim nltk numpy

... Requirement already satisfied: scikit-learn in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (1.3.2)Not

Requirement already satisfied: gensim in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (4.3.3)
Requirement already satisfied: nltk in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (3.9.1)
Requirement already satisfied: numpy in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (1.24.4)
Requirement already satisfied: scipy>=1.5.0 in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (from scik
Requirement already satisfied: joblib>=1.1.1 in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (from sci
Requirement already satisfied: smart-open>=1.8.1 in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (from
Requirement already satisfied: click in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (from nltk) (8.1
Requirement already satisfied: regex>=2021.8.3 in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (from n
Requirement already satisfied: tqdm in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (from nltk) (4.67
Requirement already satisfied: wrapt in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (from smart-open)
Requirement already satisfied: colorama in c:\users\khaye\appdata\local\programs\python\python38\lib\site-packages (from click->n

[notice] A new release of pip is available: 25.0 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip

PROBLEMS 36 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER POSTMAN CONSOLE
PS C:\Users\khaye\Desktop\Machine Learning>
```

Рис 1. Библиотеки

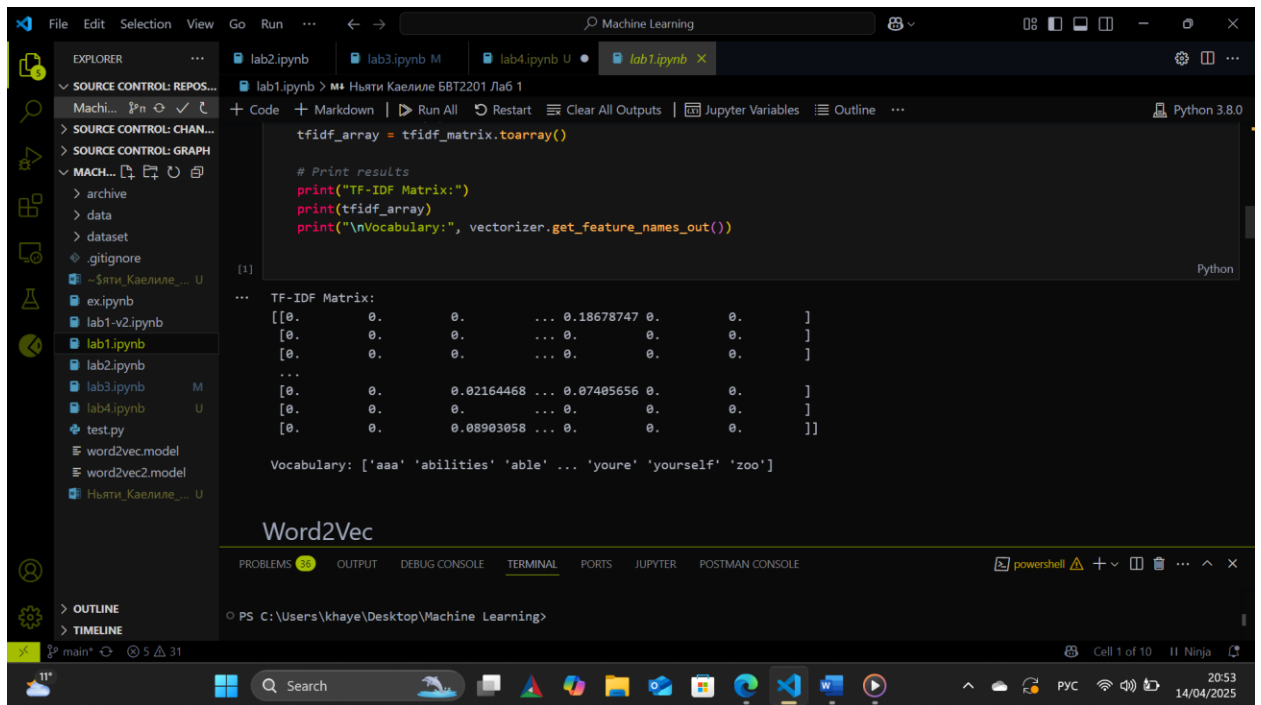
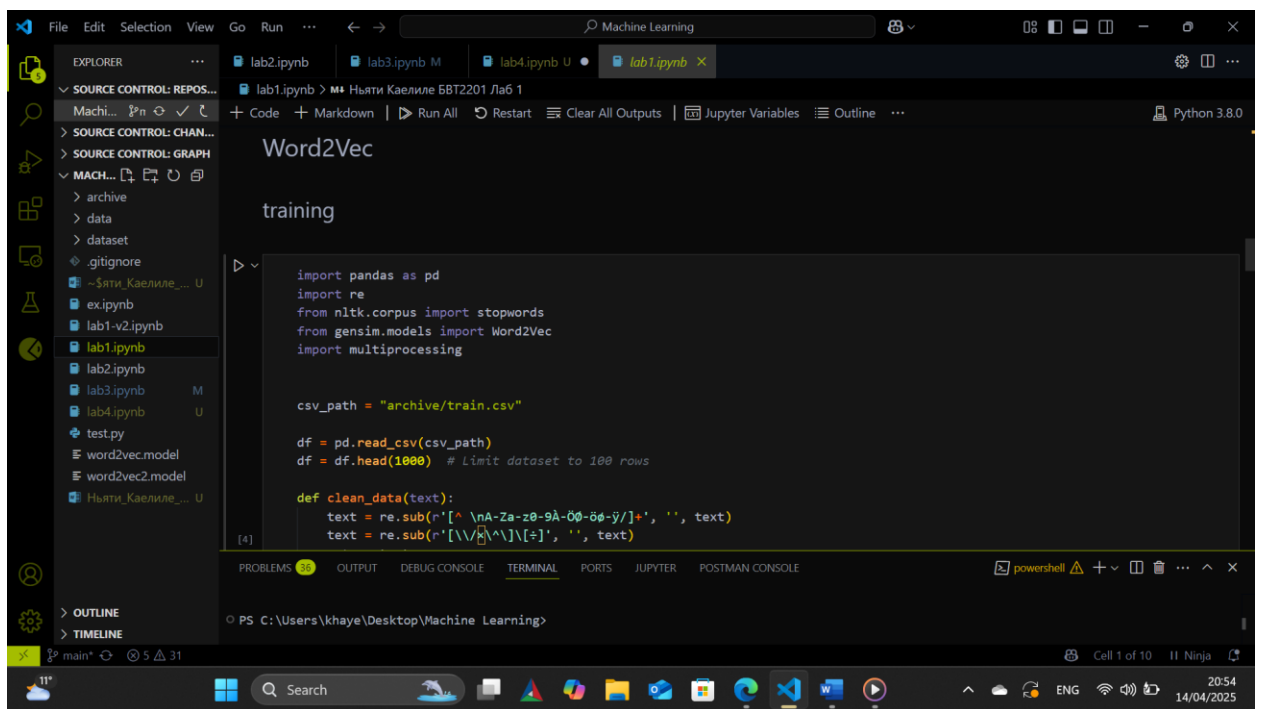


Рис 2. TF-IDF



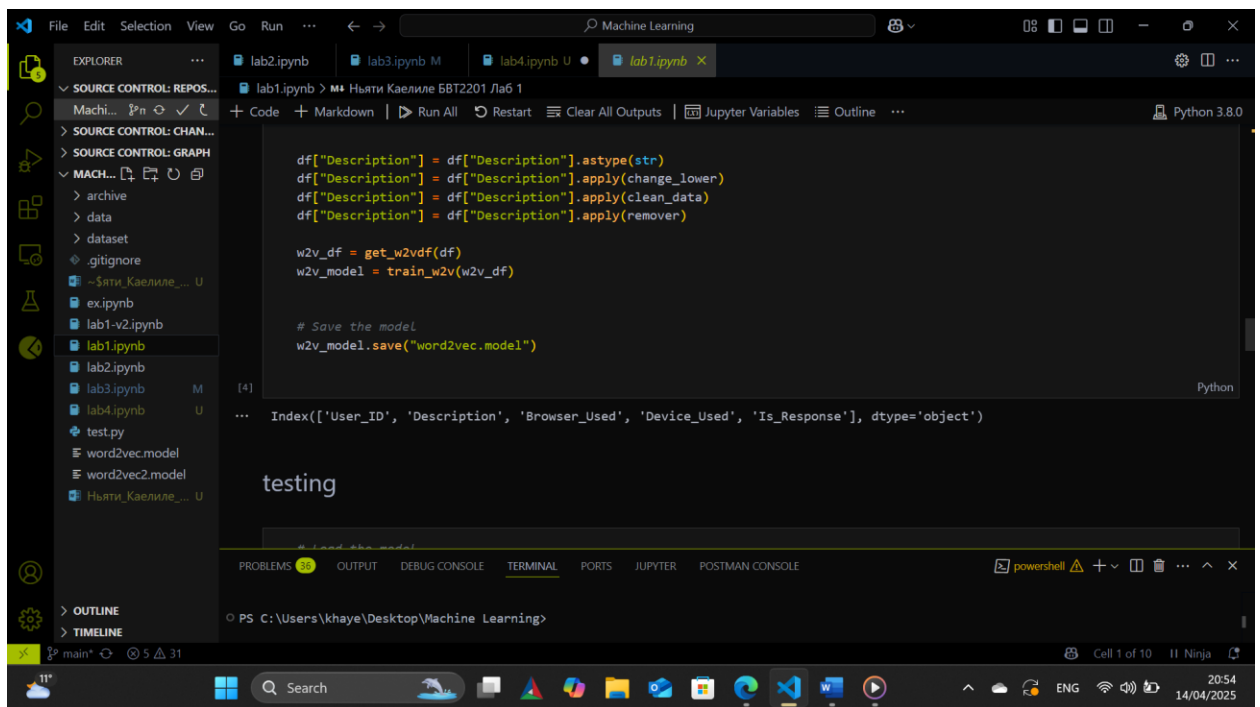


Рис 3. Word2Vec

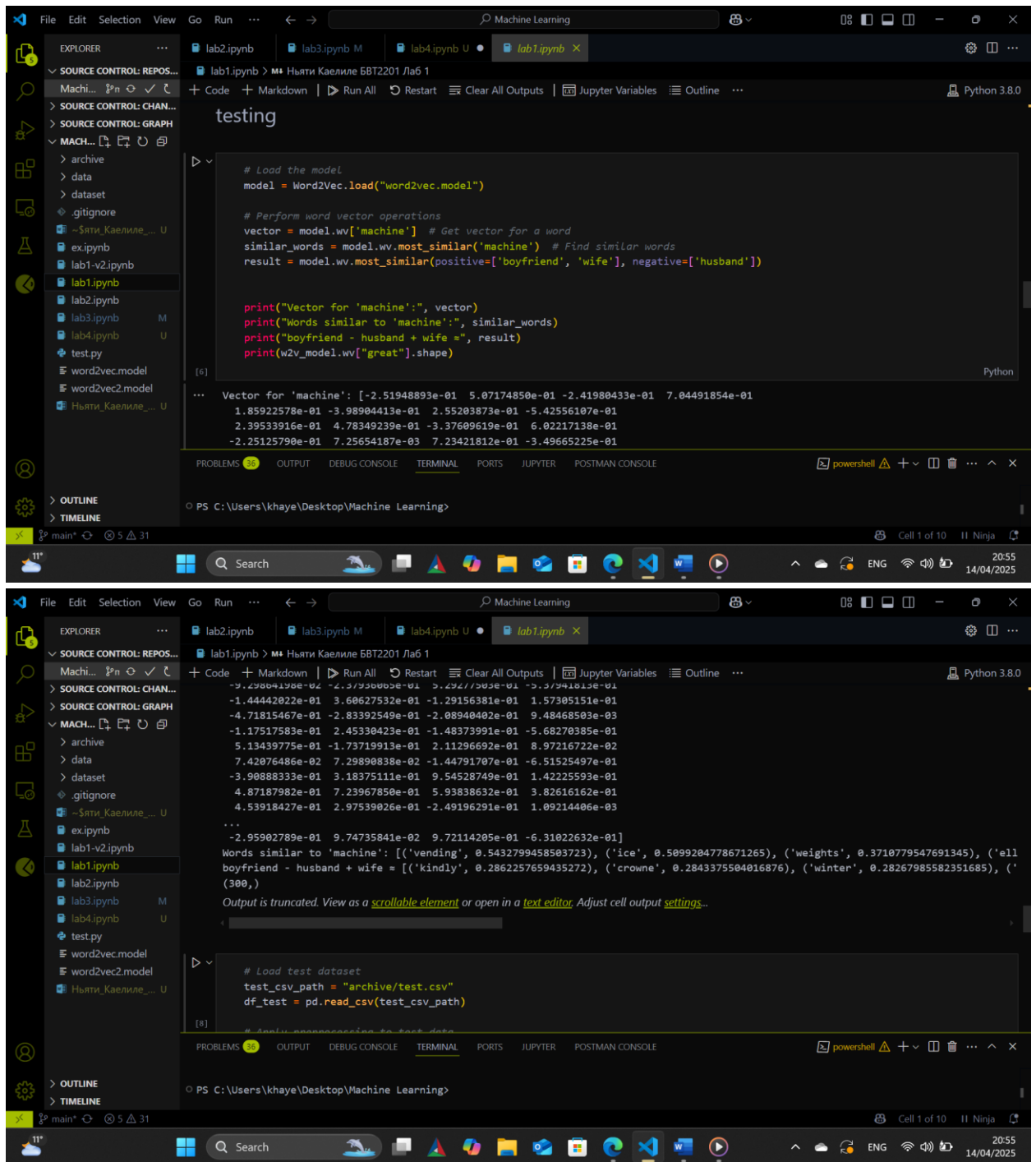


Рис 4. Тестирование

The screenshot shows a Jupyter Notebook with the following code:

```
# Load test dataset
test_csv_path = "archive/test.csv"
df_test = pd.read_csv(test_csv_path)

# Apply preprocessing to test data
df_test["Description"] = df_test["Description"].astype(str)
df_test["Description"] = df_test["Description"].apply(change_lower)
df_test["Description"] = df_test["Description"].apply(clean_data)
df_test["Description"] = df_test["Description"].apply(remove)

# Convert test data into word Lists
w2v_test_data = get_w2vdf(df_test)

# Load the trained Word2Vec model
model = Word2Vec.load("word2vec.model")

# Example: Get word vectors from test data
for sentence in w2v_test_data[:5]: # Show vectors for first 5 rows
    sentence_vectors = [model.wv[word] for word in sentence if word in model.wv]
    print(f"Sentence: {' '.join(sentence)}")
    print(f"Word vectors: {sentence_vectors}\n")
```

The interface includes a file explorer on the left, a terminal at the bottom, and a Windows taskbar at the very bottom.

The screenshot shows the continuation of the Jupyter Notebook with the following code:

```
sentence_vectors = [model.wv[word] for word in sentence if word in model.wv]
print(f"Sentence: {' '.join(sentence)}")
print(f"Word vectors: {sentence_vectors}\n")

result = model.wv.most_similar(positive=['boyfriend', 'wife'], negative=['husband'])
print("boyfriend - husband + wife =", result)

# Example: Find similar words from test data
if "machine" in model.wv:
    similar_words = model.wv.most_similar("machine")
    print(f"Words similar to 'machine': {similar_words}")
else:
    print("'machine' not in vocabulary.")
```

The output of the notebook is displayed below the code:

```
... Sentence: looking motel close proximity tv taping dr phil show chose dunes sunset blvd west hollywood although property displayed
Word vectors: [array([ 6.15962207e-01,  8.04592818e-02, -3.60216856e-01, -1.66775122e-01,
 1.32730510e-03, -1.73207089e-01,  6.84187859e-02,  3.42356294e-01,
 2.67854214e-01,  4.98764098e-01,  4.81809020e-01, -4.42093194e-01,
```

The interface includes a file explorer on the left, a terminal at the bottom, and a Windows taskbar at the very bottom.

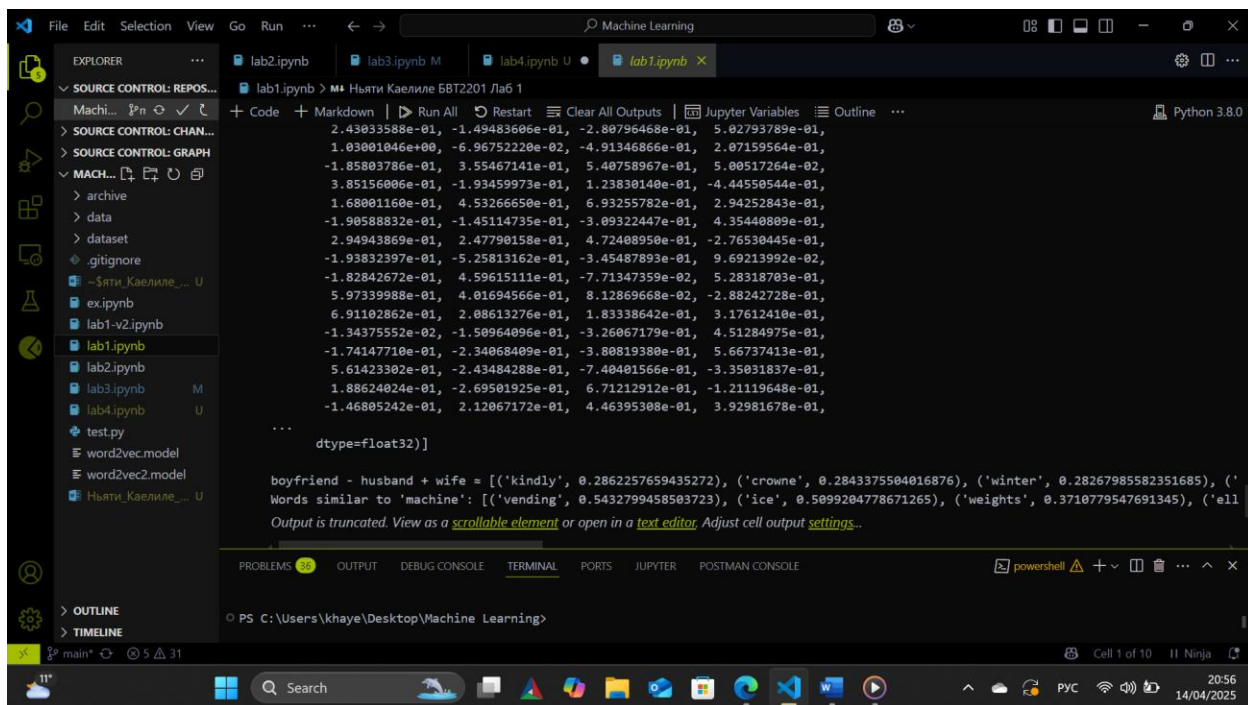


Рис 5. Результаты

3. Выводы

В рамках поставленной задачи была реализована комбинация алгоритмов TF-IDF и word2vec, позволяющая не только учитывать важность слов в контексте (TF-IDF), но и оперировать их семантическими представлениями (word2vec), что дает возможность производить операции над словами, такие как сложение и вычитание в векторном пространстве.

Для обучения и тестирования алгоритмов был подобран соответствующий датасет, обеспечивающий достаточное разнообразие лексики и контекстов. Такая интеграция подходов позволяет эффективно решать задачи семантического анализа и обработки естественного языка.