Введение в обработку естественного языка

Урок 13. Модель BERT и GPT

Практическое задание

Домашнее задание к уроку 13

Задание

1. Взять датасет
   https://huggingface.co/datasets/merionum/ru_paraphraser (https://huggingface.co/datasets/merionum/ru_paraphraser)
   решить задачу парафраза
2. (дополнительно необязательная задача)на выбор взять
   https://huggingface.co/datasets/sberquad (https://huggingface.co/datasets/sberquad)
   https://huggingface.co/datasets/blinoff/medical_qa_ru_data (https://huggingface.co/datasets/blinoff/medical_qa_ru_data)
   натренировать любую модель для вопросно ответной системы
   как альтернатива можно взять любой NER датасет из https://github.com/natasha/corus#reference (https://github.com/natasha/corus#reference) и обучить NER

Выполнил ***Соковнин ИЛ***

Загрузка данных

В [1]:

```
!pip install transformers sentencepiece
```

```
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple,) https://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-python.pkg.dev/colab-wheels/public/simple/)
Collecting transformers
  Downloading transformers-4.20.1-py3-none-any.whl (4.4 MB)
     |████████████████████████████████| 4.4 MB 35.8 MB/s
Collecting sentencepiece
  Downloading sentencepiece-0.1.96-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.2 MB)
     |████████████████████████████████| 1.2 MB 59.4 MB/s
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from transformers) (4.11.4)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (from transformers) (2022.6.2)
Collecting tokenizers!=0.11.3,<0.13,>=0.11.1
  Downloading tokenizers-0.12.1-cp37-cp37m-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (6.6 MB)
     |████████████████████████████████| 6.6 MB 57.6 MB/s
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from transformers) (1.21.6)
Collecting huggingface-hub<1.0,>=0.1.0
  Downloading huggingface_hub-0.8.1-py3-none-any.whl (101 kB)
     |████████████████████████████████| 101 kB 10.6 MB/s
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from transformers) (2.23.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from transformers) (3.7.1)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.7/dist-packages (from transformers) (21.3)
Collecting pyyaml>=5.1
  Downloading PyYAML-6.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (596 kB)
     |████████████████████████████████| 596 kB 68.6 MB/s
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.7/dist-packages (from transformers) (4.64.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0,>=0.1.0->transformers) (4.1.1)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging>=20.0->transformers) (3.0.9)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadata->transformers) (3.8.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests->transformers) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->transformers) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests->transformers) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests->transformers) (2022.6.15)
Installing collected packages: pyyaml, tokenizers, huggingface-hub, transformers, sentencepiece
  Attempting uninstall: pyyaml
    Found existing installation: PyYAML 3.13
    Uninstalling PyYAML-3.13:
      Successfully uninstalled PyYAML-3.13
Successfully installed huggingface-hub-0.8.1 pyyaml-6.0 sentencepiece-0.1.96 tokenizers-0.12.1 transformers-4.20.1
```

В [2]:

```
!pip install datasets
```

```
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple,) https://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-python.pkg.dev/colab-wheel
s/public/simple/)
Collecting datasets
  Downloading datasets-2.3.2-py3-none-any.whl (362 kB)
     |████████████████████████████████| 362 kB 18.3 MB/s
Collecting xxhash
  Downloading xxhash-3.0.0-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (212 kB)
     |████████████████████████████████| 212 kB 70.4 MB/s
Requirement already satisfied: dill<0.3.6 in /usr/local/lib/python3.7/dist-packages (from datasets) (0.3.5.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from datasets) (21.3)
Requirement already satisfied: pyarrow>=6.0.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (6.0.1)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.7/dist-packages (from datasets) (0.70.13)
Collecting responses<0.19
  Downloading responses-0.18.0-py3-none-any.whl (38 kB)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.7/dist-packages (from datasets) (4.64.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from datasets) (1.3.5)
Collecting aiohttp
  Downloading aiohttp-3.8.1-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (1.1 MB)
     |████████████████████████████████| 1.1 MB 63.0 MB/s
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from datasets) (1.21.6)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (2.23.0)
Requirement already satisfied: huggingface-hub<1.0.0,>=0.1.0 in /usr/local/lib/python3.7/dist-packages (from datasets) (0.8.1)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from datasets) (4.11.4)
Collecting fsspec[http]>=2021.05.0
  Downloading fsspec-2022.5.0-py3-none-any.whl (140 kB)
     |████████████████████████████████| 140 kB 73.0 MB/s
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0.0,>=0.1.0->datasets) (3.7.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0.0,>=0.1.0->datasets) (6.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dist-packages (from huggingface-hub<1.0.0,>=0.1.0->datasets) (4.1.1)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging->datasets) (3.0.9)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (2022.6.15)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (2.10)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests>=2.19.0->datasets) (1.24.3)
Collecting urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1
  Downloading urllib3-1.25.11-py2.py3-none-any.whl (127 kB)
     |████████████████████████████████| 127 kB 71.0 MB/s
Collecting frozenlist>=1.1.1
  Downloading frozenlist-1.3.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (144 kB)
     |████████████████████████████████| 144 kB 71.7 MB/s
Collecting yarl<2.0,>=1.0
  Downloading yarl-1.7.2-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_12_x86_64.manylinux2010_x86_64.whl (271 kB)
     |████████████████████████████████| 271 kB 71.0 MB/s
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (21.4.0)
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in /usr/local/lib/python3.7/dist-packages (from aiohttp->datasets) (2.1.0)
Collecting async-timeout<5.0,>=4.0.0a3
  Downloading async_timeout-4.0.2-py3-none-any.whl (5.8 kB)
Collecting aiosignal>=1.1.2
  Downloading aiosignal-1.2.0-py3-none-any.whl (8.2 kB)
Collecting multidict<7.0,>=4.5
  Downloading multidict-6.0.2-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (94 kB)
     |████████████████████████████████| 94 kB 3.9 MB/s
Collecting asynctest==0.13.0
  Downloading asynctest-0.13.0-py3-none-any.whl (26 kB)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadata->datasets) (3.8.0)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas->datasets) (2022.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.7.3->pandas->datasets) (1.15.0)
Installing collected packages: multidict, frozenlist, yarl, urllib3, asynctest, async-timeout, aiosignal, fsspec, aiohttp, xxhash, responses, datasets
  Attempting uninstall: urllib3
    Found existing installation: urllib3 1.24.3
    Uninstalling urllib3-1.24.3:
      Successfully uninstalled urllib3-1.24.3
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency
 conflicts.
datascience 0.10.6 requires folium==0.2.1, but you have folium 0.8.3 which is incompatible.
Successfully installed aiohttp-3.8.1 aiosignal-1.2.0 async-timeout-4.0.2 asynctest-0.13.0 datasets-2.3.2 frozenlist-1.3.0 fsspec-2022.5.0 multidict-6.0.2 responses-
0.18.0 urllib3-1.25.11 xxhash-3.0.0 yarl-1.7.2
```

В [3]:

```
from datasets import load_dataset

corpus = load_dataset('merionum/ru_paraphraser', data_files='plus.jsonl')
```

```
Using custom data configuration merionum--ru_paraphraser-4bf06233b6bfe0ef

Downloading and preparing dataset json/merionum--ru_paraphraser to /root/.cache/huggingface/datasets/merionum___json/merionum--ru_paraphraser-4bf06233b6bfe0ef/0.0.
0/da492aad5680612e4028e7f6ddc04b1dfcec4b64db470ed7cc5f2bb265b9b6b5...

Downloading data files:   0%|          | 0/1 [00:00<?, ?it/s]

Downloading data:   0%|          | 0.00/875M [00:00<?, ?B/s]

Extracting data files:   0%|          | 0/1 [00:00<?, ?it/s]

0 tables [00:00, ? tables/s]

Dataset json downloaded and prepared to /root/.cache/huggingface/datasets/merionum___json/merionum--ru_paraphraser-4bf06233b6bfe0ef/0.0.0/da492aad5680612e4028e7f6dd
c04b1dfcec4b64db470ed7cc5f2bb265b9b6b5. Subsequent calls will reuse this data.

  0%|          | 0/1 [00:00<?, ?it/s]
```

В [4]:

```
corpus
```

Out[4]:

```
DatasetDict({
    train: Dataset({
        features: ['id', 'paraphrases'],
        num_rows: 1725393
    })
})
```

В [5]:

```
corpus['train']
```

Out[5]:

```
Dataset({
    features: ['id', 'paraphrases'],
    num_rows: 1725393
})
```

В [6]:

```
corpus['train'][0:3]
```

Out[6]:

```
{'id': ['0', '1', '2'],
 'paraphrases': [['Птичий праздник: Что такое "куриное рождество" и кто его отмечает',
   'Куриное Рождество в 2019 году: что это за праздник и как стать его главным героем'],
  ['В Индии 5 млн женщин выстроились стеной в 620 км ради равенства (Фото)',
   'Индийские женщины выстроили живую стену длиной 620 километров, требуя равенства',
   'В Индии женщины встали живой стеной длиной 620 км'],
  ['Томас Бах: "Олимпийский комитет России понес достаточное наказание"',
   'Президент МОК Бах: Россия уже достаточно наказана за допинг',
   'Глава МОК заявил, что ОКР понес наказание за манипуляции с пробами на ОИ в Сочи']]}
```

В [7]:

```
corpus['train']['paraphrases'][0]
```

Out[7]:

```
['Птичий праздник: Что такое "куриное рождество" и кто его отмечает',
 'Куриное Рождество в 2019 году: что это за праздник и как стать его главным героем']
```

В [8]:

```
corpus['train']['paraphrases'][0][0]
```

Out[8]:

```
'Птичий праздник: Что такое "куриное рождество" и кто его отмечает'
```

В [9]:

```
!pip install -U sentence-transformers
```

```
Looking in indexes: https://pypi.org/simple, (https://pypi.org/simple,) https://us-python.pkg.dev/colab-wheels/public/simple/ (https://us-python.pkg.dev/colab-wheel
s/public/simple/)
Collecting sentence-transformers
  Downloading sentence-transformers-2.2.2.tar.gz (85 kB)
      |████████████████████████████████| 85 kB 3.5 MB/s
Requirement already satisfied: transformers<5.0.0,>=4.6.0 in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (4.20.1)
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (4.64.0)
Requirement already satisfied: torch>=1.6.0 in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (1.11.0+cu113)
Requirement already satisfied: torchvision in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (0.12.0+cu113)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (1.21.6)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (1.0.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (1.4.1)
Requirement already satisfied: nltk in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (3.7)
Requirement already satisfied: sentencepiece in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (0.1.96)
Requirement already satisfied: huggingface-hub>=0.4.0 in /usr/local/lib/python3.7/dist-packages (from sentence-transformers) (0.8.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.7/dist-packages (from huggingface-hub>=0.4.0->sentence-transformers) (6.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dist-packages (from huggingface-hub>=0.4.0->sentence-transformers) (4.1.1)
Requirement already satisfied: filelock in /usr/local/lib/python3.7/dist-packages (from huggingface-hub>=0.4.0->sentence-transformers) (3.7.1)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from huggingface-hub>=0.4.0->sentence-transformers) (2.23.0)
Requirement already satisfied: packaging>=20.9 in /usr/local/lib/python3.7/dist-packages (from huggingface-hub>=0.4.0->sentence-transformers) (21.3)
Requirement already satisfied: importlib-metadata in /usr/local/lib/python3.7/dist-packages (from huggingface-hub>=0.4.0->sentence-transformers) (4.11.4)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging>=20.9->huggingface-hub>=0.4.0->sentence-transforme
rs) (3.0.9)
Requirement already satisfied: tokenizers!=0.11.3,<0.13,>=0.11.1 in /usr/local/lib/python3.7/dist-packages (from transformers<5.0.0,>=4.6.0->sentence-transformers)
 (0.12.1)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.7/dist-packages (from transformers<5.0.0,>=4.6.0->sentence-transformers) (2022.6.2)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadata->huggingface-hub>=0.4.0->sentence-transformers) (3.8.0)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from nltk->sentence-transformers) (7.1.2)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from nltk->sentence-transformers) (1.1.0)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->huggingface-hub>=0.4.0->sentence-transformers) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests->huggingface-hub>=0.4.0->sentence-transformers) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests->huggingface-hub>=0.4.0->sentence-tr
ansformers) (1.25.11)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests->huggingface-hub>=0.4.0->sentence-transformers) (2022.6.1
5)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from scikit-learn->sentence-transformers) (3.1.0)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in /usr/local/lib/python3.7/dist-packages (from torchvision->sentence-transformers) (7.1.2)
Building wheels for collected packages: sentence-transformers
  Building wheel for sentence-transformers (setup.py) ... done
  Created wheel for sentence-transformers: filename=sentence_transformers-2.2.2-py3-none-any.whl size=125938 sha256=ea9d9a76fa9ddf4e68e7902d65b994cfc0712d888e89df21
6b9c5cbcdc80c348
  Stored in directory: /root/.cache/pip/wheels/bf/06/fb/d59c1e5bd1dac7f6cf61ec0036cc3a10ab8fecaa6b2c3d3ee9
Successfully built sentence-transformers
Installing collected packages: sentence-transformers
Successfully installed sentence-transformers-2.2.2
```

Используем предобученную модель: cointegrated/rut5-base-paraphraser

https://huggingface.co/cointegrated/rut5-base-paraphraser (https://huggingface.co/cointegrated/rut5-base-paraphraser)

В [10]:

```
from transformers import T5ForConditionalGeneration, T5Tokenizer
MODEL_NAME = 'cointegrated/rut5-base-paraphraser'
model = T5ForConditionalGeneration.from_pretrained(MODEL_NAME)
tokenizer = T5Tokenizer.from_pretrained(MODEL_NAME)
model.cuda();
model.eval();

def paraphrase(text, beams=5, grams=4, do_sample=False):
    x = tokenizer(text, return_tensors='pt', padding=True).to(model.device)
    max_size = int(x.input_ids.shape[1] * 1.5 + 10)
    out = model.generate(**x, encoder_no_repeat_ngram_size=grams, num_beams=beams, max_length=max_size, do_sample=do_sample)
    return tokenizer.decode(out[0], skip_special_tokens=True)

print(paraphrase('Каждый охотник желает знать, где сидит фазан.'))
# Все охотники хотят знать где фазан сидит.
```

```
Downloading:   0%|          | 0.00/724 [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/932M [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/808k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/65.0 [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/315 [00:00<?, ?B/s]

Все охотники хотят знать где фазан сидит.
```

Используем предобученную модель: cointegrated/rubert-base-cased-dp-paraphrase-detection

https://huggingface.co/cointegrated/rubert-base-cased-dp-paraphrase-detection (https://huggingface.co/cointegrated/rubert-base-cased-dp-paraphrase-detection)

В [11]:

```
import torch
from transformers import AutoModelForSequenceClassification, BertTokenizer
model_name = 'cointegrated/rubert-base-cased-dp-paraphrase-detection'

model = AutoModelForSequenceClassification.from_pretrained(model_name).cuda()
tokenizer = BertTokenizer.from_pretrained(model_name)

def compare_texts(text1, text2):
    batch = tokenizer(text1, text2, return_tensors='pt').to(model.device)
    with torch.inference_mode():
        proba = torch.softmax(model(**batch).logits, -1).cpu().numpy()

    p = proba[0]  # p(non-paraphrase), p(paraphrase)
    if p[1] > 0.5:
        paraphrase = "paraphrase"
    else:
        paraphrase = "non"

    return [text1, text2, p[0], p[1], paraphrase]


print(compare_texts('Сегодня на улице хорошая погода', 'Сегодня на улице отвратительная погода'))
# [0.7056226 0.2943774]
print(compare_texts('Сегодня на улице хорошая погода', 'Отличная погодка сегодня выдалась'))
# [0.16524374 0.8347562 ]
```

```
Downloading:   0%|          | 0.00/1.09k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/679M [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/1.57M [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/112 [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/406 [00:00<?, ?B/s]

['Сегодня на улице хорошая погода', 'Сегодня на улице отвратительная погода', 0.7056231, 0.29437694, 'non']
['Сегодня на улице хорошая погода', 'Отличная погодка сегодня выдалась', 0.1652439, 0.83475614, 'paraphrase']
```

В [12]:

```
def compare(text1, text2):
    """  """

    batch = tokenizer(text1, text2, return_tensors='pt').to(model.device)
    with torch.inference_mode():
        proba = torch.softmax(model(**batch).logits, -1).cpu().numpy()

    p = proba[0]  # p(non-paraphrase), p(paraphrase)
    paraphrase = "non"

    if p[1] > 0.5:
        paraphrase = "paraphrase"

    # row = {'текст1': text1, 'текст2': text2, 'вероятность non-paraphrase': p[0], 'вероятность paraphrase': p[1], 'paraphrase': paraphrase}
    # return row
    return [text1, text2, p[0], p[1], paraphrase]
```

В [13]:

```
import pandas as pd

pd.set_option('display.max_colwidth', 100)
# pd.set_option('display.width', 500)
```

В [15]:

```
result = []

result.append(compare('Сегодня на улице хорошая погода', 'Сегодня на улице отвратительная погода'))
result.append(compare('Цены на нефть восстанавливаются', 'Парламент Словакии поблагодарил народы'))
result.append(compare('Сегодня на улице хорошая погода', 'Сегодня на улице прекрасная погода'))
n=5
for i in range(n):
  text1 = corpus['train']['paraphrases'][i][0]
  text2 =corpus['train']['paraphrases'][i][1]
  result.append(compare(text1, text2))



col_names =  ['текст1', 'текст2', 'вероятность non-paraphrase', 'вероятность paraphrase', 'paraphrase']
df = pd.DataFrame(result, columns=col_names)
df.head(10)
```

Out[15]:

	текст1	текст2	вероятность non-paraphrase	вероятность paraphrase	paraphrase
0	Сегодня на улице хорошая погода	Сегодня на улице отвратительная погода	0.705623	0.294377	non
1	Цены на нефть восстанавливаются	Парламент Словакии поблагодарил народы	0.944955	0.055045	non
2	Сегодня на улице хорошая погода	Сегодня на улице прекрасная погода	0.035242	0.964758	paraphrase
3	Птичий праздник: Что такое "куриное рождество" и кто его отмечает	Куриное Рождество в 2019 году: что это за праздник и как стать его главным героем	0.157879	0.842121	paraphrase
4	В Индии 5 млн женщин выстроились стеной в 620 км ради равенства (Фото)	Индийские женщины выстроили живую стену длиной 620 километров, требуя равенства	0.018209	0.981791	paraphrase
5	Томас Бах: "Олимпийский комитет России понес достаточное наказание"	Президент МОК Бах: Россия уже достаточно наказана за допинг	0.028516	0.971484	paraphrase
6	Россия не представила WADA данные по пробам в срок и может быть наказана	WADA: Россия не предоставила нам доступ к допинг-пробам	0.058580	0.941420	paraphrase
7	Антидопинговое агентство США призвало ВАДА отстранить российских спортсменов от соревнований	Комиссия спортсменов WADA потребовала отстранить РУСАДА	0.428032	0.571968	paraphrase

В [ ]: