Задание

1. Взять предобученную трансформерную архитектуру и решить задачу перевода (для тогоже корпуса что вы выбрали из предыдущего дз)
2. скачиваем готовый новостной датасет
   !wget https://github.com/ods-ai-ml4sg/proj_news_viz/releases/download/data/gazeta.csv.gz (https://github.com/ods-ai-ml4sg/proj_news_viz/releases/download/data/gazeta.csv.gz)

   ```
   # пример работы с ним
   from corus import load_ods_gazeta
   path = 'gazeta.csv.gz'
   records = load_ods_gazeta(path)
   next(records)
   ```

   реализовать метод поиска ближайших статей
   (на вход метода должен приходить запрос (какой-то вопрос) и количество вариантов вывода к примеру топ 5-ть или 3-ри, ваш метод должен возвращать топ-k ближайших статей к этому запросу)
   визуально оценить качество

In [19]:

```
import transformers
from transformers import AutoTokenizer, BertTokenizer

import io

from scipy.spatial import distance
```

In [2]:

```
from transformers import pipeline
from pprint import pprint

nlp = pipeline("text-generation", model= "sberbank-ai/mGPT")
```

In [6]:

```
nlp('какие новости?', do_sample=False)[0].get('generated_text')
```

```
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
```

Out[6]:

```
'какие новости?\nВот и я не знаю, что будет дальше.\n'
```

translete

In [56]:

```
!pip install sentencepiece transformers[sentencepiece]
```

```
Requirement already satisfied: sentencepiece in c:\program files\python37\li
b\site-packages (0.1.96)
Requirement already satisfied: transformers[sentencepiece] in c:\program fil
es\python37\lib\site-packages (4.21.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.1.0 in c:\program fil
es\python37\lib\site-packages (from transformers[sentencepiece]) (0.8.1)
Requirement already satisfied: importlib-metadata in c:\program files\python
37\lib\site-packages (from transformers[sentencepiece]) (4.8.1)
Requirement already satisfied: filelock in c:\program files\python37\lib\sit
e-packages (from transformers[sentencepiece]) (3.7.1)
Requirement already satisfied: regex!=2019.12.17 in c:\program files\python3
7\lib\site-packages (from transformers[sentencepiece]) (2021.11.2)
Requirement already satisfied: tqdm>=4.27 in c:\program files\python37\lib\s
ite-packages (from transformers[sentencepiece]) (4.62.3)
Requirement already satisfied: numpy>=1.17 in c:\program files\python37\lib
\site-packages (from transformers[sentencepiece]) (1.21.4)
Requirement already satisfied: requests in c:\program files\python37\lib\sit
e-packages (from transformers[sentencepiece]) (2.26.0)
Requirement already satisfied: packaging>=20.0 in c:\program files\python37
\lib\site-packages (from transformers[sentencepiece]) (21.3)
Requirement already satisfied: tokenizers!=0.11.3,<0.13,>=0.11.1 in c:\progr
am files\python37\lib\site-packages (from transformers[sentencepiece]) (0.1
2.1)
Requirement already satisfied: pyyaml>=5.1 in c:\program files\python37\lib
\site-packages (from transformers[sentencepiece]) (6.0)
Requirement already satisfied: protobuf<=3.20.1 in c:\program files\python37
\lib\site-packages (from transformers[sentencepiece]) (3.19.4)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\program file
s\python37\lib\site-packages (from huggingface-hub<1.0,>=0.1.0->transformers
[sentencepiece]) (3.10.0.2)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in c:\program files
\python37\lib\site-packages (from packaging>=20.0->transformers[sentencepiec
e]) (2.4.7)
Requirement already satisfied: colorama in c:\program files\python37\lib\sit
e-packages (from tqdm>=4.27->transformers[sentencepiece]) (0.4.4)
Requirement already satisfied: zipp>=0.5 in c:\program files\python37\lib\si
te-packages (from importlib-metadata->transformers[sentencepiece]) (3.6.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\program files\python
37\lib\site-packages (from requests->transformers[sentencepiece]) (2021.10.
8)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\program files
\python37\lib\site-packages (from requests->transformers[sentencepiece]) (2.
0.7)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\program files\pyt
hon37\lib\site-packages (from requests->transformers[sentencepiece]) (1.26.
7)
Requirement already satisfied: idna<4,>=2.5 in c:\program files\python37\lib
\site-packages (from requests->transformers[sentencepiece]) (3.2)

[notice] A new release of pip available: 22.1.2 -> 22.2.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

Превод ru -> en

In [41]:

```
from transformers import pipeline
translator = pipeline("translation", model = 'Helsinki-NLP/opus-mt-ru-en')
print(translator("сегодня увидел за окном радугу", max_length=40))
```

```
c:\program files\python37\lib\site-packages\transformers\models\marian\token
ization_marian.py:198: UserWarning: Recommended: pip install sacremoses.
  warnings.warn("Recommended: pip install sacremoses.")

[{'translation_text': 'I saw a rainbow outside the window today.'}]
```

In [42]:

```
print(translator("Привет мир", max_length=70))
```

```
[{'translation_text': 'Hey, world.'}]
```

Превод en -> ru

In [43]:

```
from transformers import pipeline
translator_en = pipeline("translation", model = 'Helsinki-NLP/opus-mt-en-ru')
print(translator_en("I saw a rainbow outside the window today.", max_length=40))
```

```
Downloading                                293M/293M [02:04<00:00,
pytorch_model.bin: 100%                    1.95MB/s]

Downloading                                42.0/42.0 [00:00<00:00,
tokenizer_config.json: 100%                609B/s]

Downloading source.spm:                    784k/784k [00:02<00:00,
100%                                       767kB/s]

Downloading target.spm:                    1.03M/1.03M [00:02<00:00,
100%                                       652kB/s]

Downloading vocab.json:                    2.48M/2.48M [00:04<00:00,
100%                                       1.01MB/s]

[{'translation_text': 'Сегодня я видел радугу за окном.'}]
```

In [44]:

```
print(translator_en("hello world", max_length=70))
```

```
[{'translation_text': 'Приветствую мир'}]
```

In [ ]:

```
```

2

In [2]:

```
import numpy as np
import tensorflow as tf
from transformers import TFAutoModel, AutoTokenizer
```

загружаю модель и токенайзер

In [16]:

```
bert = TFAutoModel.from_pretrained("Geotrend/bert-base-ru-cased")
tokenizer = AutoTokenizer.from_pretrained("Geotrend/bert-base-ru-cased")
```

```
Some layers from the model checkpoint at Geotrend/bert-base-ru-cased were no
t used when initializing TFBertModel: ['mlm___cls']
- This IS expected if you are initializing TFBertModel from the checkpoint o
f a model trained on another task or with another architecture (e.g. initial
izing a BertForSequenceClassification model from a BertForPreTraining mode
l).
- This IS NOT expected if you are initializing TFBertModel from the checkpoi
nt of a model that you expect to be exactly identical (initializing a BertFo
rSequenceClassification model from a BertForSequenceClassification model).
All the layers of TFBertModel were initialized from the model checkpoint at
Geotrend/bert-base-ru-cased.
If your task is similar to the task the model of the checkpoint was trained
on, you can already use TFBertModel for predictions without further trainin
g.
```

In [3]:

```
tok = tokenizer(["Погода сегодня отличная в отличии от прошлой недели.", "Вечером идёт силь
                max_length=20, truncation=True, padding='max_length', return_token_type_ids=
```

In [4]:

```
tok
```

Out[4]:

```
{'input_ids': <tf.Tensor: shape=(2, 20), dtype=int32, numpy=
array([[   11,   965, 10728, 10148,   392,   625,  4266,   167,   392,
          625,  1870,   288,   392,  1436,  4668,   429, 11241,    27,
           12,     0],
       [   11,   137,  9077,  3274, 11985,  3708,  3582,   403,  1203,
         6243,  4985,   403,   184,  4539,    14,    12,     0,     0,
            0,     0]])>, 'attention_mask': <tf.Tensor: shape=(2, 20), dtype
=int32, numpy=
array([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0],
       [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0]])>}
```

In [5]:

```
out = bert(**tok)
```

In [6]:

```
out
```

Out[6]:

```
TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=<tf.Tensor:
shape=(2, 20, 768), dtype=float32, numpy=
array([[[ 0.03819018,  0.07156943,  0.09131993, ..., -0.1835718 ,
          0.01867675, -0.22367653],
        [ 0.14797847, -0.17483613,  0.17382078, ..., -0.41834867,
         -0.21098769, -0.5299227 ],
        [ 0.05723458, -0.5699748 ,  0.49159086, ..., -0.30470586,
          0.0797454 , -0.6216985 ],
        ...,
        [-0.01815331, -0.1583628 ,  0.698445  , ..., -0.4177136 ,
         -0.21134536, -0.17493863],
        [ 0.20085657, -0.08146185,  0.7731839 , ..., -0.570418  ,
         -0.07559601, -0.45889935],
        [ 0.0134001 , -0.17168729,  0.67724174, ..., -0.31340474,
         -0.16356654, -0.41178197]],

       [[ 0.06399527,  0.1544463 , -0.13583916, ...,  0.15447147,
          0.46281058, -0.09160701],
        [ 0.2685944 , -0.20166266,  0.37964725, ...,  0.33930385,
          0.65223634, -0.41767874],
        [-0.23570526,  0.10125846,  0.02285784, ..., -0.5501144 ,
          0.8288276 ,  0.06447223],
        ...,
        [-0.01996499,  0.20887825,  0.14996056, ...,  0.17891084,
          0.34426045, -0.18655458],
        [ 0.04150454,  0.15244213,  0.15443751, ..., -0.04336701,
          0.36998278,  0.1036798 ],
        [ 0.21945743,  0.04559727,  0.23187432, ..., -0.09085312,
          0.3441566 , -0.2172923 ]]], dtype=float32)>, pooler_output=<tf.Ten
sor: shape=(2, 768), dtype=float32, numpy=
array([[-0.3267656 ,  0.1162425 , -0.08803337, ...,  0.3300892 ,
         0.08718193, -0.00109257],
       [-0.34685668,  0.22584419,  0.09379828, ...,  0.19157435,
         0.32265192, -0.27708763]], dtype=float32)>, past_key_values=None, h
idden_states=None, attentions=None, cross_attentions=None)
```

Загружаю данные

In [8]:

```
# по сути данные те же можно пользоваться любым способом загрузки
from datasets import load_dataset

dataset = load_dataset('IlyaGusev/gazeta', revision="v1.0")["train"]
```

```
No config specified, defaulting to: gazeta/default
Reusing dataset gazeta (C:\Users\VoronkovSergey\.cache\huggingface\datasets
\IlyaGusev___gazeta\default\1.0.0\ef9349c3c0f3112ca4036520d76c4bc1b8a79d30bc
29643c6cae5a094d44e457)

  0%|          | 0/3 [00:00<?, ?it/s]
```

In [3]:

```
import pandas as pd
```

In [10]:

```
df = pd.DataFrame(dataset)
```

In [22]:

```
# df.to_pickle('news.pkl')
```

In [11]:

```
df.head()
```

Out[11]:

	text	summary	title	date	
0	«По итогам 2011 года чистый отток может состав...	В 2011 году из России уйдет $80 млрд, считают ...	Прогноз не успевает за оттоком	2011-11-30 18:33:39	https://www.gazeta.ru/financial/2011/
1	Российское подразделение интернет-корпорации G...	Юлия Соловьева, экс-директор холдинга «Профмед...	Google закончил поиск	2013-01-24 18:20:09	https://www.gazeta.ru/business/2013/0
2	Басманный районный суд Москвы вечером 6 феврал...	Суд арестовал на два месяца четверых экс-чинов...	«Фигуранты дела могут давить на свидетелей»	2018-02-06 21:21:14	https://www.gazeta.ru/social/2018/02/(
3	Как повлияло вступление в ВТО на конкурентносп...	Мнения предпринимателей по поводу вступления в...	«С последних традиционно «отжимают» больше»	2013-06-21 17:43:50	https://www.gazeta.ru/business/2013/0
4	К третьему сезону «Голос» на Первом канале ста...	На Первом канале завершился третий сезон шоу «...	Третий «Голос» за Градского	2014-12-27 01:10:01	https://www.gazeta.ru/culture/2014/12

In [72]:

```
data = df[['text', 'summary']]
```

In [73]:

```
data.head()
```

Out[73]:

	text	summary
0	«По итогам 2011 года чистый отток может состав...	В 2011 году из России уйдет $80 млрд, считают ...
1	Российское подразделение интернет-корпорации G...	Юлия Соловьева, экс-директор холдинга «Профмед...
2	Басманный районный суд Москвы вечером 6 феврал...	Суд арестовал на два месяца четверых экс-чинов...
3	Как повлияло вступление в ВТО на конкурентносп...	Мнения предпринимателей по поводу вступления в...
4	К третьему сезону «Голос» на Первом канале ста...	На Первом канале завершился третий сезон шоу «...

функция для преобразования в вектор

вектора будут длины самой длинной последовательности

In [70]:

```
count = 0
for i, data in data.iterrows():
#     print(len(data['summary'].split()))
    if count < len(data['summary'].split()):
        count = len(data['summary'].split())
print(count)
```

```
73
```

In [14]:

```
def emb(data):
    tokinization = tokenizer([data],
                max_length=73, truncation=True, padding='max_length', return_token_type_ids=
    out_emb = bert(**tokinization)
    out_emb = out_emb[1].numpy()
    return out_emb
```

In [75]:

```
data.shape
```

Out[75]:

```
(52400, 2)
```

In [76]:

```
data = data.sample(1000)
```

In [77]:

```
data.shape
```

Out[77]:

```
(1000, 2)
```

делаю эмбединги каждого текста в колонке summary

In [78]:

```
data['embrding'] = data['summary'].apply(emb)
```

In [79]:

```
data.head()
```

Out[79]:

	text	summary	embrding
9758	Две главные точки притяжения россиян, покупающ...	Курортную недвижимость на берегу океана в США ...	[[-0.34230134, 0.082322955, 0.06668175, -0.155...
43431	Министр обороны Сергей Шойгу обрушился с крити...	Министр обороны Сергей Шойгу потребовал вернут...	[[-0.22858617, 0.0655227, 0.024531558, -0.1688...
26232	Общество защиты прав потребителей (ОЗПП) подал...	Компания МТС нарушала закон, посылая абонентам...	[[-0.3335981, 0.06889029, -0.027939957, -0.138...
40167	В центре Москвы произошел крупный пожар. Огонь...	Крупный пожар вспыхнул утром в центре Москвы, ...	[[-0.24017225, 0.08606251, -0.06971901, -0.005...
16675	В России, увы, пока до подобных стандартов кач...	В странах Европейского союза действует «Директ...	[[-0.30010387, 0.031129533, 0.07411454, -0.060...

In [80]:

```
# data.to_pickle('data_emb.pkl')
```

In [4]:

```
data = pd.read_pickle('data_emb.pkl')
```

In [5]:

```
from scipy.cluster.hierarchy import linkage, dendrogram
import matplotlib.pyplot as plt
```

делаю список эмбедингов

In [6]:

```
samples = data['embrding'].values
```

In [7]:

```
samples = [i[0] for i in samples]
```

In [18]:

```

```

In [129]:

```
distance.cosine(samples[0], samples[1])
```

Out[129]:

```
0.2532772421836853
```

In [9]:

```
data['embrding'] = data['embrding'].apply(lambda x: x[0])
```

In [10]:

```
data.head()
```

Out[10]:

	text	summary	embrding
9758	Две главные точки притяжения россиян, покупающ...	Курортную недвижимость на берегу океана в США ...	[-0.34230134, 0.082322955, 0.06668175, -0.1551...
43431	Министр обороны Сергей Шойгу обрушился с крити...	Министр обороны Сергей Шойгу потребовал вернут...	[-0.22858617, 0.0655227, 0.024531558, -0.16884...
26232	Общество защиты прав потребителей (ОЗПП) подал...	Компания МТС нарушала закон, посылая абонентам...	[-0.3335981, 0.06889029, -0.027939957, -0.1389...
40167	В центре Москвы произошел крупный пожар. Огонь...	Крупный пожар вспыхнул утром в центре Москвы, ...	[-0.24017225, 0.08606251, -0.06971901, -0.0053...
16675	В России, увы, пока до подобных стандартов кач...	В странах Европейского союза действует «Директ...	[-0.30010387, 0.031129533, 0.07411454, -0.0606...

для поиска похожих 5 текстов я буду использовать косинусное расстояние

In [11]:

```
result = {}
output_text = []
def distance_cos(x, samples):
    x = emb(x)
    for i in samples:
        dis = distance.cosine(i, x)
        if len(result) < 5:
            result[dis] = i
        elif max(result) > dis:
            result[dis] = result.pop(max(result))
            result[dis] = i
    return result

#     for i in result.values():
```

In [29]:

```
test = data['summary'].iloc[1]
```

In [30]:

```
# test = 'Министр обороны Сергей Шойгу потребовал вернуть госпиталю им. Бурденко репутацию
res = distance_cos(test, samples)
```

In [31]:

```
res.keys()
```

Out[31]:

```
dict_keys([0.06345844268798828, 0.05766087770462036, 0.0590246319770813, 0.0
6165957450866699, 0])
```

функция возврата предложения по вектору

In [37]:

```
result_text = []
def equality():
    for i, d in data.iterrows():
        for j in res.values():
            x = set(d['embrding'])
            j = set(j)
            if x ==  j:
                result_text.append(d['summary'])
    return result_text
```

In [38]:

```
text_predict = equality()
```

In [39]:

```
test
```

Out[39]:

```
'Министр обороны Сергей Шойгу потребовал вернуть госпиталю им. Бурденко репу
тацию «образцовой клиники». Для этого его руководству одновременно выделили
средства и пригрозили кадровыми решениями. По мнению же экспертов, несмотря
на потерю былого ореола, у главного госпиталя страны дела не так плохи. А в
первую очередь спасать надо небольшие госпитали в регионах.'
```

In [40]:

```
print("\n\n".join(text_predict))
```

```
Министр обороны Сергей Шойгу потребовал вернуть госпиталю им. Бурденко репут
ацию «образцовой клиники». Для этого его руководству одновременно выделили с
редства и пригрозили кадровыми решениями. По мнению же экспертов, несмотря н
а потерю былого ореола, у главного госпиталя страны дела не так плохи. А в п
ервую очередь спасать надо небольшие госпитали в регионах.

Немецкий телеканал ответил на требование Киева выгнать солиста группы Scoote
r из состава жюри одной из популярных программ после того, как коллектив выс
тупил с концертом в Крыму.

Музыкант Сергей Шнуров рассказал, что из совместно нажитого имущества достал
ось его бывшей супруге Матильде Мозговой после развода. По словам артиста, ж
енщине отошло «больше половины».

С помощью молотков и дымовых шашек в Петербурге поменяли гендиректора сети г
ипермакетов «Лента». Прежний руководитель считает свою отставку незаконной и
намерен обратиться в прокуратуру.

В Армении уволен мэр Еревана. Решение было принято на следующий день после д
раки, которую устроил градоначальник армянской столицы: он вступился за чест
ь супруги на концерте Пласидо Доминго. Теперь следователи ищут состав престу
пления в его действиях.
```

Вывод

в целом текста которые по косинусному расстоянию оказались ближе можно отнести в одной тематике, думаю если делать на полных текстах точность будет лучше