Задание

Взять тот же датасет, который был на вебинаре и предобученную модель для задачи суммаризации

1. Проверить насколько хорошо она суммаризирует
2. (дополнительно) Сделать генерацию заголовков для статьи (обучить модель для генерации заголовков)

In [1]:

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

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

```
Using the latest cached version of the module from C:\Users\VoronkovSergey\.
cache\huggingface\modules\datasets_modules\datasets\IlyaGusev--gazeta\ef9349
c3c0f3112ca4036520d76c4bc1b8a79d30bc29643c6cae5a094d44e457 (last modified on
Wed Aug 10 21:41:36 2022) since it couldn't be found locally at IlyaGusev/ga
zeta., or remotely on the Hugging Face Hub.
No config specified, defaulting to: gazeta/default
Reusing dataset gazeta (C:\Users\VoronkovSergey\.cache\huggingface\datasets
\IlyaGusev___gazeta\default\1.0.0\ef9349c3c0f3112ca4036520d76c4bc1b8a79d30bc
29643c6cae5a094d44e457)
```

100% 3/3 [00:00<00:00, 3.54it/s]

In [2]:

```
import pandas as pd
```

In [3]:

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

In [4]:

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

In [5]:

```
data = data.sample(100)
```

In [6]:

```
from transformers import AutoTokenizer, MBartForConditionalGeneration #AutoModel

model_name = "IlyaGusev/mbart_ru_sum_gazeta"

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = MBartForConditionalGeneration.from_pretrained(model_name)
```

In [7]:

```
print(type(tokenizer), type(model))
```

```
<class 'transformers.models.mbart.tokenization_mbart_fast.MBartTokenizerFas
t'> <class 'transformers.models.mbart.modeling_mbart.MBartForConditionalGene
ration'>
```

In [8]:

```
article_text = data['text'].iloc[1]
article_text
```

Out[8]:

'Директор Федеральной службы исполнения наказаний (ФСИН) Александр Реймер рассказал о начале глобального российского тюремного эксперимента: более 153 тыс. заключенных переместили по российским тюрьмам, отделив тех, кто отбывает наказание впервые, от рецидивистов. О перемещениях в местах лишения свободы Реймер сообщил в интервью « Российской газете », которое будет опубликовано в среду, 25 августа. Разделение заключенных сделано для того, чтобы оградить впервые попавших на зону людей «от влияния криминала» и «ограничить развитие уголовных порядков», уточнил Реймер. В законе есть несколько категорий лиц, которых надо держать отдельно, пояснил главный тюремщик страны. Так, начиная со следственного изолятора предусмотрено раздельное содержание несовершеннолетних, женщин, а также тех, кто впервые попал в места лишения свободы. Во время переселений «никаких бунтов, никаких массовых жалоб не было», рассказал Реймер, и не пришлось применять ни спецсредства, ни физическую силу. О результатах перестановок Реймер говорить пока не готов: «Для того чтобы оценить, насколько она (обстановка в колониях — «Газета.Ru») оздоровилась, нужно время. Мы эту работу практически только завершили». Точные сроки начала и конца переселений Реймер не назвал. На момент написания материала ФСИН была не доступна для комментария. Согласно статистике, опубликованной на сайте службы, к 1 августа 2010 года в учреждениях уголовно-исполнительной системы содержались 843,2 тыс. человек. В 2009 году из 724 тыс. взрослых заключенных, содержащихся в исправительных колониях, 377 тыс. отбывали наказание впервые. В марте 2010-го Реймер обещал, что к концу года сортировка заключенных будет закончена. Эксперты разделение преступников на тех, кто впервые попал в колонию, и тех, кто совершил рецидив, поддерживают. По словам члена Общественной палаты Марии Каннабих, подобная изоляция позволяет предотвратить влияние матерых преступников на впервые оказавшихся в тюрьме граждан. «Особенно это касается молодежи», — полагает она. Эксперт подчеркивает, что правозащитное сообщество давно выступало за разделение преступников, однако отмечает, что процесс не может пройти безболезненно. «У меня огромное количество жалоб от заключенных и их родственников на то, что людей переводят с насиженных мест. Это понятно — человек ведь обживается, привыкает, так что процесс очень болезненный», — говорит Каннабих. Официальный представитель правительства в высших судебных инстанциях страны Михаил Барщевский заявил «Газете.Ru», что процесс полностью поддерживает. «Очень хорошо, что слова не расходятся с делом», — сказал он. Разделение «новичков» и рецидивистов — один из этапов многосоставной тюремной реформы, объявленной в 2009 году с назначением Реймера во ФСИН. В перспективе у разделенных может смениться не только компания, но и место отбывания наказания: предполагается, что после реформы ФСИН останутся только тюрьмы (для особо тяжких статей, рецидивистов) и колонии-поселения (для остальных), исправительные колонии общего и строгого режима будут ликвидированы. Также в рамках изменения пенитенциарной системы Минюст разрабатывает законопроект о наказании в виде принудительных работ на специальных предприятиях. В рамках реформы же в марте 2010 года во ФСИН появилось управление собственной безопасности (УСБ), призванное бороться с коррупцией среди тюремщиков. На полную реформу системы исполнения наказаний отведено десять лет.'

In [9]:

```
input_ids = tokenizer.prepare_seq2seq_batch(
    [article_text],
    src_lang="ru_XX",
    return_tensors="pt",
    padding="max_length",
    truncation=True,
    max_length=600
)["input_ids"]

output_ids = model.generate(
    input_ids=input_ids,
    max_length=180,
    no_repeat_ngram_size=3,
    num_beams=5,
    top_k=0
)[0]

summary = tokenizer.decode(output_ids, skip_special_tokens=True, clean_up_tokenization_spac
```

```
c:\program files\python37\lib\site-packages\transformers\tokenization_utils_
base.py:3579: FutureWarning:
`prepare_seq2seq_batch` is deprecated and will be removed in version 5 of Hu
ggingFace Transformers. Use the regular
`__call__` method to prepare your inputs and the tokenizer under the `as_tar
get_tokenizer` context manager to prepare
your targets.

Here is a short example:

model_inputs = tokenizer(src_texts, ...)
with tokenizer.as_target_tokenizer():
    labels = tokenizer(tgt_texts, ...)
model_inputs["labels"] = labels["input_ids"]

See the documentation of your specific tokenizer for more details on the spe
cific arguments to the tokenizer of choice.
For a more complete example, see the implementation of `prepare_seq2seq_batc
h`.

  warnings.warn(formatted_warning, FutureWarning)
```

In [12]:

```
print('original: \n', data['summary'].iloc[1], '\n')
print('predict model: \n', summary)
```

```
original:
 ФСИН провел массовое переселение зэков в российских колониях. Более 150 тыс
яч отбывающих первый срок отделены от рецидивистов. Еще свыше 200 тысяч зэко
в должны быть перемещены, и служба исполнения наказаний будет готова к ликви
дации исправительных колоний, с тем чтобы оставить лишь тюрьмы и поселения.

predict model:
 В России началась глобальная тюремная реформа: более 153 тыс. заключенных п
ереместили по российским тюрьмам, отделив тех, кто отбывает наказание впервы
е, от рецидивистов. Разделение заключенных сделано для того, чтобы оградить
впервые попавших на зону людей «от влияния криминала» и «ограничить развитие
уголовных порядков». О результатах перестановок Реймер говорить пока не гото
в: «Для того чтобы оценить, насколько она (обстановка в колониях — «Газета.R
u») оздоровилась, нужно время».
```

Модель составила пересказ как мне кажется лучше оригинала

In [13]:

```
from nltk.translate.bleu_score import corpus_bleu
from rouge import Rouge

def calc_scores(references, predictions, metric="all"):
    print("Count:", len(predictions))
    print("Ref:", references[-1])
    print("Hyp:", predictions[-1])

    if metric in ("bleu", "all"):
        print("BLEU: ", corpus_bleu([[r] for r in references], predictions))
    if metric in ("rouge", "all"):
        rouge = Rouge()
        scores = rouge.get_scores(predictions, references, avg=True)
        print("ROUGE: ", scores)
```

In [20]:

```
record = data.iloc[1]
```

In [24]:

```
import razdel

def calc_lead_n_score(records, n=3, lower=True, nrows=10):
    references = []
    predictions = []

    for i, record in enumerate(records):
        if i >= nrows:
            break

        input_ids = tokenizer.prepare_seq2seq_batch(
            [data['text'].iloc[47]],
            src_lang="ru_XX",
            return_tensors="pt",
            padding="max_length",
            truncation=True,
            max_length=600
        )["input_ids"]

        output_ids = model.generate(
            input_ids=input_ids,
            max_length=162,
            no_repeat_ngram_size=3,
            num_beams=5,
            top_k=0
        )[0]

        summary = tokenizer.decode(output_ids, skip_special_tokens=True, clean_up_tokenizat

        references.append(summary)

        text = data['summary'].iloc[47]
        text = text if not lower else text.lower()
        sentences = [sentence.text for sentence in razdel.sentenize(text)]
        prediction = " ".join(sentences[:n])
        predictions.append(prediction)

    calc_scores(references, predictions)

calc_lead_n_score(data, n=1)
```

```
c:\program files\python37\lib\site-packages\transformers\tokenization_uti
ls_base.py:3579: FutureWarning:
`prepare_seq2seq_batch` is deprecated and will be removed in version 5 of
HuggingFace Transformers. Use the regular
`__call__` method to prepare your inputs and the tokenizer under the `as_
target_tokenizer` context manager to prepare
your targets.

Here is a short example:

model_inputs = tokenizer(src_texts, ...)
with tokenizer.as_target_tokenizer():
    labels = tokenizer(tgt_texts, ...)
model_inputs["labels"] = labels["input_ids"]

See the documentation of your specific tokenizer for more details on the
```

```
specific arguments to the tokenizer of choice.
For a more complete example, see the implementation of `prepare_seq2seq_b
atch`.

  warnings.warn(formatted_warning, FutureWarning)
```

```
Count: 2
Ref: Украинская артистка Анна Корсун, известная под псевдонимом Maruv, стала
жертвой критики своих соотечественников за «развратное поведение» на выступл
ениях в России. Подписчики певицы в инстаграме осудили ее якобы двусмысленны
е позы на фотографиях с концертов — на некоторых кадрах она стоит на коленях
перед российской публикой.
Hyp: украинскую певицу maruv раскритиковали за «развратное поведение» на выс
туплениях в россии. подписчикам исполнительницы не понравилось, что на многи
х снимках она стоит перед российской публикой на коленях.
BLEU:  0.3950590830610852
ROUGE:  {'rouge-1': {'r': 0.3, 'p': 0.5, 'f': 0.3749999953125}, 'rouge-2':
{'r': 0.20930232558139536, 'p': 0.36, 'f': 0.26470587770328724}, 'rouge-l':
{'r': 0.275, 'p': 0.4583333333333333, 'f': 0.3437499953125}}
```

ВЫВОД

модель замечательно ищет смысл текста, я считаю даже лучше чем оригинал написанный человеком

In [ ]: