Тематическая классификация длинных текстов - TFIDF и LogReg

```
# Если Вы запускаете ноутбук на colab или kaggle,
# выполните следующие строчки, чтобы подгрузить библиотеку dlnlputils:

!git clone https://github.com/Samsung-IT-Academy/stepik-dl-nlp.git &&
pip install -r stepik-dl-nlp/requirements.txt
import sys; sys.path.append('./stepik-dl-nlp')
```

```
Cloning into 'stepik-dl-nlp'...
remote: Enumerating objects: 296, done.ote: Counting objects: 100%
(3/3), done.ote: Compressing objects: 100% (3/3), done.ote: Total 296
(delta 0), reused 1 (delta 0), pack-reused 293ent already satisfied:
scikit-learn in /usr/local/lib/python3.10/dist-packages (from -r
stepik-dl-nlp/requirements.txt (line 1)) (1.2.2)
Collecting spacy-udpipe (from -r stepik-dl-nlp/requirements.txt (line
2))
  Downloading spacy_udpipe-1.0.0-py3-none-any.whl (11 kB)
Collecting pymorphy2 (from -r stepik-dl-nlp/requirements.txt (line 3))
  Downloading pymorphy2-0.9.1-py3-none-any.whl (55 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 55.5/55.5 kB 1.7 MB/s eta
0:00:00
ent already satisfied: torch>=1.2 in /usr/local/lib/python3.10/dist-
packages (from -r stepik-dl-nlp/requirements.txt (line 4))
(2.3.1+cu121)
Requirement already satisfied: matplotlib in
/usr/local/lib/python3.10/dist-packages (from -r
stepik-dl-nlp/requirements.txt (line 5)) (3.7.1)
Collecting ipymarkup (from -r stepik-dl-nlp/requirements.txt (line 6))
  Downloading ipymarkup-0.9.0-py3-none-any.whl (14 kB)
Requirement already satisfied: lxml in /usr/local/lib/python3.10/dist-
packages (from -r stepik-dl-nlp/requirements.txt (line 7)) (4.9.4)
Requirement already satisfied: scipy in
/usr/local/lib/python3.10/dist-packages (from -r
stepik-dl-nlp/requirements.txt (line 8)) (1.11.4)
Requirement already satisfied: pandas in
/usr/local/lib/python3.10/dist-packages (from -r
stepik-dl-nlp/requirements.txt (line 9)) (2.0.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-
packages (from -r stepik-dl-nlp/requirements.txt (line 10)) (4.66.4)
Collecting youtokentome (from -r stepik-dl-nlp/requirements.txt (line
11))
  Downloading youtokentome-1.0.6.tar.gz (86 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 86.7/86.7 kB 6.0 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: seaborn in
```

```
/usr/local/lib/python3.10/dist-packages (from -r
stepik-dl-nlp/requirements.txt (line 12)) (0.13.1)
Requirement already satisfied: ipykernel in
/usr/local/lib/python3.10/dist-packages (from -r
stepik-dl-nlp/requirements.txt (line 13)) (5.5.6)
Requirement already satisfied: ipython in
/usr/local/lib/python3.10/dist-packages (from -r
stepik-dl-nlp/requirements.txt (line 14)) (7.34.0)
Collecting pyconll (from -r stepik-dl-nlp/requirements.txt (line 15))
  Downloading pyconll-3.2.0-py3-none-any.whl (27 kB)
Collecting gensim==3.8.1 (from -r stepik-dl-nlp/requirements.txt (line
16))
  Downloading gensim-3.8.1.tar.gz (23.4 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 23.4/23.4 MB 40.0 MB/s eta
0:00:00
etadata (setup.py) ...  -r stepik-dl-nlp/requirements.txt (line 17))
  Downloading wget-3.2.zip (10 kB)
  Preparing metadata (setup.py) ...  -r stepik-dl-nlp/requirements.txt
(line 18))
  Downloading livelossplot-0.5.3-py3-none-any.whl (30 kB)
Requirement already satisfied: numpy>=1.11.3 in
/usr/local/lib/python3.10/dist-packages (from gensim==3.8.1->-r
stepik-dl-nlp/requirements.txt (line 16)) (1.25.2)
Requirement already satisfied: six>=1.5.0 in
/usr/local/lib/python3.10/dist-packages (from gensim==3.8.1->-r
stepik-dl-nlp/requirements.txt (line 16)) (1.16.0)
Requirement already satisfied: smart_open>=1.8.1 in
/usr/local/lib/python3.10/dist-packages (from gensim==3.8.1->-r
stepik-dl-nlp/requirements.txt (line 16)) (7.0.4)
Requirement already satisfied: bokeh in
/usr/local/lib/python3.10/dist-packages (from livelossplot==0.5.3->-r
stepik-dl-nlp/requirements.txt (line 18)) (3.3.4)
Requirement already satisfied: joblib>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn->-r stepik-
dl-nlp/requirements.txt (line 1)) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn->-r stepik-
dl-nlp/requirements.txt (line 1)) (3.5.0)
Requirement already satisfied: spacy<4.0.0,>=3.0.0 in
/usr/local/lib/python3.10/dist-packages (from spacy-udpipe->-r stepik-
dl-nlp/requirements.txt (line 2)) (3.7.5)
Collecting ufal.udpipe>=1.2.0 (from spacy-udpipe->-r
stepik-dl-nlp/requirements.txt (line 2))
  Downloading ufal.udpipe-1.3.1.1-cp310-cp310-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (936 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 936.8/936.8 kB 59.7 MB/s eta
0:00:00
 pymorphy2->-r stepik-dl-nlp/requirements.txt (line 3))
  Downloading DAWG_Python-0.7.2-py2.py3-none-any.whl (11 kB)
```

```
Collecting pymorphy2-dicts-ru<3.0,>=2.4 (from pymorphy2->-r stepik-dl-
nlp/requirements.txt (line 3))
  Downloading pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-
any.whl (8.2 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.2/8.2 MB 82.8 MB/s eta
0:00:00
 pymorphy2->-r stepik-dl-nlp/requirements.txt (line 3))
  Downloading docopt-0.6.2.tar.gz (25 kB)
  Preparing metadata (setup.py) ... ent already satisfied: filelock in
/usr/local/lib/python3.10/dist-packages (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4)) (3.15.4)
Requirement already satisfied: typing-extensions>=4.8.0 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4)) (4.12.2)
Requirement already satisfied: sympy in
/usr/local/lib/python3.10/dist-packages (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4)) (1.13.0)
Requirement already satisfied: networkx in
/usr/local/lib/python3.10/dist-packages (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4)) (3.3)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4)) (3.1.4)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.10/dist-packages (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4)) (2023.6.0)
Collecting nvidia-cuda-nvrtc-cu12==12.1.105 (from torch>=1.2->-r
stepik-dl-nlp/requirements.txt (line 4))
  Using cached nvidia_cuda_nvrtc_cu12-12.1.105-py3-none-
manylinux1_x86_64.whl (23.7 MB)
Collecting nvidia-cuda-runtime-cu12==12.1.105 (from torch>=1.2->-r
stepik-dl-nlp/requirements.txt (line 4))
  Using cached nvidia_cuda_runtime_cu12-12.1.105-py3-none-
manylinux1_x86_64.whl (823 kB)
Collecting nvidia-cuda-cupti-cu12==12.1.105 (from torch>=1.2->-r
stepik-dl-nlp/requirements.txt (line 4))
  Using cached nvidia_cuda_cupti_cu12-12.1.105-py3-none-
manylinux1_x86_64.whl (14.1 MB)
Collecting nvidia-cudnn-cu12==8.9.2.26 (from torch>=1.2->-r stepik-dl-
nlp/requirements.txt (line 4))
  Using cached nvidia_cudnn_cu12-8.9.2.26-py3-none-
manylinux1_x86_64.whl (731.7 MB)
Collecting nvidia-cublas-cu12==12.1.3.1 (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4))
  Using cached nvidia_cublas_cu12-12.1.3.1-py3-none-
manylinux1_x86_64.whl (410.6 MB)
Collecting nvidia-cufft-cu12==11.0.2.54 (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4))
  Using cached nvidia_cufft_cu12-11.0.2.54-py3-none-
```

```
manylinux1_x86_64.whl (121.6 MB)
Collecting nvidia-curand-cu12==10.3.2.106 (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4))
  Using cached nvidia_curand_cu12-10.3.2.106-py3-none-
manylinux1_x86_64.whl (56.5 MB)
Collecting nvidia-cusolver-cu12==11.4.5.107 (from torch>=1.2->-r
stepik-dl-nlp/requirements.txt (line 4))
  Using cached nvidia_cusolver_cu12-11.4.5.107-py3-none-
manylinux1_x86_64.whl (124.2 MB)
Collecting nvidia-cusparse-cu12==12.1.0.106 (from torch>=1.2->-r
stepik-dl-nlp/requirements.txt (line 4))
  Using cached nvidia_cusparse_cu12-12.1.0.106-py3-none-
manylinux1_x86_64.whl (196.0 MB)
Collecting nvidia-nccl-cu12==2.20.5 (from torch>=1.2->-r stepik-dl-
nlp/requirements.txt (line 4))
  Using cached nvidia_nccl_cu12-2.20.5-py3-none-
manylinux2014_x86_64.whl (176.2 MB)
Collecting nvidia-nvtx-cu12==12.1.105 (from torch>=1.2->-r stepik-dl-
nlp/requirements.txt (line 4))
  Using cached nvidia_nvtx_cu12-12.1.105-py3-none-
manylinux1_x86_64.whl (99 kB)
Requirement already satisfied: triton==2.3.1 in
/usr/local/lib/python3.10/dist-packages (from torch>=1.2->-r stepik-
dl-nlp/requirements.txt (line 4)) (2.3.1)
Collecting nvidia-nvjitlink-cu12 (from nvidia-cusolver-
cu12==11.4.5.107->torch>=1.2->-r stepik-dl-nlp/requirements.txt (line
4))
  Downloading nvidia_nvjitlink_cu12-12.5.82-py3-none-
manylinux2014_x86_64.whl (21.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.3/21.3 MB 56.5 MB/s eta
0:00:00
ent already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->-r stepik-
dl-nlp/requirements.txt (line 5)) (1.2.1)
Requirement already satisfied: cycler>=0.10 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->-r stepik-
dl-nlp/requirements.txt (line 5)) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->-r stepik-
dl-nlp/requirements.txt (line 5)) (4.53.1)
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->-r stepik-
dl-nlp/requirements.txt (line 5)) (1.4.5)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->-r stepik-
dl-nlp/requirements.txt (line 5)) (24.1)
Requirement already satisfied: pillow>=6.2.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->-r stepik-
dl-nlp/requirements.txt (line 5)) (9.4.0)
```

```
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->-r stepik-
dl-nlp/requirements.txt (line 5)) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->-r stepik-
dl-nlp/requirements.txt (line 5)) (2.8.2)
Collecting intervaltree>=3 (from ipymarkup->-r
stepik-dl-nlp/requirements.txt (line 6))
  Downloading intervaltree-3.1.0.tar.gz (32 kB)
  Preparing metadata (setup.py) ... ent already satisfied:
pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas-
>-r stepik-dl-nlp/requirements.txt (line 9)) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in
/usr/local/lib/python3.10/dist-packages (from pandas->-r stepik-dl-
nlp/requirements.txt (line 9)) (2024.1)
Requirement already satisfied: Click>=7.0 in
/usr/local/lib/python3.10/dist-packages (from youtokentome->-r stepik-
dl-nlp/requirements.txt (line 11)) (8.1.7)
Requirement already satisfied: ipython-genutils in
/usr/local/lib/python3.10/dist-packages (from ipykernel->-r stepik-dl-
nlp/requirements.txt (line 13)) (0.2.0)
Requirement already satisfied: traitlets>=4.1.0 in
/usr/local/lib/python3.10/dist-packages (from ipykernel->-r stepik-dl-
nlp/requirements.txt (line 13)) (5.7.1)
Requirement already satisfied: jupyter-client in
/usr/local/lib/python3.10/dist-packages (from ipykernel->-r stepik-dl-
nlp/requirements.txt (line 13)) (6.1.12)
Requirement already satisfied: tornado>=4.2 in
/usr/local/lib/python3.10/dist-packages (from ipykernel->-r stepik-dl-
nlp/requirements.txt (line 13)) (6.3.3)
Requirement already satisfied: setuptools>=18.5 in
/usr/local/lib/python3.10/dist-packages (from ipython->-r stepik-dl-
nlp/requirements.txt (line 14)) (67.7.2)
Collecting jedi>=0.16 (from ipython->-r stepik-dl-nlp/requirements.txt
(line 14))
  Downloading jedi-0.19.1-py2.py3-none-any.whl (1.6 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.6/1.6 MB 69.6 MB/s eta
0:00:00
ent already satisfied: decorator in /usr/local/lib/python3.10/dist-
packages (from ipython->-r stepik-dl-nlp/requirements.txt (line 14))
(4.4.2)
Requirement already satisfied: pickleshare in
/usr/local/lib/python3.10/dist-packages (from ipython->-r stepik-dl-
nlp/requirements.txt (line 14)) (0.7.5)
Requirement already satisfied: prompt-toolkit!=3.0.0,!
=3.0.1,<3.1.0,>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from
ipython->-r stepik-dl-nlp/requirements.txt (line 14)) (3.0.47)
Requirement already satisfied: pygments in
/usr/local/lib/python3.10/dist-packages (from ipython->-r stepik-dl-
```

```
nlp/requirements.txt (line 14)) (2.16.1)
Requirement already satisfied: backcall in
/usr/local/lib/python3.10/dist-packages (from ipython->-r stepik-dl-
nlp/requirements.txt (line 14)) (0.2.0)
Requirement already satisfied: matplotlib-inline in
/usr/local/lib/python3.10/dist-packages (from ipython->-r stepik-dl-
nlp/requirements.txt (line 14)) (0.1.7)
Requirement already satisfied: pexpect>4.3 in
/usr/local/lib/python3.10/dist-packages (from ipython->-r stepik-dl-
nlp/requirements.txt (line 14)) (4.9.0)
Requirement already satisfied: sortedcontainers<3.0,>=2.0 in
/usr/local/lib/python3.10/dist-packages (from intervaltree>=3-
>ipymarkup->-r stepik-dl-nlp/requirements.txt (line 6)) (2.4.0)
Requirement already satisfied: parso<0.9.0,>=0.8.3 in
/usr/local/lib/python3.10/dist-packages (from jedi>=0.16->ipython->-r
stepik-dl-nlp/requirements.txt (line 14)) (0.8.4)
Requirement already satisfied: ptyprocess>=0.5 in
/usr/local/lib/python3.10/dist-packages (from pexpect>4.3->ipython->-r
stepik-dl-nlp/requirements.txt (line 14)) (0.7.0)
Requirement already satisfied: wcwidth in
/usr/local/lib/python3.10/dist-packages (from prompt-toolkit!=3.0.0,!
=3.0.1,<3.1.0,>=2.0.0->ipython->-r stepik-dl-nlp/requirements.txt
(line 14)) (0.2.13)
Requirement already satisfied: wrapt in
/usr/local/lib/python3.10/dist-packages (from smart_open>=1.8.1-
>gensim==3.8.1->-r stepik-dl-nlp/requirements.txt (line 16)) (1.14.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (3.0.9)
Requirement already satisfied: thinc<8.3.0,>=8.2.2 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (8.2.5)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
```

```
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (2.0.10)
Requirement already satisfied: weasel<0.5.0,>=0.1.0 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (0.4.1)
Requirement already satisfied: typer<1.0.0,>=0.3.0 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (0.12.3)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (2.31.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (2.8.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in
/usr/local/lib/python3.10/dist-packages (from spacy<4.0.0,>=3.0.0-
>spacy-udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (3.4.0)
Requirement already satisfied: PyYAML>=3.10 in
/usr/local/lib/python3.10/dist-packages (from bokeh-
>livelossplot==0.5.3->-r stepik-dl-nlp/requirements.txt (line 18))
(6.0.1)
Requirement already satisfied: xyzservices>=2021.09.1 in
/usr/local/lib/python3.10/dist-packages (from bokeh-
>livelossplot==0.5.3->-r stepik-dl-nlp/requirements.txt (line 18))
(2024.6.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->torch>=1.2->-r
stepik-dl-nlp/requirements.txt (line 4)) (2.1.5)
Requirement already satisfied: jupyter-core>=4.6.0 in
/usr/local/lib/python3.10/dist-packages (from jupyter-client-
>ipykernel->-r stepik-dl-nlp/requirements.txt (line 13)) (5.7.2)
Requirement already satisfied: pyzmq>=13 in
/usr/local/lib/python3.10/dist-packages (from jupyter-client-
>ipykernel->-r stepik-dl-nlp/requirements.txt (line 13)) (24.0.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.2->-r
stepik-dl-nlp/requirements.txt (line 4)) (1.3.0)
Requirement already satisfied: platformdirs>=2.5 in
/usr/local/lib/python3.10/dist-packages (from jupyter-core>=4.6.0-
>jupyter-client->ipykernel->-r stepik-dl-nlp/requirements.txt (line
13)) (4.2.2)
Requirement already satisfied: language-data>=1.2 in
/usr/local/lib/python3.10/dist-packages (from langcodes<4.0.0,>=3.2.0-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (1.2.0)
Requirement already satisfied: annotated-types>=0.4.0 in
/usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!
```

```
=1.8.1,<3.0.0,>=1.7.4->spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-
dl-nlp/requirements.txt (line 2)) (0.7.0)
Requirement already satisfied: pydantic-core==2.20.1 in
/usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!
=1.8.1,<3.0.0,>=1.7.4->spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-
dl-nlp/requirements.txt (line 2)) (2.20.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (2024.7.4)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in
/usr/local/lib/python3.10/dist-packages (from thinc<8.3.0,>=8.2.2-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in
/usr/local/lib/python3.10/dist-packages (from thinc<8.3.0,>=8.2.2-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (0.1.5)
Requirement already satisfied: shellingham>=1.3.0 in
/usr/local/lib/python3.10/dist-packages (from typer<1.0.0,>=0.3.0-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (1.5.4)
Requirement already satisfied: rich>=10.11.0 in
/usr/local/lib/python3.10/dist-packages (from typer<1.0.0,>=0.3.0-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (13.7.1)
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in
/usr/local/lib/python3.10/dist-packages (from weasel<0.5.0,>=0.1.0-
>spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-nlp/requirements.txt
(line 2)) (0.18.1)
Requirement already satisfied: marisa-trie>=0.7.7 in
/usr/local/lib/python3.10/dist-packages (from language-data>=1.2-
>langcodes<4.0.0,>=3.2.0->spacy<4.0.0,>=3.0.0->spacy-udpipe->-r
stepik-dl-nlp/requirements.txt (line 2)) (1.2.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.10/dist-packages (from rich>=10.11.0-
>typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.0.0->spacy-udpipe->-r stepik-dl-
```

```
nlp/requirements.txt (line 2)) (3.0.0)
Requirement already satisfied: mdurl~=0.1 in
/usr/local/lib/python3.10/dist-packages (from markdown-it-py>=2.2.0-
>rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.0.0->spacy-
udpipe->-r stepik-dl-nlp/requirements.txt (line 2)) (0.1.2)
Building wheels for collected packages: gensim, youtokentome, wget,
docopt, intervaltree
  Building wheel for gensim (setup.py) ... : filename=gensim-3.8.1-
cp310-cp310-linux_x86_64.whl size=24682929
sha256=22c4d1a57fc8b124a673c85e89478a7e30ff7379c8b82547b03a78b13dfbb9c
0
  Stored in directory:
/root/.cache/pip/wheels/92/23/5d/b5ce54b3760acfebee170a8fe4d91cb303faf
befd8f93f3723
  Building wheel for youtokentome (setup.py) ... e:
filename=youtokentome-1.0.6-cp310-cp310-linux_x86_64.whl size=1951499
sha256=3c830b76c4048d00b28602a704a0c8c33ba1b10560dbacf01dd2e105ac2e839
6
  Stored in directory:
/root/.cache/pip/wheels/df/85/f8/301d2ba45f43f30bed2fe413efa760bc726b8
b660ed9c2900c
  Building wheel for wget (setup.py) ... e=wget-3.2-py3-none-any.whl
size=9656
sha256=ae079c5d788152a12277de2ae637a1224065b54bc5fb809f5079d38cda1f435
9
  Stored in directory:
/root/.cache/pip/wheels/8b/f1/7f/5c94f0a7a505ca1c81cd1d9208ae2064675d9
7582078e6c769
  Building wheel for docopt (setup.py) ... e=docopt-0.6.2-py2.py3-
none-any.whl size=13706
sha256=667320b33d4149e31bb413c64be64b2fc6897f070321f9caa654116bc48bb1b
9
  Stored in directory:
/root/.cache/pip/wheels/fc/ab/d4/5da2067ac95b36618c629a5f93f8094257005
06f72c9732fac
  Building wheel for intervaltree (setup.py) ... e=intervaltree-3.1.0-
py2.py3-none-any.whl size=26096
sha256=abb9ddfb25cbe71906b2f00faf1e9000693bdd9aed7eef22cf20f4fb6a3e6d2
4
  Stored in directory:
/root/.cache/pip/wheels/fa/80/8c/43488a924a046b733b64de3fac99252674c89
2a4c3801c0a61
Successfully built gensim youtokentome wget docopt intervaltree
Installing collected packages: wget, ufal.udpipe, pymorphy2-dicts-ru,
docopt, dawg-python, youtokentome, pymorphy2, pyconll, nvidia-nvtx-
cu12, nvidia-nvjitlink-cu12, nvidia-nccl-cu12, nvidia-curand-cu12,
nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12,
nvidia-cuda-cupti-cu12, nvidia-cublas-cu12, jedi, intervaltree,
nvidia-cusparse-cu12, nvidia-cudnn-cu12, ipymarkup, gensim, nvidia-
```

```
cusolver-cu12, livelossplot, spacy-udpipe
  Attempting uninstall: gensim
    Found existing installation: gensim 4.3.2
    Uninstalling gensim-4.3.2:
      Successfully uninstalled gensim-4.3.2
Successfully installed dawg-python-0.7.2 docopt-0.6.2 gensim-3.8.1
intervaltree-3.1.0 ipymarkup-0.9.0 jedi-0.19.1 livelossplot-0.5.3
nvidia-cublas-cu12-12.1.3.1 nvidia-cuda-cupti-cu12-12.1.105 nvidia-
cuda-nvrtc-cu12-12.1.105 nvidia-cuda-runtime-cu12-12.1.105 nvidia-
cudnn-cu12-8.9.2.26 nvidia-cufft-cu12-11.0.2.54 nvidia-curand-cu12-
10.3.2.106 nvidia-cusolver-cu12-11.4.5.107 nvidia-cusparse-cu12-
12.1.0.106 nvidia-nccl-cu12-2.20.5 nvidia-nvjitlink-cu12-12.5.82
nvidia-nvtx-cu12-12.1.105 pyconll-3.2.0 pymorphy2-0.9.1 pymorphy2-
dicts-ru-2.4.417127.4579844 spacy-udpipe-1.0.0 ufal.udpipe-1.3.1.1
wget-3.2 youtokentome-1.0.6
```

```
import warnings
warnings.filterwarnings('ignore')

from sklearn.datasets import fetch_20newsgroups
from sklearn.metrics import accuracy_score

import numpy as np

import matplotlib.pyplot as plt
%matplotlib inline

import collections

import torch
from torch import nn
from torch.nn import functional as F

import dlnlputils
from dlnlputils.data import tokenize_text_simple_regex,
tokenize_corpus, build_vocabulary, \
    vectorize_texts, SparseFeaturesDataset
from dlnlputils.pipeline import train_eval_loop, predict_with_model,
init_random_seed

init_random_seed()

import re
from collections import Counter
# коллекция
c  = [
'Казнить нельзя, помиловать. Нельзя наказывать.',
'Казнить, нельзя помиловать. Нельзя освободить.',
'Нельзя не помиловать.',
'Обязательно освободить.'
```

```
]

# токенизация + приведение к нижнему регистру
TOKENIZE_RE = re.compile(r'[\w\d]+', re.I)
c = [x.lower() for x in c]
def tokenize(txt):
   return TOKENIZE_RE.findall(txt)
c_tokens = [tokenize(x) for x in c]
print("c_tokens_____________________________________")
print(c_tokens)
print('-----------------------------------------')


# встроенная функция для формирования словаря и подсчета частотности
MAX_DF = 1
MIN_COUNT = 1
vocabulary, word_doc_freq = build_vocabulary(c_tokens,
max_doc_freq=MAX_DF, min_count=MIN_COUNT)
UNIQUE_WORDS_N = len(vocabulary)

# из словаря и частотности получаем пары
word_df = [(word, "{:.2f}".format(word_doc_freq[i])) for i, (word, _)
in enumerate(vocabulary.items())]
print('0000', word_df)
answer = sorted(word_df, key=lambda x: (x[1], x[0]), reverse=False)
print('1111', answer)


answer_1 = []
answer_2 = []
for k, v in list(answer):
  answer_1.append(k)
  answer_2.append(str(v))

print(" ".join(answer_1))
print(" ".join(answer_2))
```

```
c_tokens_____________________________________
[['казнить', 'нельзя', 'помиловать', 'нельзя', 'наказывать'],
['казнить', 'нельзя', 'помиловать', 'нельзя', 'освободить'],
['нельзя', 'не', 'помиловать'], ['обязательно', 'освободить']]
-----------------------------------------
0000 [('помиловать', '0.75'), ('нельзя', '0.75'), ('казнить', '0.50'),
('освободить', '0.50'), ('наказывать', '0.25'), ('не', '0.25'),
('обязательно', '0.25')]
1111 [('наказывать', '0.25'), ('не', '0.25'), ('обязательно', '0.25'),
('казнить', '0.50'), ('освободить', '0.50'), ('нельзя', '0.75'),
('помиловать', '0.75')]
наказывать не обязательно казнить освободить нельзя помиловать
0.25 0.25 0.25 0.50 0.50 0.75 0.75
```

```
# import re

# c = [
#     'Казнить нельзя, помиловать. Нельзя наказывать.',
#     'Казнить, нельзя помиловать. Нельзя освободить.',
#     'Нельзя не помиловать.',
#     'Обязательно освободить.'
# ]

# def build_vocabulary(tokenized_texts, max_size=1000000,
max_doc_freq=1, min_count=1, pad_word=None):
#     word_counts = collections.defaultdict(int)
#     doc_n = 0

#     # посчитать количество документов, в которых употребляется
каждое слово
#     # а также общее количество документов
#     for txt in tokenized_texts:
#         doc_n += 1
#         unique_text_tokens = set(txt)
#         for token in unique_text_tokens:
#             word_counts[token] += 1

#     # убрать слишком редкие и слишком частые слова
#     word_counts = {word: cnt for word, cnt in word_counts.items()
#                    if cnt >= min_count and cnt / doc_n <=
max_doc_freq}

#     # отсортировать слова по убыванию частоты
#     sorted_word_counts = sorted(word_counts.items(),
#                                 reverse=False,
#                                 key=lambda pair: pair[1])

#     # добавим несуществующее слово с индексом 0 для удобства
пакетной обработки
#     if pad_word is not None:
#         sorted_word_counts = [(pad_word, 0)] + sorted_word_counts

#     # если у нас по прежнему слишком много слов, оставить только
max_size самых частотных
#     if len(word_counts) > max_size:
#         sorted_word_counts = sorted_word_counts[:max_size]

#     # нумеруем слова
#     word2id = {word: i for i, (word, _) in
enumerate(sorted_word_counts)}

#     # нормируем частоты слов
#     word2freq = np.array([cnt / doc_n for _, cnt in
sorted_word_counts], dtype='float32')
```

```
#     return word2id, word2freq

# c = [x.lower() for x in c]

# TOKENIZE_RE = re.compile(r'[\w\d]+')
# def tokenize(txt):
#     return TOKENIZE_RE.findall(txt)

# c_tokens = [tokenize(x) for x in c]
# MAX_DF = 1
# MIN_COUNT = 1
# vocabulary, word_doc_freq = build_vocabulary(c_tokens,
max_doc_freq=MAX_DF, min_count=MIN_COUNT)
# UNIQUE_WORDS_N = len(vocabulary)
# word_df = [(word, "{:.2f}".format(word_doc_freq[i])) for i, (word,
_) in enumerate(vocabulary.items())]
# answer = sorted(word_df, key=lambda x: (float(x[1]), x[0]),
reverse=False)
# answer_1 = []
# answer_2 = []
# for k, v in answer:
#     answer_1.append(k)
#     answer_2.append(str(v))
# print(" ".join(answer_1))
# print(" ".join(answer_2))
```

```
наказывать не обязательно казнить освободить нельзя помиловать
0.25 0.25 0.25 0.50 0.50 0.75 0.75
```

```
from sklearn.feature_extraction.text import
CountVectorizer,TfidfVectorizer
from sklearn.preprocessing import StandardScaler
import numpy as np

corpus = [
    'Казнить нельзя, помиловать. Нельзя наказывать.',
    'Казнить, нельзя помиловать. Нельзя освободить.',
    'Нельзя не помиловать.',
    'Обязательно освободить.']

#Получаем счетчики слов
TF = CountVectorizer().fit_transform(corpus)

#Строим IDF. К сожалению, в этом задании нам нужно только
vectorizer.idf_
#Для стандартных случаев на этой строке все вычисления и
заканчиваются.
#Обычно  TFIDF = vectorizer.fit_transform(corpus)
vectorizer = TfidfVectorizer(smooth_idf=False, use_idf=True)
vectorizer.fit_transform(corpus)
```

```
## из IDF  в DF
word_doc_freq = 1/np.exp(vectorizer.idf_ - 1)

#TF нормируем и сглаживаем логарифмом (требование задания)
print(TF.toarray())

TFIDF = np.log(np.array(TF/TF.sum(axis=1) + 1)) * vectorizer.idf_

#Масштабируем признаки
scaledTFIDF = StandardScaler().fit_transform(TFIDF)

#Домножаем на np.sqrt((4-1)/4) для перевода из DDOF(0) в DDOF(1) для 4 текстов
#(требование задания)
scaledTFIDF *= np.sqrt(3/4)

#Вывод в порядке возрастания DF
for l in scaledTFIDF[:,np.argsort(word_doc_freq)]:
    print (" ".join([ "%.2f" % d for d in l]))
```

```
[[1 1 0 2 0 0 1]
 [1 0 0 2 0 1 1]
 [0 0 1 1 0 0 1]
 [0 0 0 0 1 1 0]]
```

```
---------------------------------------------------------------------------
NotImplementedError                       Traceback (most recent call last)
<ipython-input-11-e06b549392f0> in <cell line: 26>()
     24 print(TF.toarray())
     25
---> 26 TFIDF = np.log(np.array(TF/TF.sum(axis=1) + 1)) * vectorizer.idf_
     27
     28 #Масштабируем признаки

/usr/local/lib/python3.10/dist-packages/scipy/sparse/_base.py in __add__(self, other)
    460                 return self.copy()
    461             # Now we would add this scalar to every element.
--> 462             raise NotImplementedError('adding a nonzero scalar to a '
    463                                       'sparse array is not supported')
    464         elif issparse(other):

NotImplementedError: adding a nonzero scalar to a sparse array is not supported
```

Предобработка текстов и подготовка признаков

```
train_source = fetch_20newsgroups(subset='train')
test_source = fetch_20newsgroups(subset='test')

print('Количество обучающих текстов', len(train_source['data']))
print('Количество тестовых текстов', len(test_source['data']))
print()
print(train_source['data'][0].strip())

print()
print('Метка', train_source['target'][0])
```

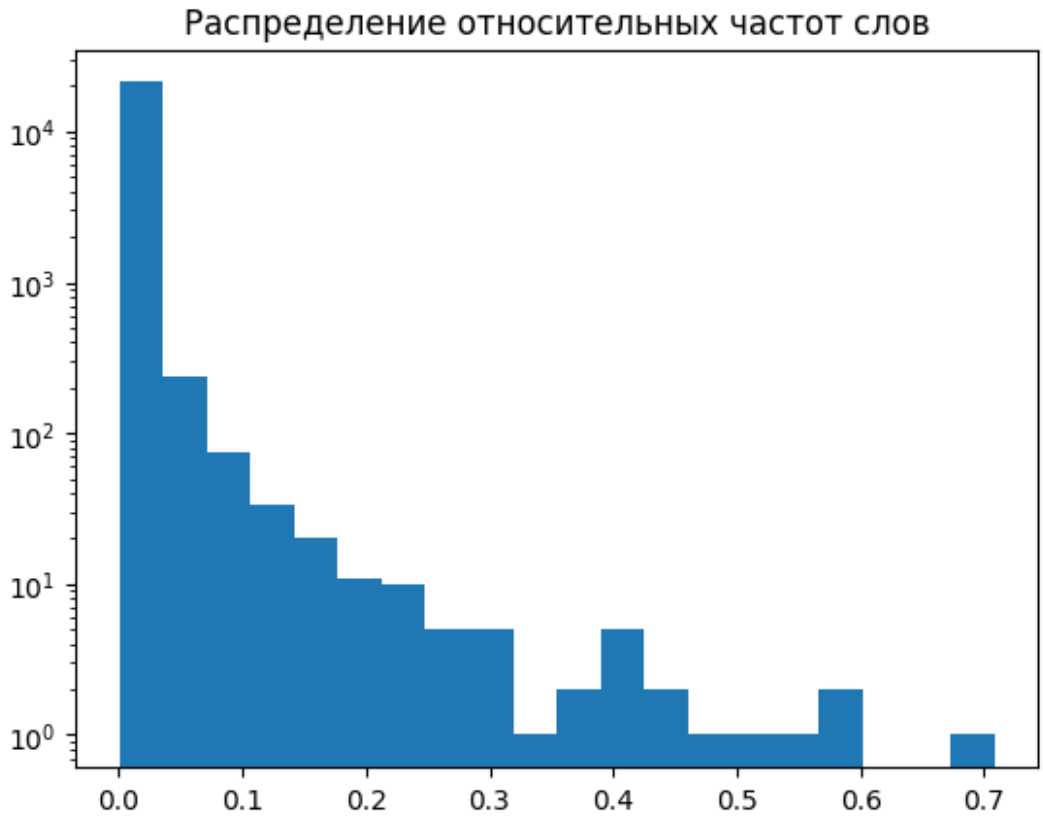
```
Количество обучающих текстов 11314
Количество тестовых текстов 7532

From: lerxst@wam.umd.edu (where's my thing)
Subject: WHAT car is this!?
Nntp-Posting-Host: rac3.wam.umd.edu
Organization: University of Maryland, College Park
Lines: 15

 I was wondering if anyone out there could enlighten me on this car I
saw
the other day. It was a 2-door sports car, looked to be from the late
60s/
early 70s. It was called a Bricklin. The doors were really small. In
addition,
the front bumper was separate from the rest of the body. This is
all I know. If anyone can tellme a model name, engine specs, years
of production, where this car is made, history, or whatever info you
have on this funky looking car, please e-mail.

Thanks,
- IL
   ---- brought to you by your neighborhood Lerxst ----

Метка 7
```

Подготовка признаков

```
train_tokenized = tokenize_corpus(train_source['data'])
test_tokenized = tokenize_corpus(test_source['data'])

print(' '.join(train_tokenized[0]))
```

```
from lerxst where thing subject what this nntp posting host rac3
organization university maryland college park lines wondering anyone
there could enlighten this other door sports looked from late early
called bricklin doors were really small addition front bumper separate
from rest body this know anyone tellme model name engine specs years
```

```
production where this made history whatever info have this funky
looking please mail thanks brought your neighborhood lerxst

MAX_DF = 0.8
MIN_COUNT = 5
vocabulary, word_doc_freq = build_vocabulary(train_tokenized,
max_doc_freq=MAX_DF, min_count=MIN_COUNT)
UNIQUE_WORDS_N = len(vocabulary)
print('Количество уникальных токенов', UNIQUE_WORDS_N)
print(list(vocabulary.items())[:10])

Количество уникальных токенов 21628
[('that', 0), ('this', 1), ('have', 2), ('with', 3), ('writes', 4),
('article', 5), ('posting', 6), ('host', 7), ('nntp', 8), ('there',
9)]

plt.hist(word_doc_freq, bins=20)
plt.title('Распределение относительных частот слов')
plt.yscale('log');
```

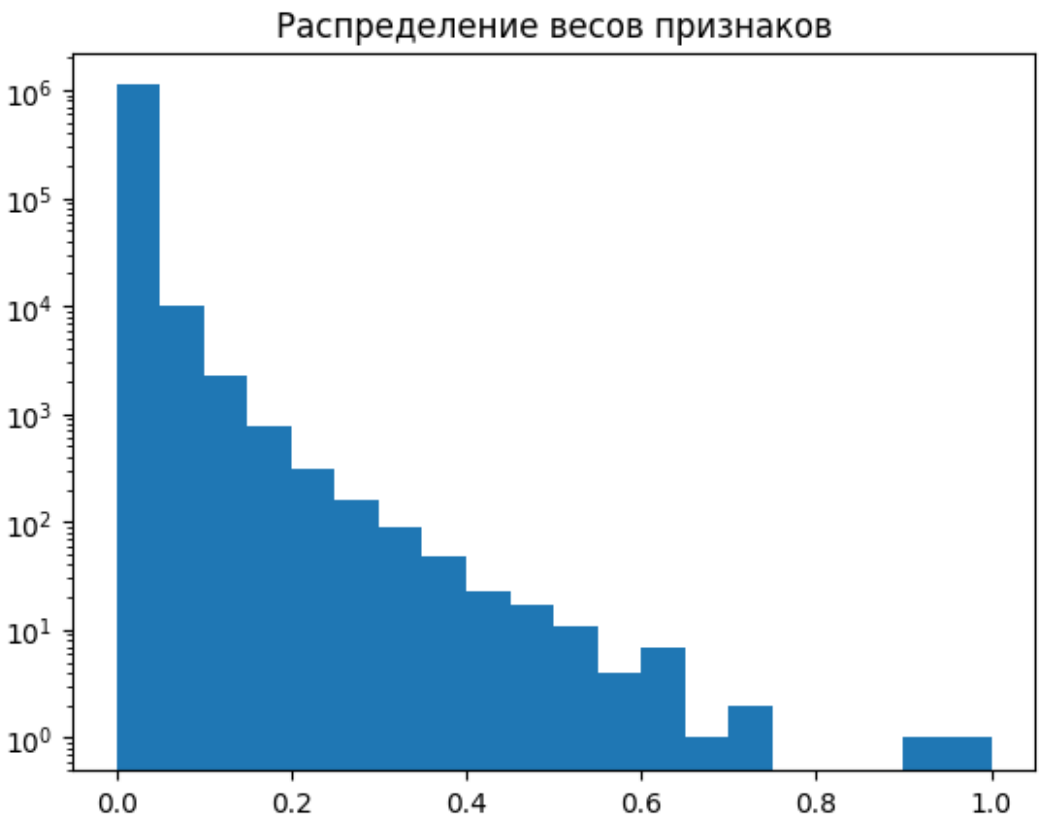



```
VECTORIZATION_MODE = 'tfidf'
train_vectors = vectorize_texts(train_tokenized, vocabulary,
word_doc_freq, mode=VECTORIZATION_MODE)
test_vectors = vectorize_texts(test_tokenized, vocabulary,
```

```
word_doc_freq, mode=VECTORIZATION_MODE)

print('Размерность матрицы признаков обучающей выборки',
train_vectors.shape)
print('Размерность матрицы признаков тестовой выборки',
test_vectors.shape)
print()
print('Количество ненулевых элементов в обучающей выборке',
train_vectors.nnz)
print('Процент заполненности матрицы признаков {:.2f}
%'.format(train_vectors.nnz * 100 / (train_vectors.shape[0] *
train_vectors.shape[1])))
print()
print('Количество ненулевых элементов в тестовой выборке',
test_vectors.nnz)
print('Процент заполненности матрицы признаков {:.2f}
%'.format(test_vectors.nnz * 100 / (test_vectors.shape[0] *
test_vectors.shape[1])))
```

```
Размерность матрицы признаков обучающей выборки (11314, 21628)
Размерность матрицы признаков тестовой выборки (7532, 21628)

Количество ненулевых элементов в обучающей выборке 1126792
Процент заполненности матрицы признаков 0.46%

Количество ненулевых элементов в тестовой выборке 721529
Процент заполненности матрицы признаков 0.44%
```

```
plt.hist(train_vectors.data, bins=20)
plt.title('Распределение весов признаков')
plt.yscale('log');
```

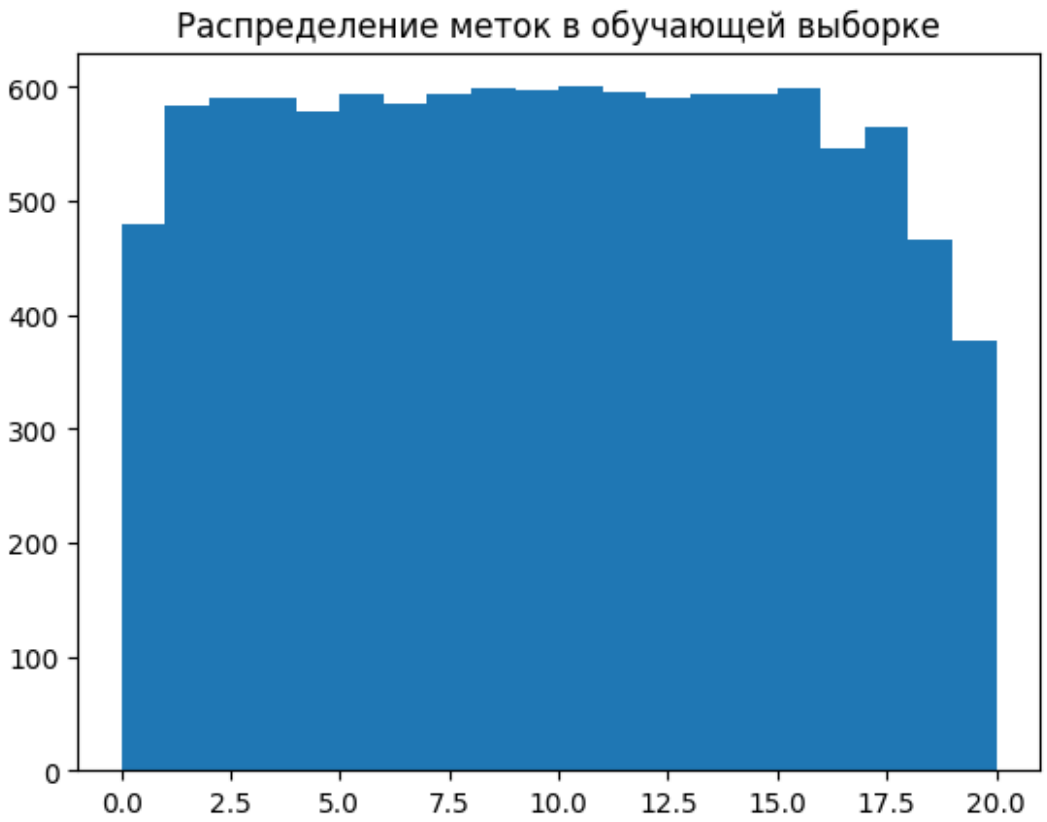


Распределение классов

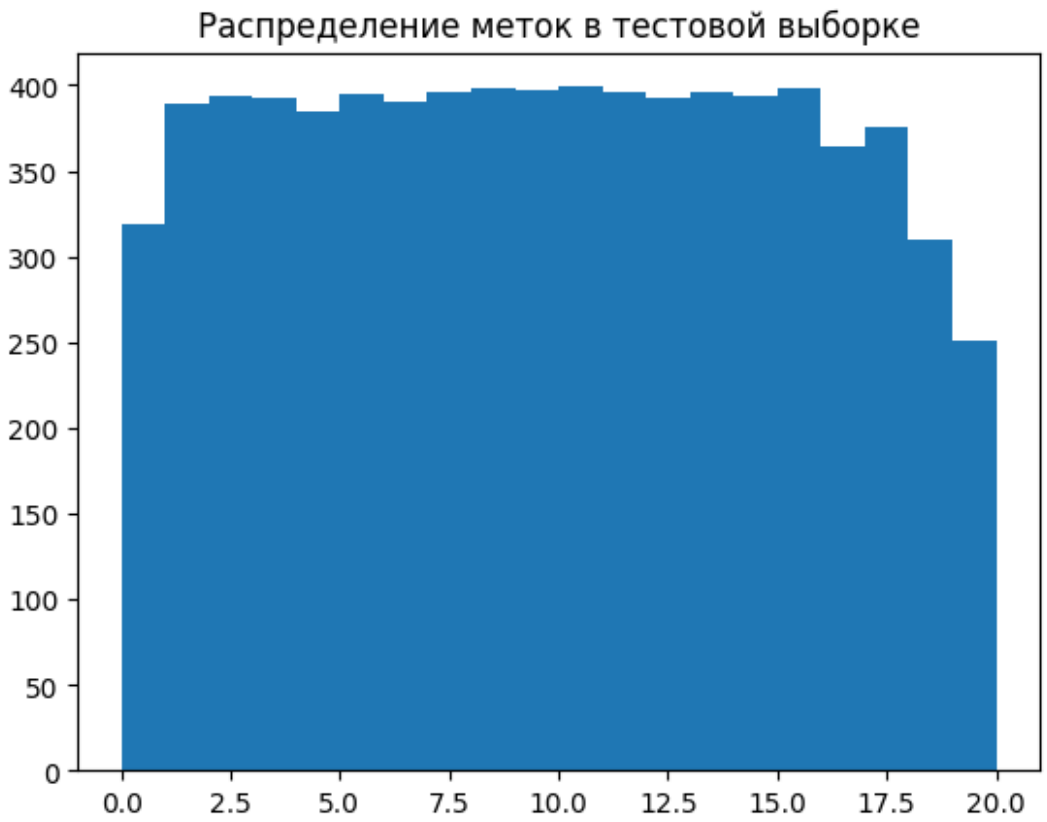
```
UNIQUE_LABELS_N = len(set(train_source['target']))
print('Количество уникальных меток', UNIQUE_LABELS_N)
```

```
Количество уникальных меток 20
```

```
plt.hist(train_source['target'], bins=np.arange(0, 21))
plt.title('Распределение меток в обучающей выборке');
```




```
plt.hist(test_source['target'], bins=np.arange(0, 21))
plt.title('Распределение меток в тестовой выборке');
```



PyTorch Dataset

```
train_dataset = SparseFeaturesDataset(train_vectors,
train_source['target'])
test_dataset = SparseFeaturesDataset(test_vectors,
test_source['target'])
```

Обучение модели на PyTorch

```
model = nn.Linear(UNIQUE_WORDS_N, UNIQUE_LABELS_N)

scheduler = lambda optim: \
    torch.optim.lr_scheduler.ReduceLROnPlateau(optim, patience=5,
factor=0.5, verbose=True)

best_val_loss, best_model = train_eval_loop(model=model,

train_dataset=train_dataset,
                                            val_dataset=test_dataset,
                                            criterion=F.cross_entropy,
                                            lr=1e-1,
                                            epoch_n=200,
                                            batch_size=32,
                                            l2_reg_alpha=0,
```

```
lr_scheduler_ctor=scheduler)

Эпоха 0
Эпоха: 354 итераций, 5.56 сек
Среднее значение функции потерь на обучении 2.2252588447204418
Среднее значение функции потерь на валидации 2.113699140690141
Новая лучшая модель!

Эпоха 1
Эпоха: 354 итераций, 3.97 сек
Среднее значение функции потерь на обучении 0.9165915277718151
Среднее значение функции потерь на валидации 1.6821942723403542
Новая лучшая модель!

Эпоха 2
Эпоха: 354 итераций, 4.46 сек
Среднее значение функции потерь на обучении 0.4660174340683188
Среднее значение функции потерь на валидации 1.464388514474287
Новая лучшая модель!

Эпоха 3
Эпоха: 354 итераций, 3.96 сек
Среднее значение функции потерь на обучении 0.28363614996611064
Среднее значение функции потерь на валидации 1.3450368306899474
Новая лучшая модель!

Эпоха 4
Эпоха: 354 итераций, 4.05 сек
Среднее значение функции потерь на обучении 0.1907244601660529
Среднее значение функции потерь на валидации 1.2602406106257842
Новая лучшая модель!

Эпоха 5
Эпоха: 354 итераций, 4.42 сек
Среднее значение функции потерь на обучении 0.13643471676708951
Среднее значение функции потерь на валидации 1.2005351966215392
Новая лучшая модель!

Эпоха 6
Эпоха: 354 итераций, 3.92 сек
Среднее значение функции потерь на обучении 0.10175162716900225
Среднее значение функции потерь на валидации 1.1543818995103998
Новая лучшая модель!

Эпоха 7
Эпоха: 354 итераций, 5.07 сек
Среднее значение функции потерь на обучении 0.07793642956721412
Среднее значение функции потерь на валидации 1.1209699140261795
Новая лучшая модель!
```

```
Эпоха 8
Эпоха: 354 итераций, 4.00 сек
Среднее значение функции потерь на обучении 0.061769798013913094
Среднее значение функции потерь на валидации 1.0886663978887816
Новая лучшая модель!

Эпоха 9
Эпоха: 354 итераций, 4.85 сек
Среднее значение функции потерь на обучении 0.04921774062627958
Среднее значение функции потерь на валидации 1.0575762226925058
Новая лучшая модель!

Эпоха 10
Эпоха: 354 итераций, 4.28 сек
Среднее значение функции потерь на обучении 0.04008678682043222
Среднее значение функции потерь на валидации 1.0522257650302629
Новая лучшая модель!

Эпоха 11
Эпоха: 354 итераций, 5.03 сек
Среднее значение функции потерь на обучении 0.03305450567481039
Среднее значение функции потерь на валидации 1.0252110523692632
Новая лучшая модель!

Эпоха 12
Эпоха: 354 итераций, 3.91 сек
Среднее значение функции потерь на обучении 0.027422088429878998
Среднее значение функции потерь на валидации 1.0233956320306001
Новая лучшая модель!

Эпоха 13
Эпоха: 354 итераций, 4.48 сек
Среднее значение функции потерь на обучении 0.023065803551488677
Среднее значение функции потерь на валидации 1.0005402216466808
Новая лучшая модель!

Эпоха 14
Эпоха: 354 итераций, 3.98 сек
Среднее значение функции потерь на обучении 0.019530126435608513
Среднее значение функции потерь на валидации 0.9915590105673014
Новая лучшая модель!

Эпоха 15
Эпоха: 354 итераций, 4.19 сек
Среднее значение функции потерь на обучении 0.016497758172431404
Среднее значение функции потерь на валидации 0.9758748836436514
Новая лучшая модель!

Эпоха 16
```

```
Эпоха: 354 итераций, 4.19 сек
Среднее значение функции потерь на обучении 0.01395973838447451
Среднее значение функции потерь на валидации 0.9762108462341761

Эпоха 17
Эпоха: 354 итераций, 3.94 сек
Среднее значение функции потерь на обучении 0.012367766882926963
Среднее значение функции потерь на валидации 0.9614278031355243
Новая лучшая модель!

Эпоха 18
Эпоха: 354 итераций, 4.81 сек
Среднее значение функции потерь на обучении 0.010551420524505908
Среднее значение функции потерь на валидации 0.9536916459515944
Новая лучшая модель!

Эпоха 19
Эпоха: 354 итераций, 4.00 сек
Среднее значение функции потерь на обучении 0.009276754325413602
Среднее значение функции потерь на валидации 0.9553852299765005

Эпоха 20
Эпоха: 354 итераций, 5.06 сек
Среднее значение функции потерь на обучении 0.008060229120060465
Среднее значение функции потерь на валидации 0.9670787574881214

Эпоха 21
Эпоха: 354 итераций, 4.01 сек
Среднее значение функции потерь на обучении 0.007328641356895073
Среднее значение функции потерь на валидации 0.9403575849482568
Новая лучшая модель!

Эпоха 22
Эпоха: 354 итераций, 5.04 сек
Среднее значение функции потерь на обучении 0.006772741428076548
Среднее значение функции потерь на валидации 0.9329051228903108
Новая лучшая модель!

Эпоха 23
Эпоха: 354 итераций, 3.97 сек
Среднее значение функции потерь на обучении 0.005589817616415057
Среднее значение функции потерь на валидации 0.932979334966611

Эпоха 24
Эпоха: 354 итераций, 4.66 сек
Среднее значение функции потерь на обучении 0.005256614731993042
Среднее значение функции потерь на валидации 0.9690085734098645

Эпоха 25
Эпоха: 354 итераций, 4.07 сек
```

```
Среднее значение функции потерь на обучении 0.004769208558702231
Среднее значение функции потерь на валидации 0.9437574386849241

Эпоха 26
Эпоха: 354 итераций, 4.19 сек
Среднее значение функции потерь на обучении 0.004408718667704728
Среднее значение функции потерь на валидации 0.9343373862616087

Эпоха 27
Эпоха: 354 итераций, 4.53 сек
Среднее значение функции потерь на обучении 0.004121565427381208
Среднее значение функции потерь на валидации 0.9609629836375431

Эпоха 28
Эпоха: 354 итераций, 3.98 сек
Среднее значение функции потерь на обучении 0.004164909231552123
Среднее значение функции потерь на валидации 0.9498716054831521

Эпоха 29
Эпоха: 354 итераций, 4.60 сек
Среднее значение функции потерь на обучении 0.0032320540977311767
Среднее значение функции потерь на валидации 0.9344251049777209

Эпоха 30
Эпоха: 354 итераций, 4.02 сек
Среднее значение функции потерь на обучении 0.002957674951176159
Среднее значение функции потерь на валидации 0.9424987504795447

Эпоха 31
Эпоха: 354 итераций, 5.02 сек
Среднее значение функции потерь на обучении 0.0028751842024154687
Среднее значение функции потерь на валидации 0.934858668911255

Эпоха 32
Эпоха: 354 итераций, 3.93 сек
Среднее значение функции потерь на обучении 0.0026018112765672473
Среднее значение функции потерь на валидации 0.9306279965123888
Новая лучшая модель!

Эпоха 33
Эпоха: 354 итераций, 4.85 сек
Среднее значение функции потерь на обучении 0.002713252267718715
Среднее значение функции потерь на валидации 0.9447229037345466

Эпоха 34
Эпоха: 354 итераций, 3.99 сек
Среднее значение функции потерь на обучении 0.002534759648395492
Среднее значение функции потерь на валидации 0.9425551963307089

Эпоха 35
```

```
Эпоха: 354 итераций, 4.18 сек
Среднее значение функции потерь на обучении 0.0026106528503643373
Среднее значение функции потерь на валидации 0.9399326563639155

Эпоха 36
Эпоха: 354 итераций, 5.24 сек
Среднее значение функции потерь на обучении 0.002410933136608515
Среднее значение функции потерь на валидации 0.938659789324817

Эпоха 37
Эпоха: 354 итераций, 4.51 сек
Среднее значение функции потерь на обучении 0.0023538285292853405
Среднее значение функции потерь на валидации 0.9418990392301042

Эпоха 38
Эпоха: 354 итераций, 4.03 сек
Среднее значение функции потерь на обучении 0.0023111264638771764
Среднее значение функции потерь на валидации 0.9422686339940055

Эпоха 39
Эпоха: 354 итераций, 3.95 сек
Среднее значение функции потерь на обучении 0.0019421890409815252
Среднее значение функции потерь на валидации 0.9435674703474772

Эпоха 40
Эпоха: 354 итераций, 4.44 сек
Среднее значение функции потерь на обучении 0.0018815712933066843
Среднее значение функции потерь на валидации 0.9396190942596581

Эпоха 41
Эпоха: 354 итераций, 3.90 сек
Среднее значение функции потерь на обучении 0.0018609663672428868
Среднее значение функции потерь на валидации 0.944632682128478

Эпоха 42
Эпоха: 354 итераций, 5.30 сек
Среднее значение функции потерь на обучении 0.0018222677341111842
Среднее значение функции потерь на валидации 0.9422169335312762

Эпоха 43
Эпоха: 354 итераций, 4.39 сек
Среднее значение функции потерь на обучении 0.001788784245051783
Среднее значение функции потерь на валидации 0.9431613667284028
Модель не улучшилась за последние 10 эпох, прекращаем обучение
```

Оценка качества

```
train_pred = predict_with_model(best_model, train_dataset)

train_loss = F.cross_entropy(torch.from_numpy(train_pred),
```

```
torch.from_numpy(train_source['target']).long())

print('Среднее значение функции потерь на обучении',
float(train_loss))
print('Доля верных ответов', accuracy_score(train_source['target'],
train_pred.argmax(-1)))
print()



test_pred = predict_with_model(best_model, test_dataset)

test_loss = F.cross_entropy(torch.from_numpy(test_pred),

torch.from_numpy(test_source['target']).long())

print('Среднее значение функции потерь на валидации',
float(test_loss))
print('Доля верных ответов', accuracy_score(test_source['target'],
test_pred.argmax(-1)))
```

```
100%|██████████| 354/353.5625 [00:02<00:00, 134.87it/s]

Среднее значение функции потерь на обучении 0.002232222817838192
Доля верных ответов 0.9994696835778681


236it [00:01, 136.99it/s]

Среднее значение функции потерь на валидации 0.92894047498703
Доля верных ответов 0.76805629314923
```

Альтернативная реализация на scikit-learn

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression

sklearn_pipeline = Pipeline((('vect',
TfidfVectorizer(tokenizer=tokenize_text_simple_regex,
                                                    max_df=MAX_DF,

min_df=MIN_COUNT)),
                             ('cls', LogisticRegression())))
sklearn_pipeline.fit(train_source['data'], train_source['target']);
```

Оценка качества

```
sklearn_train_pred =
sklearn_pipeline.predict_proba(train_source['data'])
sklearn_train_loss =
F.cross_entropy(torch.from_numpy(sklearn_train_pred),

torch.from_numpy(train_source['target']))
print('Среднее значение функции потерь на обучении',
float(sklearn_train_loss))
print('Доля верных ответов', accuracy_score(train_source['target'],
sklearn_train_pred.argmax(-1)))
print()

sklearn_test_pred =
sklearn_pipeline.predict_proba(test_source['data'])
sklearn_test_loss =
F.cross_entropy(torch.from_numpy(sklearn_test_pred),

torch.from_numpy(test_source['target']))
print('Среднее значение функции потерь на валидации',
float(sklearn_test_loss))
print('Доля верных ответов', accuracy_score(test_source['target'],
sklearn_test_pred.argmax(-1)))
```

```
Среднее значение функции потерь на обучении 2.4954788918567647
Доля верных ответов 0.9716280714159449

Среднее значение функции потерь на валидации 2.6539022582327556
Доля верных ответов 0.8190387679235263
```