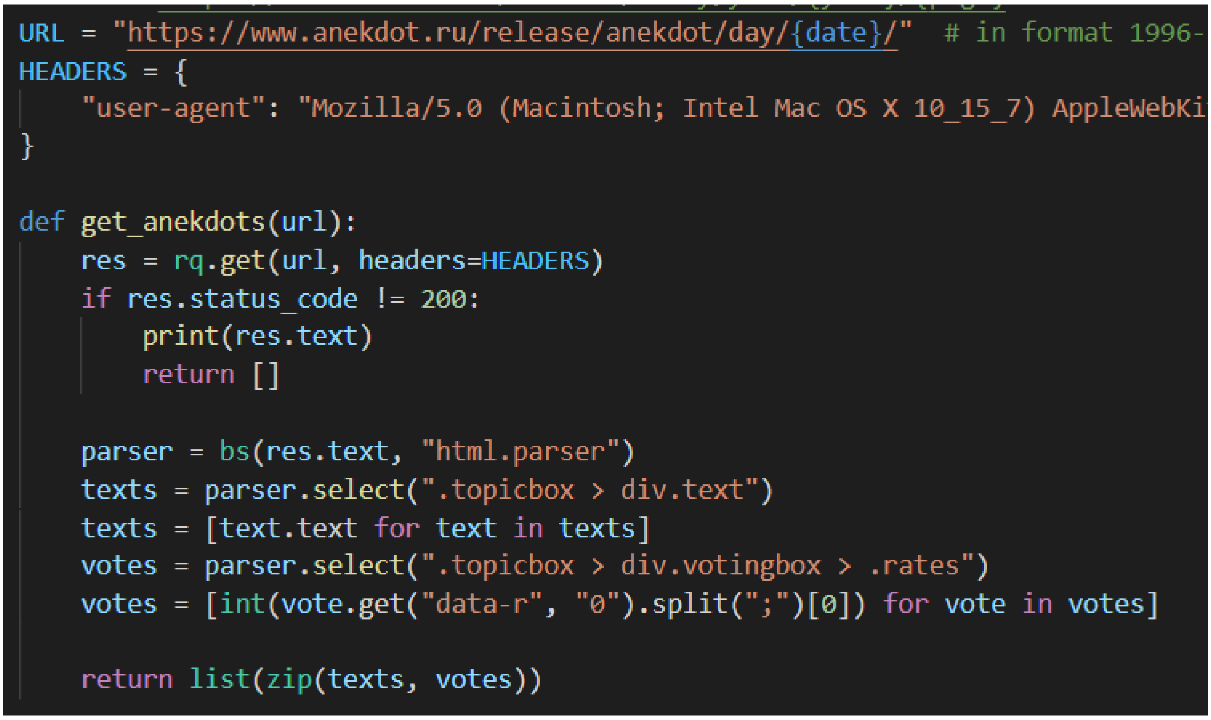
PMDL Project Deliverable 1.2

The first week of the project was dedicated to the acquisition of the jokes dataset. A Python based crawler using the library BeautifulSoup was developed to parse the website (anekdot.ru).

Once the parser has run its course we have approximately 300 000 anekdotes. Each resulting anekdote has the text and an associated ranking value that the users are assigning based on whether they like the joke or not. The negative value represents a substandard anekdote, while a large positive one means the users really like the joke.

The following screenshot is taken from the code for the crawler:



```python
URL = "https://www.anekdot.ru/release/anekdot/day/{date}/"  # in format 1996-
HEADERS = {
    "user-agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKi
}

def get_anekdots(url):
    res = rq.get(url, headers=HEADERS)
    if res.status_code != 200:
        print(res.text)
        return []

    parser = bs(res.text, "html.parser")
    texts = parser.select(".topicbox > div.text")
    texts = [text.text for text in texts]
    votes = parser.select(".topicbox > div.votingbox > .rates")
    votes = [int(vote.get("data-r", "0").split(";")[0]) for vote in votes]

    return list(zip(texts, votes))
```

Our plans for the next week include further work on the dataset, figuring out ways to improve it or find additional sources to parse. Moreover, we plan to develop a baseline model for anekdote generation.