Test task for the Data Science (NLP) Intern

1. The following task would be in English in order to check whether the candidate can operate with the up-to-date documentation.
2. The following task includes theoretical questions and practical task.
3. The following task should be submitted via git repository and include the structured project + answers in the separate file.

Part 1. Theory

1. What are the main problems of modern NLP and NLU?
2. Which libraries would you pick to use for the following cases and why (all problems should be solved for the Russian)
    - Sentiment analysis
    - Multi-label classification
    - Dependency parsing
    - POS-tagging
    - NER
3. How would you evaluate a classification model, which metrics would you use?
4. Main pipeline for the text pre-processing.
5. Microservices or monoliths? Why.
6. Describe the hardest programming task you've been facing with. It's not necessarily ML task, could be just a programming. Why this task was hard to accomplish? What was your solution for the task? Can you share a github project?
7. Did you work with VCS? Which one?
8. Did you work with Github Actions?
9. How familiar are you with Docker and other orchestration tools?
10. What is ed25519 and why is it concerning to be better than ecdsa?
11. Do you have any experience in data mining?

Part 2. Practice

1. Given https://github.com/yutkin/Lenta.Ru-News-Dataset, perform EDA on it focusing on the following:

- Provide descriptive statistics
- Anomaly detection

2. Given the same dataset,

- extract the most syntactically weighted N-grams, omitting nonsense ('казалось бы', 'возможно предположить', etc). The main idea is to extract the most valuable data from the text.
- Try different models for a topic extraction. Which one performs better? What metrics were used to evaluate the model?