1. What are the main problems of modern NLP and NLU?

If we consider NLU as one of the constituent parts of NLP, then understanding natural language is an open source of NLP problems. In addition, one of the problems of modern NLU what biases and structure should we build explicitly into our models to get closer to NLU. Also will be difficult to incorporate a human element relating to emotion. And one of the problems of NLU is the use of knowledge about the cognitive and neuroscience nature in models.

If we talk about NLP in general, then dealing with low-data settings may be open problem of NLP, bcs we don't have a universal solution to this universal problem. Moreover, is there a need to develop special tools for specific languages or it is enough to work on general NLP. In addition, one of the problems is the lexical polysemy of terms. Multi-document summarization and multi-document question answering – problem of the modern NLP. Finally, that a big issue is that there are no datasets available for low-resource languages.

2. Which libraries would you pick to use for the following cases and why (all problems should be solved for the Russian)

- *Sentiment analysis*

Dostoevsky lib, may be used Multinomial Naive Bayes model from sklearn and others classification algorithms (if we have labeled dataset).

- *Multi-label classification*

DeepPavlov, Keras.

- *Dependency parsing*

SyntaxNet (it is component of ISANLP lib).

- *POS-tagging*

MyStem (it is component of ISANLP lib), may be DeepPavlov.

- *NER*

Polyglot (it is component of ISANLP lib), DeepPavlov, Natasha.

3. How would you evaluate a classification model, which metrics would you use?

Precision or recall (depends on the goal: to estimate how many objects of class 1 the model guesses or how much the classifier can be trusted), F-measure (harmonic mean of precision or recall), accuracy, PR-curve or ROC-curve.

4. Main pipeline for the text pre-processing.

Tokenization, stemming or lemmatization, tagging (POS-tagging), stop-words removal, words embeddings. Then we can do input transformation (for example – bag or words, TF-IDF and etc).

5. Microservices or monoliths? Why.

Microservices, bsc in a microservice architecture, the individual processes are broken out into independent services. Microservices (ms) lose to monolithics (ml) in latency and complexity, but ms win to ml in reliability, scalability, resource usage.

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?

It was my bachelor's work. I needed to build an algorithm for selecting personal recommendations for an expert recommendation system using NLP algorithms. The dataset was taken from kaggle and contained textual descriptions of movie plots (from wiki). I used the topic modeling algorithm (LDA, using Gensim lib), but not for the entire corpus, but for individual descriptions. Now I think this is a bad idea. I compared the resulting sets of topics for semantic proximity. To do this, I first used an algorithm to remove lexical ambiguity (Adapted Lesk algorithm, which differs from the classical use of hyperonyms), using pyWSD lib. The definition of semantic similarity was carried out by the wu-palmer measure also using hyperonyms (using WordNet).

I think that the thesaurus approach is not quite a good idea and now I would carry out the search for semantic similarity with other tools, but then I set myself a specific task and implemented it in this way.

Github: https://github.com/ptfrwrd/MovieRetrieval

7. Did you work with VCS? Which one?

Only git.

8. Did you work with Github Actions?

I know what it is, but have no experience of using.

9. How familiar are you with Docker and other orchestration tools?

I have only theoretical knowledge about Docker.

10. What is ed25519 and why is it concerning to be better than ecdsa?

These are keys generated using different encryption algorithms. ECDSA is the elliptic curve implementation of DSA, and it is sensitive to bad RNGs. Ed25519 - this is EdDSA signature scheme, but using SHA-512/256 and Curve25519; it is a protected elliptical curve that provides better security than ECDSA and has better performance.

11. Do you have any experience in data mining?

Yes, I have some experience with topic modeling (it is my bachelor's work), I take courses on the coursera.org completing assignments (and I take some courses on the stepik.org) and my master's program at university – data science. I have a good theoretical base (in machine learning and mathematics), but not much experience.