Summarcha

Chat Summarizer using Machine Learning

Progress report

During past weeks we decided to explore existing solutions in order to better understand the topic and build our own on top of the relevant approaches in the field. As a result we managed to implement a minimal working prototype, allowing us to summarize messages sent to it.

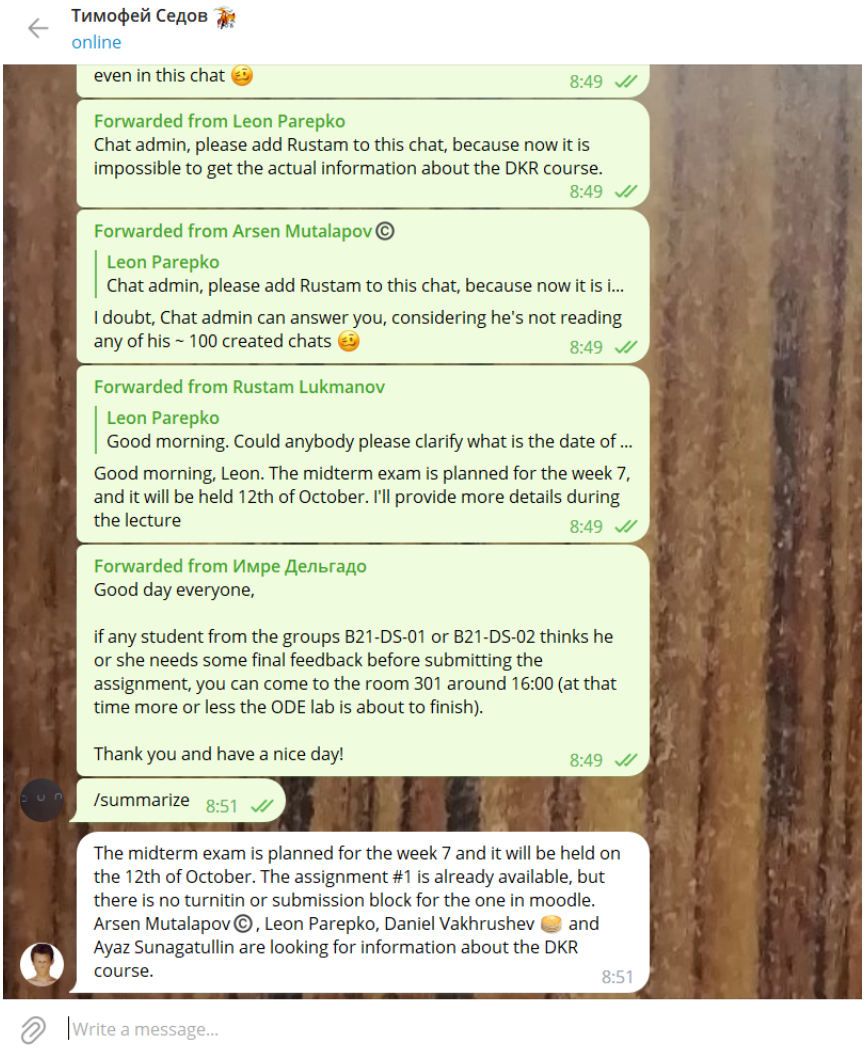
Creating dataset

Creating a dataset was one of the tasks to complete in this time interval. Therefore, as promised, we put our efforts into getting as much information as possible from the telegram chat. So these are the messages with their metadata such as the person, who sent it, timestamp of the message, number of reactions etc. Going further, we plan to extend it with more information using other models to understand toxicity, emotionality, sarcasm, complexity and other useful aspects of the sentences.

	index	text	a..	author_id	reactions_count
94	147	like if you like to r...	Fu...	529046301	6
275	416	demolish window...	M...	324137190	6
280	421	and there s no po...	y4...	321599704	6
336	482	offer your candid...	M...	324137190	6
399	552	i got it right	Le...	1049972060	6
471	654	with every questi...	M...	694220158	6
493	685	to shrek in the sw...	M...	324137190	6
538	739	by the way yes w...	cu...	582021012	6
83	135	i thought	sn...	61829189	5
223	343	this is a zip bomb...	cu...	582021012	5
305	449	you just have to f...	Fu...	529046301	5
486	677	our	M...	324137190	5

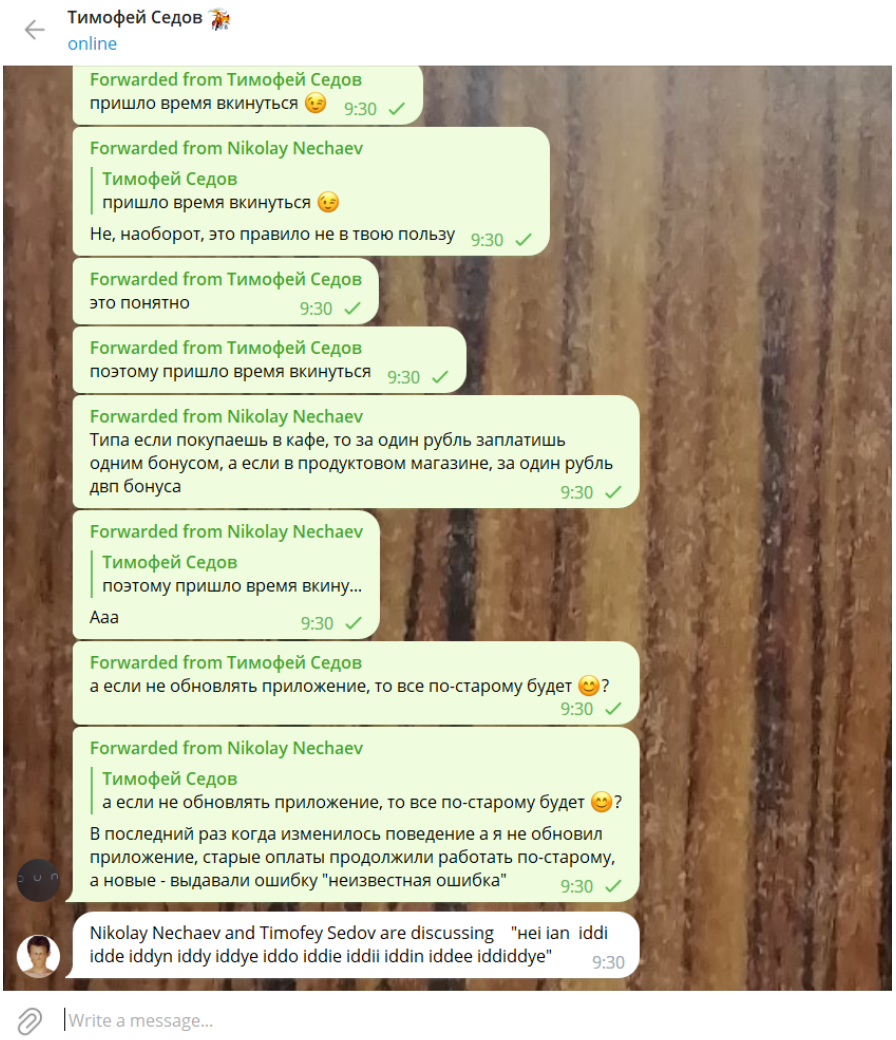
Model choosing

Searching for the model fitting our purposes we stuck with the field of the abstract dialogue summarization. Under further investigation, we found one promising solution involving using a fine-tuned BART model. We decided to use it for our very first solution. We created a telegram bot serving HuggingFace API which can summarize up to 20 forwarded messages at a time.



Testing

Our telegram bot worked smoothly, except for Russian symbols, which our dataset was mostly created of. So from this moment we have 2 options. Either fine-tune the model for Russian language, or translate the messages using telegram API (using premium-user feature). For now we stick with the 2nd option, but till the next project report, we plan to try the first approach too.



Further work

As we previously said, our further work consists of pre-cleaning the data using metrics from companion models and fine-tuning our model for datasets. Additionally, we plan to work more on user-friendly experience, allowing customization, scheduling, etc.

Team members

Full name	**Innopolis email**	**Tasks**
Timofey Sedov	t.sedov@innopolis.university	Models research, ML, backend
Grigorii Fil	g.fil@innopolis.university	ML, Dataset search, backend, testing
Andrei Markov	a.markov@innopolis.university	Models training, full-stack, deploy

Git repository

https://github.com/markovav-official/summarcha