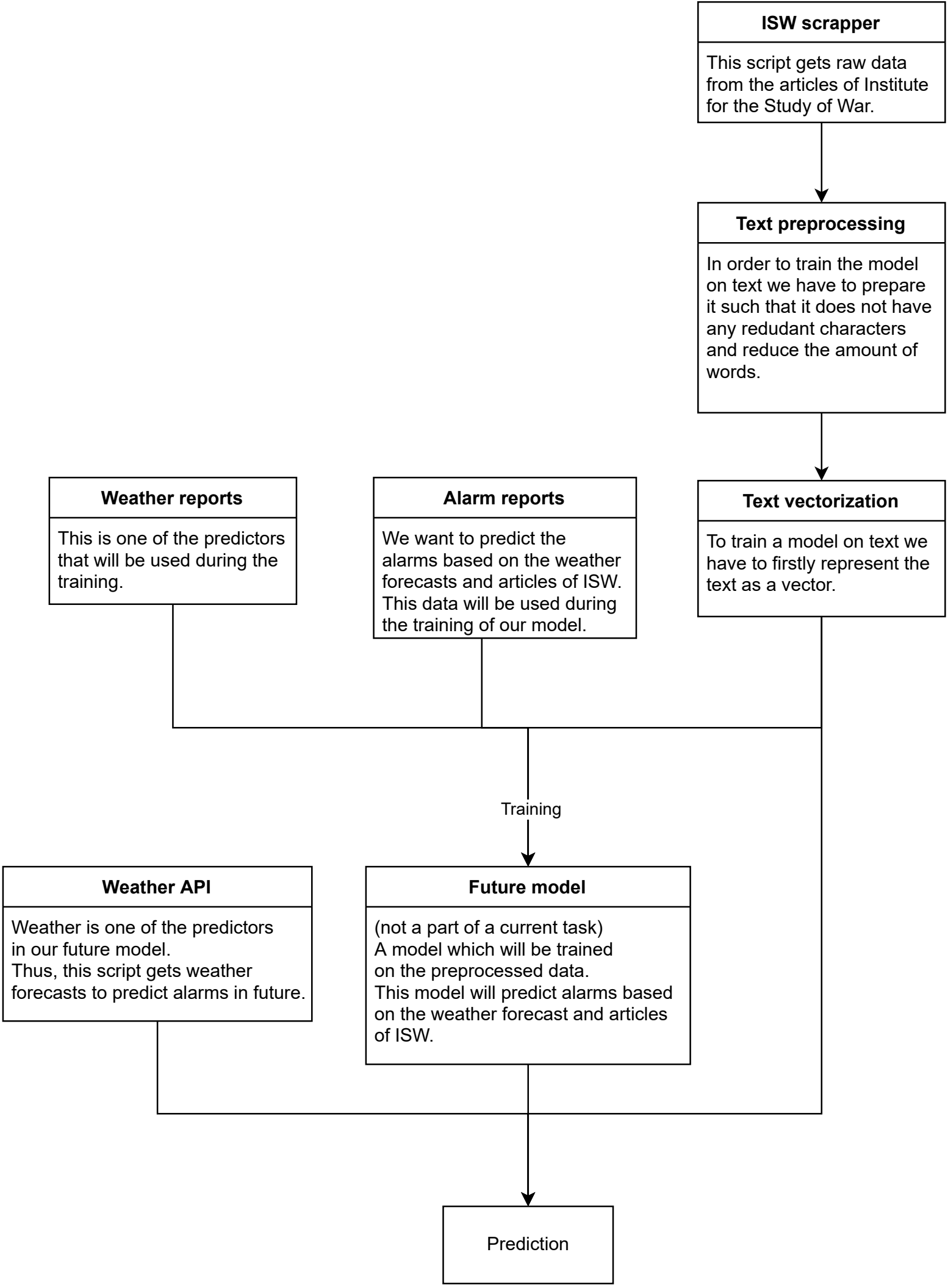
Data preprocessing with NLTK
Plan

group 9
Tymur Mykhalievskyi,
Zasyadko Matviy,
Kysilenko Danyil,
Zakutin Tymofii

March 15, 2023

System diagram

This diagram shows the relations between the different parts of the task.


ISW scrapper
This script gets raw data from the articles of Institute for the Study of War.
Text preprocessing
In order to train the model on text we have to prepare it such that it does not have any redudant characters and reduce the amount of words.
Weather reports
This is one of the predictors that will be used during the training.
Alarm reports
We want to predict the alarms based on the weather forecasts and articles of ISW. This data will be used during the training of our model.
Text vectorization
To train a model on text we have to firstly represent the text as a vector.
Training
Weather API
Weather is one of the predictors in our future model.
Thus, this script gets weather forecasts to predict alarms in future.
Future model
(not a part of a current task)
A model which will be trained on the preprocessed data.
This model will predict alarms based on the weather forecast and articles of ISW.
Prediction

Text data preparation

The main reason of preparing text is to get rid of irrelevant data for the predictions and vectorize the text to train the model on it. This task can be split into multiple subtasks:

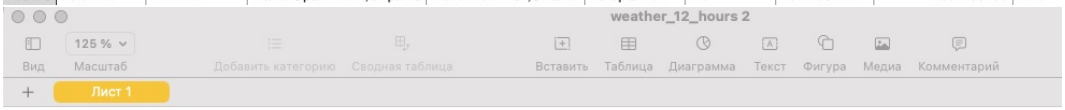
- using patterns we can delete characters that are irrelevant for the predictions. Also we should delete links, enters and non-breaking spaces.
- stemming and lemmatization to make the words with different endings the same words.
- when the text is prepared to be trained on we have to vectorize it to be able to perform mathematical operations on it during the training.

After performing these steps with a good selection of a model we will be able to predict the alarms based on weather forecasts and articles from the ISW.

Screenshots of results

Weather forecasts for the next 12 hours

	city_latitude	city_longitude	city_resolvedAddress	city_address	city_timezone	city_tzoffset	day_datetime	day_datetimeEpoch	day_te
163234	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163235	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163236	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163237	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163238	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163239	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163240	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163241	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163242	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163243	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163244	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6
163245	48.9226	24.7147	Івано-Франківськ, Україна	Ivano-Frankivsk,Ukraine	Europe/Kiev	2.0	2022-08-24	1661288400	29.6



icon	day_source	day_preciptype	day_stations	hour_datetime	hour_datetimeEpoch	hour_temp	hour_feelslike	hour_humidity	hour_dew	hou
y-cloudy-day	obs		33000599999;remote	11:00:00	1661328000	23.4	23.4	73.06	18.3	0.0
y-cloudy-day	obs		33000599999;remote	12:00:00	1661331600	24.7	24.7	61.86	16.9	0.0
y-cloudy-day	obs		33000599999;remote	13:00:00	1661335200	26.6	26.6	60.37	18.3	0.0
y-cloudy-day	obs		33000599999;remote	14:00:00	1661338800	27.9	28.7	54.2	17.8	0.0
y-cloudy-day	obs		33000599999;remote	15:00:00	1661342400	28.6	28.8	47.62	16.4	0.0
y-cloudy-day	obs		33000599999;remote	16:00:00	1661346000	29.3	29.4	45.15	16.2	0.0
y-cloudy-day	obs		33000599999;remote	17:00:00	1661349600	29.5	29.7	45.79	16.6	0.0
y-cloudy-day	obs		33000599999;remote	18:00:00	1661353200	29.6	29.6	44.1	16.1	0.0
y-cloudy-day	obs		33000599999;remote	19:00:00	1661356800	28.3	28.1	43.18	14.6	0.0
y-cloudy-day	obs		33000599999;remote	20:00:00	1661360400	26.3	26.3	55.19	16.6	0.0
y-cloudy-day	obs		33000599999;remote	21:00:00	1661364000	22.7	22.7	75.75	18.2	0.0
y-cloudy-day	obs		33000599999;remote	22:00:00	1661367600	21.1	21.1	81.44	17.8	0.0

Vectorized data from ISW

Out[4]:

	01	05000600	08000830	10	100	1000	10000	100000	1000000	100000rubl	...	zurab	zusko	zvanivka	zvezda	zvinchuk	zyabrovka	zybyn	zymoł
0	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
1	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
2	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
3	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
4	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
320	0	0	0	1	0	0	0	1	0	0	...	0	0	0	0	0	0	0	
321	0	0	0	1	0	0	1	0	0	0	...	0	0	0	0	0	0	0	
322	0	0	0	1	1	0	1	0	0	0	...	0	0	0	0	0	0	0	
323	0	0	0	2	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
324	0	0	0	3	1	0	2	0	0	0	...	0	0	0	0	0	0	0	