National University of “Kyiv-Mohyla Academy”
Faculty of Informatics

Homework #2: Project sturcutre and data

Group #8:
Bohdan Shevchenko
Vadym Zaitsev
Danyil Melnyk
Artur Yurchenko

Kyiv 2023

Content

The project plan 3

Link to the repo..... 6

Link to the diagram..... 6

The project plan

At this stage our project consists of 2 parts:

1. Collecting and preprocessing the text data from the ISW Website for every day of the full-scale invasion of russia in Ukraine.
2. Getting the weather forecast for the next 12 hours using Visual Crossing weather API. Later this will be used to create a model that will be able to forecast air alarms.

1. The realization of this part consists of 4 files:
   - Proga_geyHTML.ipynb – saves the HTML files of the Webpages in a folder fff.

     Developed by Bohdan Shevchenko

     Example of an HTML file:

```
<!DOCTYPE html>
<html xmlns:og="http://opengraphprotocol.org/schema/" lang="en" dir="ltr">

<head>
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="Generator" content="Drupal 7 (http://drupal.org)" />
<link rel="canonical" href="/backgrounder/russia-ukraine-warning-update-russian-offensive-campaign-assessment-february-25-2022" />
<link rel="shortlink" href="/node/4465" />
<link rel="shortcut icon" href="https://www.understandingwar.org/sites/default/themes/isw/isw.ico" type="image/vnd.microsoft.icon" />
<meta name="keywords" content="Ukraine Project,Institute for the Study of War, ISW, War, Institute, Study, ORBAT, Report, Iraq, Iran, Afghanistan, Middle East, Libya, Security
<meta name="description" content="Russian forces entered major Ukrainian cities—including Kyiv and Kherson—for the first time on February 25. Russian forces' main axes of adva
  <title>Russia-Ukraine Warning Update: Russian Offensive Campaign Assessment, February 25, 2022 | Institute for the Study of War</title>
  <style type="text/css" media="all">
@import url("https://www.understandingwar.org/modules/system/system.base.css?rayj4d");
@import url("https://www.understandingwar.org/modules/system/system.menus.css?rayj4d");
@import url("https://www.understandingwar.org/modules/system/system.messages.css?rayj4d");
@import url("https://www.understandingwar.org/modules/system/system.theme.css?rayj4d");
</style>
<style type="text/css" media="all">
```

   - Proga_text_processing.ipynb – iterates over the HTML files, divides the information there into title, full_title, date and main_text, stores this in a list of dictionaries, which is later cast to a Dataframe and saved in a table_of_data.csv file.

     Developed by Bohdan Shevchenko and Danyil Melnyk

table_of_data.csv:

	date	title	text_title	text_main
0	2022-02-25	Russia-Ukraine Warning Update: Russian Offensive Campaign Assessment, February 25, 2022 | Institute for the Study of War	Russian Offensive Campaign Assessment, January 20, 2023	<div class="field-items' February 25, 3:00 pm EST Russian forces entered major Ukrainian cities—including Kyiv and Khers Key Takeaways <ul><li>Russian forces entered the outskirts of Kyiv on the west bank of the Dnipro on February 25. Russian <li>Russian forces have so far failed to enter Kyiv's eastern outskirts. Ukrainian forces have successfully slowe <li>Elements of the Russian 76th VDV (Airborne) division have concentrated in southeastern Belarus likely fo <li>Russian forces will likely envelop Kharkhiv in the next 24 hours after failing to enter the city through fron <li>Russian forces have achieved little success on frontal assaults or envelopments against Ukrainian forces i <li>North of Crimea, Russian forces fully captured Kherson and are likely on the verge of seizing Melitopol ir <li>Russian forces may be assembling in Stolin, Belarus, to open a new line of advance against Rivne in west </ul> <a href="https://www.understandingwar.org/sites/default/files/DraftUkraineCoTFeb25%2C2022.png"> <img alt="Control of terrain map of Ukraine showing Russian lines of advance" height="794" src="http Russian forces carried out additional air and missile strikes on Kyiv and Russian ground forces are advancing on four primary axes, discussed in <ol><li>Belarus/Kyiv;</li> <li>Kharkiv;</li> <li>Donbas; and</li> <li>Crimea-Kherson.</li> </ol> 1) Belarus/Kyiv axis: Russian forces entered the outskirts of Kyiv <ul><li>Russian forces entered the northwest Kyiv suburb of Obolon, on the western bank of the Dnipro Riv

- 2_isw_reports_preparetion.ipynb – the first part of filtering the text, the filtered data is stored in the table_of_data_preprocessed.csv

Developed by Danyil Melnyk with some help from Bohdan Shevchenko.

table_of_data_preprocessed.csv:

	date	title	text_title	main_html_v8	text_main
0	2022-02-25	Russia-Ukraine Warning Update: Russian Offensive Campaign Assessment, February 25, 2022 | Institute for the Study of War	Russian Offensive Campaign Assessment, January 20, 2023	Russian forces entered major Ukrainian cities—including Kyiv and Kherson—for the first time on February 25. Russian forces' main axes of advance focused on Kyiv (successfully isolating the city on both banks of the Dnipro River). Russian military operations along Ukraine's northern border have been less well-planned, organized, and conducted than those emanating from Crimea. They have also been less successful so far. The divergence in performance likely arises in part from differences in the composition and organization of the Russian ground forces elements in the Western Military District and Belarus (to Ukraine's north) and Southern Military District and Black Sea Fleet (to its south and east), as ISW has previously observed. Determined and well-organized Ukrainian	<div class="field field-name-body field-type-text <stron Russian forc Key Takeawa <ul><li>Russian forces entered the outskirts of K <li>Russian forces have so far failed to enter Kyiv <li>Elements of the Russian 76th VDV (Airborne) <li>Russian forces will likely envelop Kharkhiv in <li>Russian forces have achieved little success or <li>North of Crimea, Russian forces fully capture <li>Russian forces may be assembling in Stolin, B </ul> <a href="https://www.understandingwar.or <img alt="Control of terrain map of Ukrain Russian forc Russian grou <ol><li>Belarus/Kyiv;</li> <li>Kharkiv;</li> <li>Donbas; and</li> <li>Crimea-Kherson.</li> </ol> 1) B <ul><li>Russian forces entered the northwest Ky

- Processing.ipynb – the second part of filtering the text + stemming, lemmatizing. This data is stored in table_of_data_processed.csv. After that, the text is vectorized using TF-IDF, the keywords are extracted and then the data is stored in table_vectorized.csv. This is exactly the data that then will be used for the forecasting.

Developed by Artur Yurchenko and Danyil Melnyk

table_of_data_processed.csv:

te	title	text_title	text_main
!022-02-27	Russia-Ukraine Warning Update: Russian Offensi...	Russian Offensive Campaign Assessment, January...	<div class
!022-02-28	Russian Offensive Campaign Assessment, Februar...	Russian Offensive Campaign Assessment, January...	<div class
!022-03-01	Russian Offensive Campaign Assessment, March 1...	Russian Offensive Campaign Assessment, January...	<div class
!022-03-02	Russian Offensive Campaign Assessment, March 2...	Russian Offensive Campaign Assessment, January...	<div class
!022-03-03	Ukraine Conflict Update 14 | Institute for the...	Russian Offensive Campaign Assessment, January...	<div class
!022-03-04	Ukraine Conflict Update 15 | Institute for the...	Russian Offensive Campaign Assessment, January...	<div class
!022-03-05	Explainer on Russian Conscription, Reserve, an...	Russian Offensive Campaign Assessment, January...	<div class
!022-03-06	Ukraine Conflict Update 16 | Institute for the...	Russian Offensive Campaign Assessment, January...	<div class

10 rows × 6 columns Open in new tab

main_html4	data_stemmed	data_lemmatized	keywords
russian forces main axes advance last twenty ...	russian forc main axe advanc last twenti four...	russian force main ax advance last twenty fou...	{'february': 0.409, 'kyiv': 0.367, 'twenty': 0.
russian military likely recognized initial ex...	russian militari like recogn initi expect lim...	russian military likely recognized initial ex...	{'february': 0.496, 'seven': 0.35, 'twenty': 0.
russian military reorganizing military effort...	russian militari reorgan militari effort atte...	russian military reorganizing military effort...	{'february': 0.553, 'eight': 0.348, 'twenty': .
russian forces completing reinforcement resup...	russian forc complet reinforc resuppli troop ...	russian force completing reinforcement resupp...	{'kyiv': 0.359, 'chernihiv': 0.219, 'hour': 0..
russian forces resumed offensive operations s...	russian forc resum offens oper support envelo...	russian force resumed offensive operation sup...	{'march': 0.561, 'kyiv': 0.392, 'chernihiv': 0.
critical threats project aei march isw publis...	critic threat project aei march isw publish r...	critical threat project aei march isw publish...	{'march': 0.589, 'foreign': 0.179, 'eu': 0.179.
critical threats project aei march isw publis...	critic threat project aei march isw publish r...	critical threat project aei march isw publish...	{'march': 0.601, 'nato': 0.161, 'lukashenko': .
early announcement spring draft unlikely incr...	earli announc spring draft unlik increas russ...	early announcement spring draft unlikely incr...	{'reservist': 0.394, 'reserve': 0.269, 'conscr.

table_vectorized.csv:

Unnamed: 0;date;title;text_title;text_main;main_html_v8;main_html;main_html1;main_html2;main_html3;main_html4;data_stemmed;data_lemmatized;keywords

0;2022-0 February 25 2022 | Ins January 2(2023;"<di George Ba and Kateryna Stepanenko</p>

<p align= 3:00 pm EST</p>

<p style= organized and condu as ISW ha likely beyc to offset t however. Russia will likely defeat Ukrainian regular military forces and secure their territ

<p style=""text-align: left;"">Key Takeaways</p>

<ul><li>Russian forces entered the outskirts of Kyiv on the west bank of the Dnipro on February 25. Russian sabotage groups in civilian clothes are reportedly active in downtown K

<li>Russia which have temporarily abandoned the failed attempt to take the city of Chernihiv and are instead bypassing it.</li>

<li>Elements of the Russian 76th VDV (Airborne) division have concentrated in southeastern Belarus likely for use along the Chernihiv-bypass axis toward Kyiv in the next 24 hours.<

<li>Russian forces will likely envelop Kharkhiv in the next 24 hours after failing to enter the city through frontal assaults on February 24.</li>

<li>Russian forces have achieved little success on frontal assaults or envelopments against Ukrainian forces in Donbas but may not have intended to do more than pin Ukrainian fo

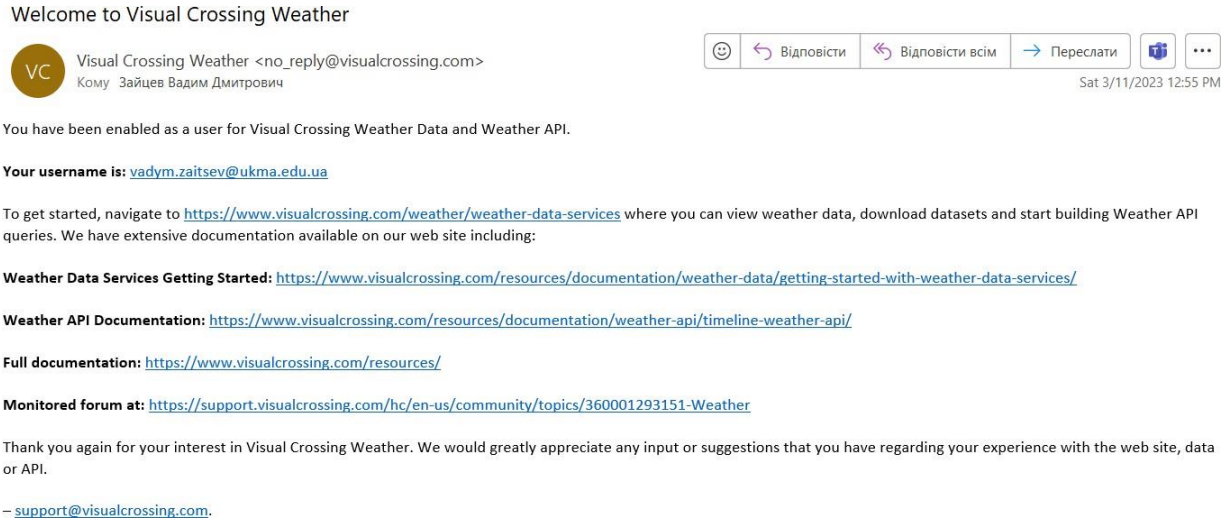
<li>North Russian forces fully captured Kherson and are likely on the verge of seizing Melitopol in the east. Unconfirmed reports indicate that Russian forces had bypassed Kherson

<li>Russia Belarus to open a new line of advance against Rivne in western Ukraine.</li>

</ul> Â

<a href=""https://www.understandingwar.org/sites/default/files/DraftUkraineCoTFeb25%2C2022.png"">Click here to enlarge the map below.</a>

2. The weather part is considerably simpler than the previous one. It consists of only one file weather.ipynb and uses Visual Crossing Weather API



Welcome to Visual Crossing Weather

VC Visual Crossing Weather <no_reply@visualcrossing.com>
Кому Зайцев Вадим Дмитрович

Відповісти | Відповісти всім | Переслати

Sat 3/11/2023 12:55 PM

You have been enabled as a user for Visual Crossing Weather Data and Weather API.

Your username is: vadym.zaitsev@ukma.edu.ua

To get started, navigate to https://www.visualcrossing.com/weather/weather-data-services where you can view weather data, download datasets and start building Weather API queries. We have extensive documentation available on our web site including:

Weather Data Services Getting Started: https://www.visualcrossing.com/resources/documentation/weather-data/getting-started-with-weather-data-services/

Weather API Documentation: https://www.visualcrossing.com/resources/documentation/weather-api/timeline-weather-api/

Full documentation: https://www.visualcrossing.com/resources/

Monitored forum at: https://support.visualcrossing.com/hc/en-us/community/topics/360001293151-Weather

Thank you again for your interest in Visual Crossing Weather. We would greatly appreciate any input or suggestions that you have regarding your experience with the web site, data or API.

– support@visualcrossing.com.

The function get_weather_forecast() returns a list of dictionaries representing the weather forecast for the next 12 hours, which can be transformed into a Dataframe using the vectorize() function.

```
ready_data = vectorize(output)
ready_data.head(1)
```

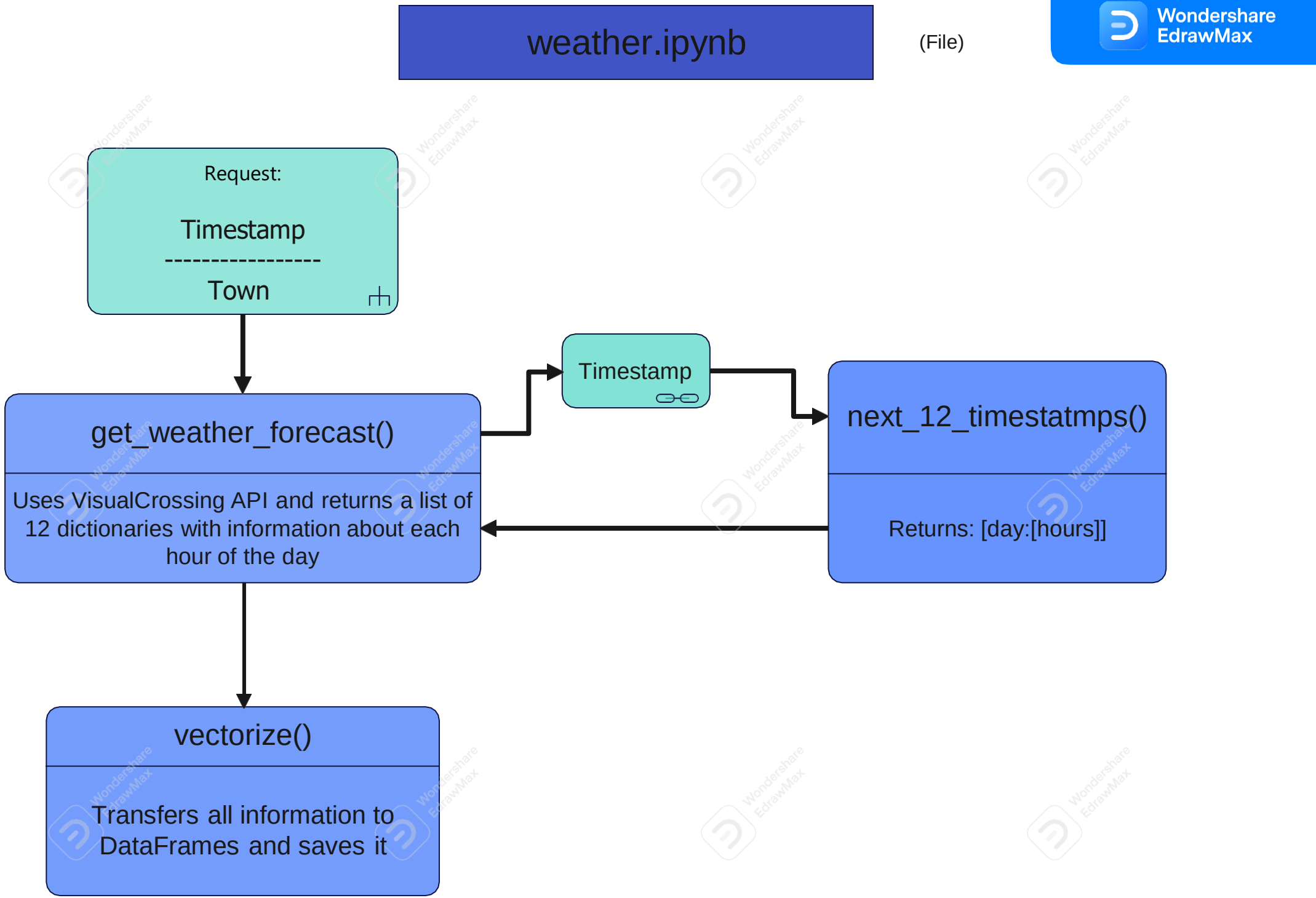
	city_latitude	city_longitude	city_resolvedAddress	city_address	city_timezone	city_tzoffset	day_datetime	day_datetimeEpoch	day_tempmax	day_tempmin	...	hour_visibility
0	49.5879	34.5517	Полтава, Україна	Poltava,UA	Europe/Kiev	2.0	2023-03-15	1678831200	10.9	1.4	...	24.1

Developed by Vadym Zaitsev

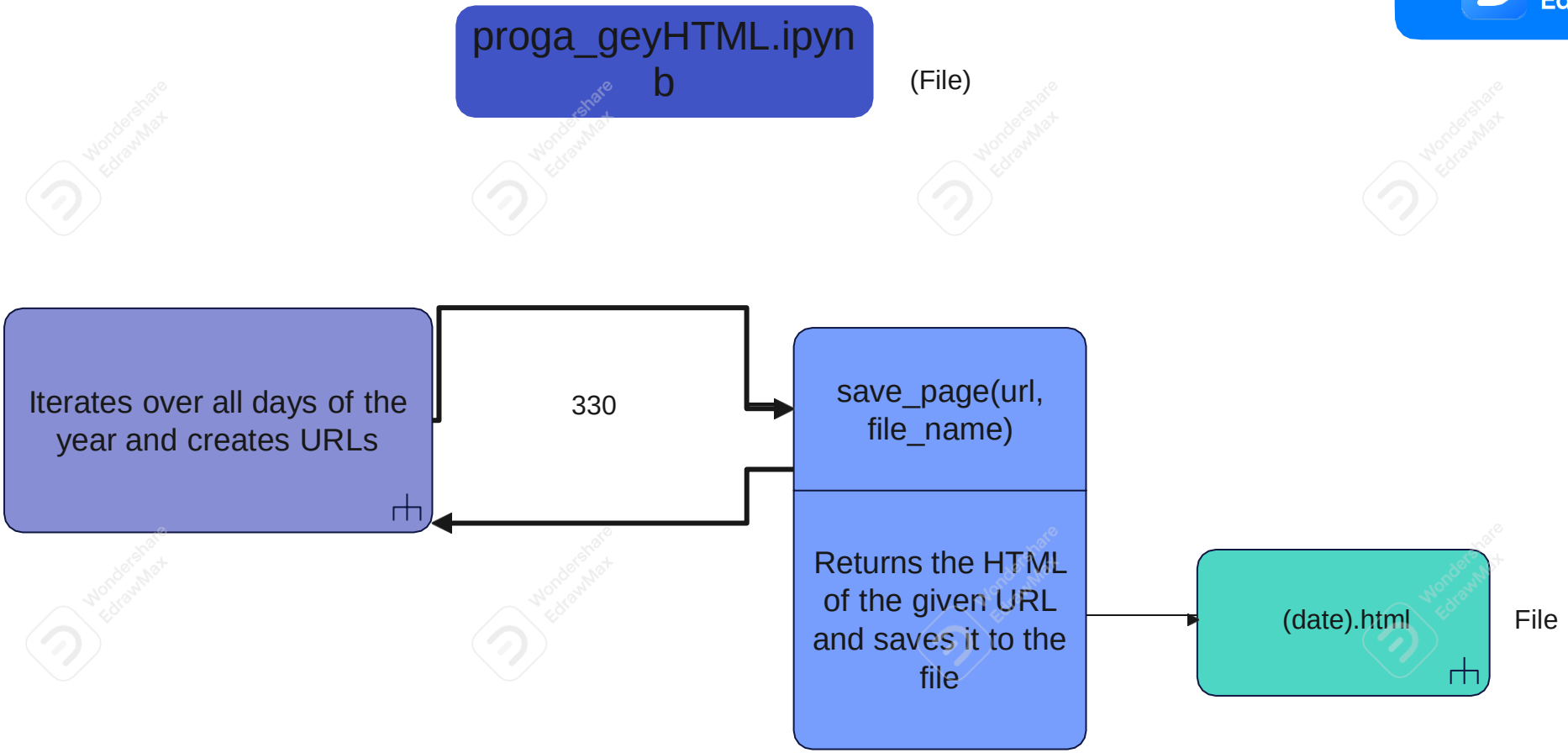
Link to the repo

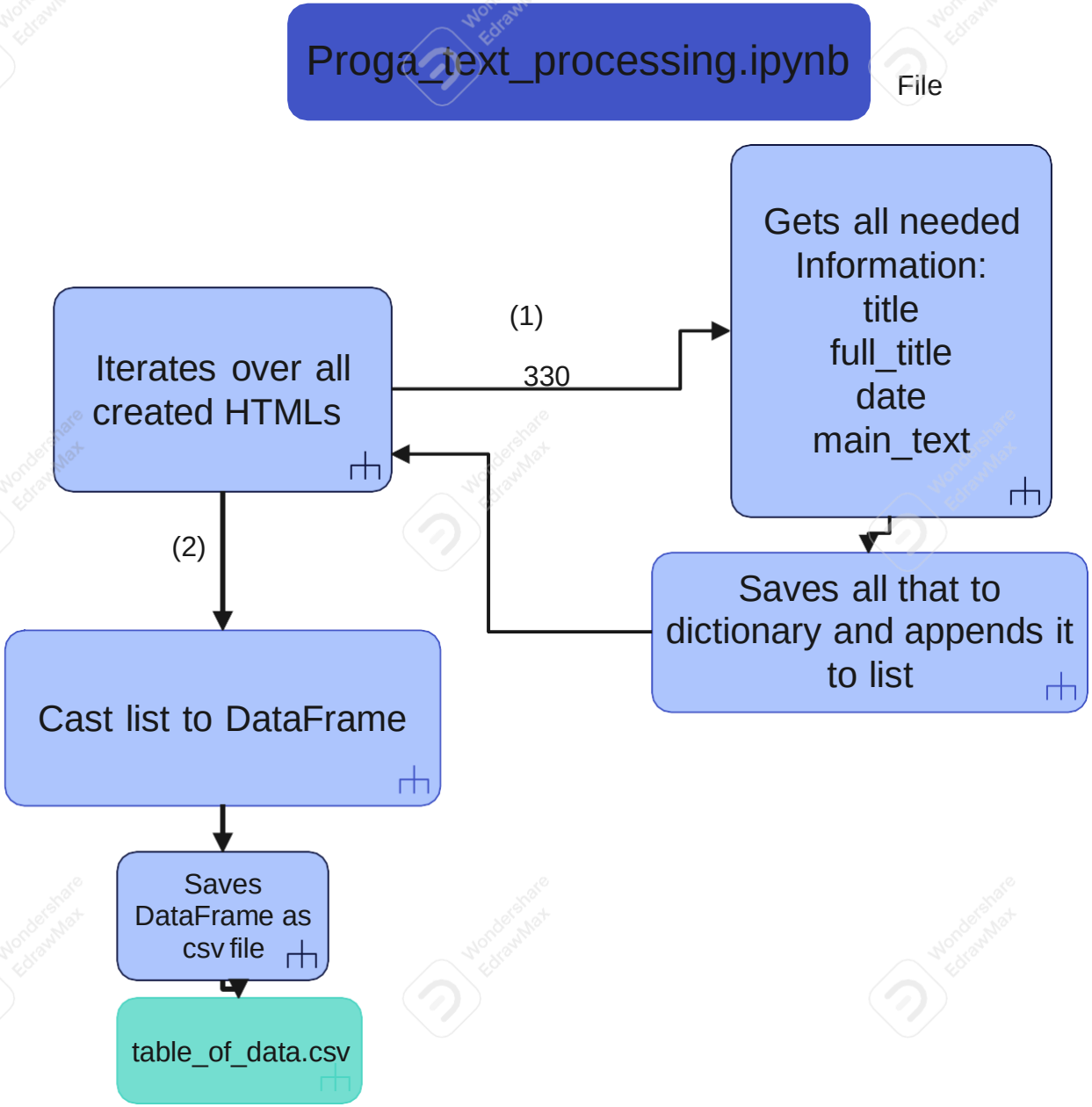
The report has been done by Vadym Zaitsev

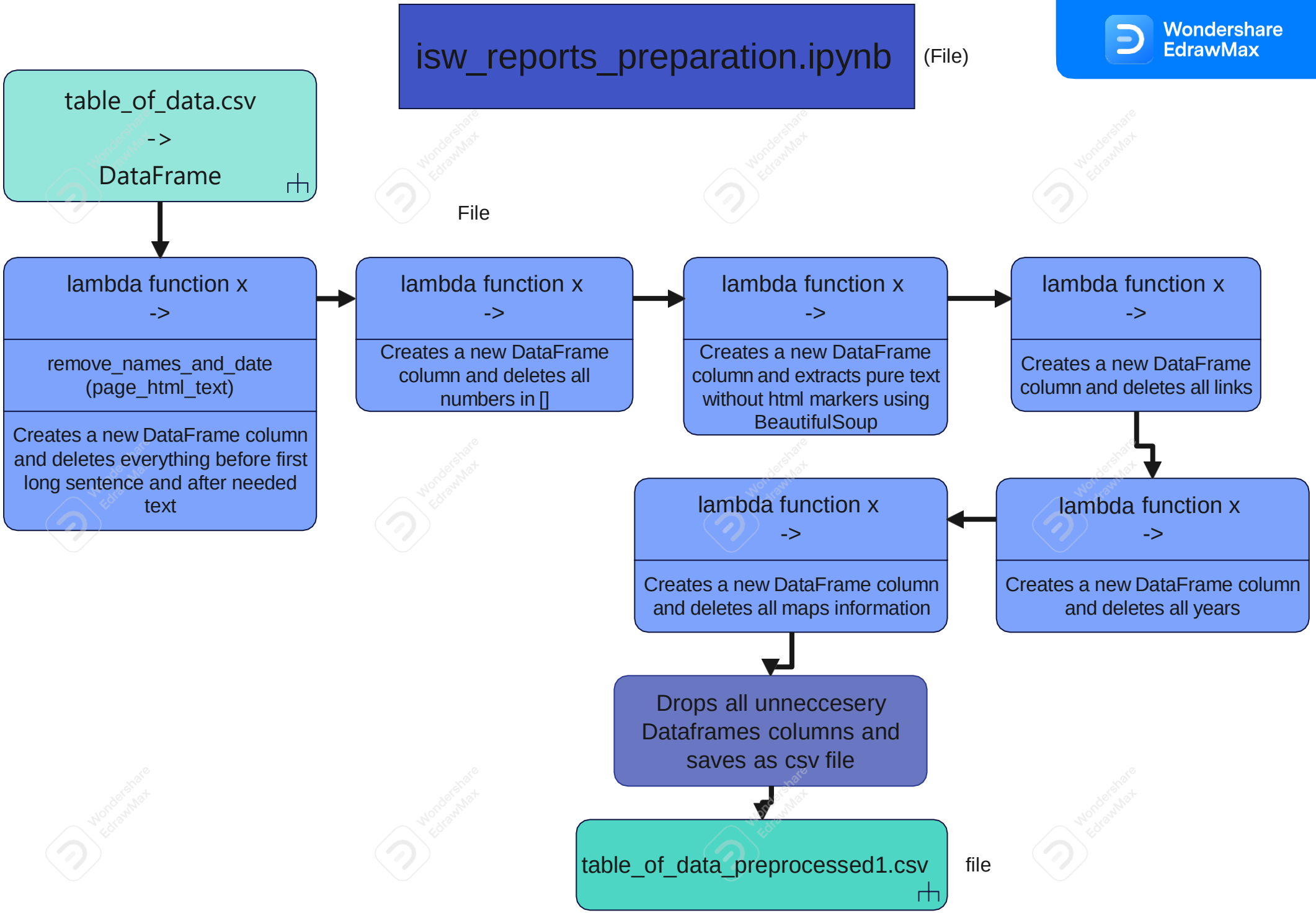
The diagram was created by Bohdan Shevchenko:


weather.ipynb
(File)
Request:
Timestamp

Town
Timestamp
get_weather_forecast()
Uses VisualCrossing API and returns a list of 12 dictionaries with information about each hour of the day
next_12_timestatmps()
Returns: [day:[hours]]
vectorize()
Transfers all information to DataFrames and saves it


proga_geyHTML.ipynb
(File)
Iterates over all days of the year and creates URLs
330
save_page(url, file_name)
Returns the HTML of the given URL and saves it to the file
(date).html
File


Proga_text_processing.ipynb
File
Iterates over all created HTMLs
(1)
330
Gets all needed Information:
title
full_title
date
main_text
(2)
Saves all that to dictionary and appends it to list
Cast list to DataFrame
Saves DataFrame as csv file
table_of_data.csv


isw_reports_preparation.ipynb
(File)
table_of_data.csv
->
DataFrame
File
lambda function x
->
remove_names_and_date (page_html_text)
Creates a new DataFrame column and deletes everything before first long sentence and after needed text
lambda function x
->
Creates a new DataFrame column and deletes all numbers in []
lambda function x
->
Creates a new DataFrame column and extracts pure text without html markers using BeautifulSoup
lambda function x
->
Creates a new DataFrame column and deletes all links
lambda function x
->
Creates a new DataFrame column and deletes all years
lambda function x
->
Creates a new DataFrame column and deletes all maps information
Drops all unneccesery Dataframes columns and saves as csv file
table_of_data_preprocessed1.csv
file

Processing.ipynb (File)

table_of_data.csv -> DataFrame

All characters are made lower case using nltk

lambda function x -> creates a new column->
remove_any_punct(q)
Returns text with deleted punctuation

lambda function x ->creates a new column->
remove_special_characters(data)
Returns text with deleted special characters.

lambda function x ->creates a new column->
remove_one_letter_word (data)
Returns text with removed short words such as: a I...

lambda function x ->creates a new column->
lemmatizing(data)
Returns text with switched any kind of a word to its base root mode. Using lemmatizer and nltk tokenizer

lambda function x ->creates a new column->
stemming(data)
Returns text with changed suffixes and prefixes and some roots. Using stemmer and nltk tokenizer

lambda function x ->creates a new column
convert_numbers(x)
Returns text with all relatively small numbers transfered to words. Using nltk tokenizer
lambda function y ->
remove_stopwords(data)
Returns text with removed stopwords from nltk library except for "no" "not"

Throws away all other colums of DataFrame and saves it as csv file

table_of_data_processed.csv

lambda function x ->creates a new column->
conver_doc_to_vector(doc)
Transforms text into TF-IDF vector representation using the transform method of a TfidfVectorizer.
Then it sorts the vector by values
Then extracts Keywords

sort_coo(coo_matrix)
Sorts the vector by values

extract_topn_from_vector()
Extracts the Keywords

Saves vector to a new file

text_vectorized.csv