Assignment 3 Final Report

Team A++: Ziyue Chen, Ying Wang, Cheng Shi, Zichao Wang, Ziquan Chen, Wenting Shang

Abstract: In this assignment, we focused on practicing Optical Character Recognition (OCR), Named Entities Recognition (NER), and Geoparsing and Geolocation of extracted entities. The news images were first cropped into smaller sizes so that the python package Tesseract can be used for OCR. Then, we performed NER on the news content and obtained five entities, which were PERSON, NORP, ORG, GPE and LOC, and we plotted the top 20 mentioned words for each entity in the entire dataset. In addition, we also compared the distribution of each entity for the individual news sources, providing another 15 barplots. In the end, GeoPy was used to find the longitude and latitude of GPE and LOC entities to generate dynamic World and Ukraine maps.

1. Introduction

In this report, we will first discuss challenges faced when we performed the Optical Character Recognition technology and what methods we took to solve the problems. Then we will illustrate our observations from the plots of the distribution of the entities in the text that we extracted from all news images. And finally, we will discuss difficulties encountered when performing the geoparsing.

2. How difficult was performing the OCR on provided images? Was there any step in the OCR process that was particularly challenging?

It was challenging when we tried to split the images and cut the ads and menus. We were first thinking of cutting the images manually to remove the menus and ads, but it was too tedious for hundreds of images. Therefore, we came up with the idea of using the crop function in python to cut the top part of each image where the menus are usually located. Other than that, we found there still were some ads in the middle of some news images, which was the most challenging part for us and we couldn't figure out a way to let the machine do the work.

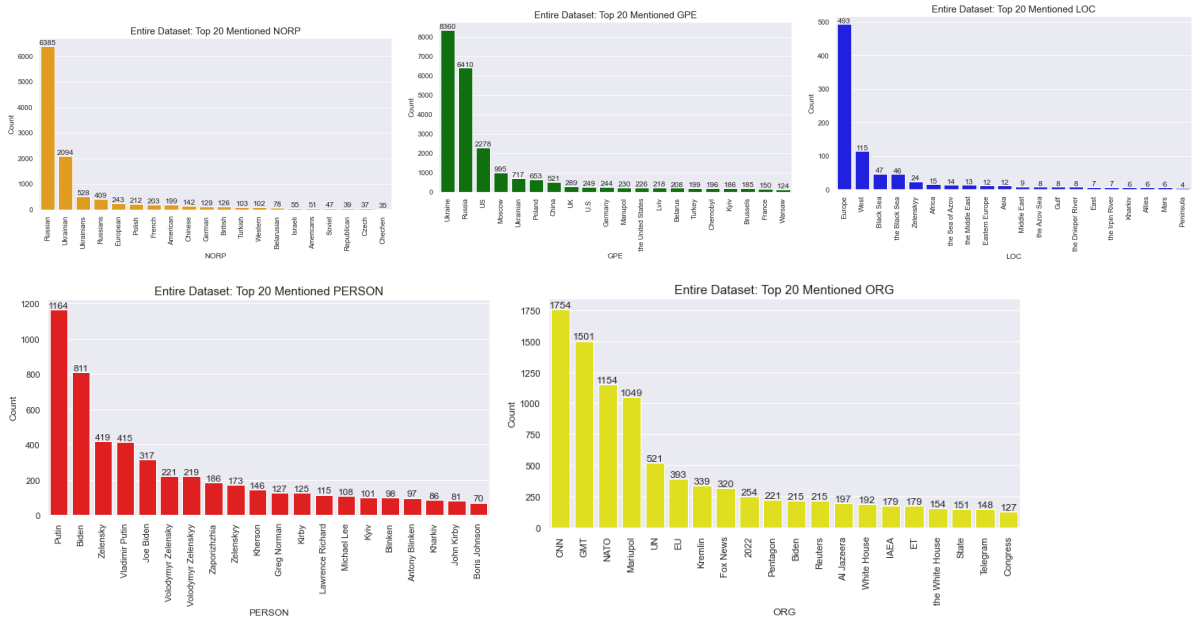
Basically, the challenging part of the OCR process is image preprocessing. Besides, the OCR package, pytesseract, was well-documented and easy to use. We found no trouble using it.

3. What can we say from observing the distributions of the entities in the text? Are there any differences between the news outlets?

To answer this question, we have plotted the distribution of top 20 mentioned entities for each of the 5 entity types in the entire dataset combining three news outlets. We made the same bar plots for each of the three news outlets separately as well. In this question, we will

mainly focus on observing the 5 bar plots for the entire dataset, and we will talk about the differences between the news outlets in section 5.

Observation 1: Among all five types of entities for the entire dataset, “Ukraine” is mentioned the most, which has 8360 times. It is easy to understand the reason it's mentioned way more than other entities that this is the most heated region since the Ukraine-Russia War has begun.



Observation 2: From the observations of the extracted Top 20 mentioned GPE (bar chart 1 attached below), except Ukraine and Russia, we could see that there are other countries being mentioned hundreds of times, such as the US, Poland, China, and UK. Therefore, the news paid much attention to these countries’ actions and their opinions, which meant those countries also played important roles in the Ukraine-Russia conflict.

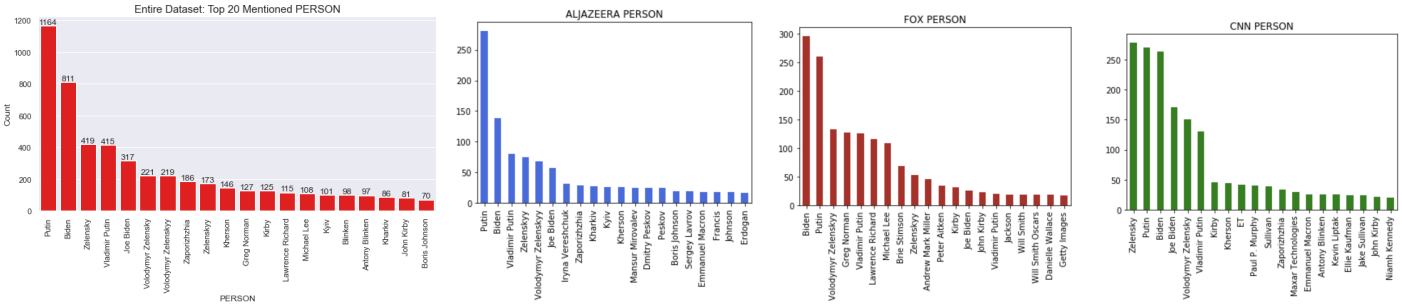
4. How difficult was the geoparsing? Did you encounter any issues while performing this step?

We used GeoPy to parse the geolocations. When we used Nominatim to get the geocode, we did not verify the permission of SSL, and we spent some effort on solving this issue. While running the geolocator, the connection broke several times, and sometimes returned status code 500, which we could not locate the error. Thus we set a timeout in each loop, and it took us several hours to process all of the locations that appeared in the news. Overall GeoPy is not hard to use, but sometimes the error message does not provide enough information for us to solve the problem, so we had to try out several times.

Also when we were drawing the dynamic map for Ukraine, it was hard to identify whether the location is in Ukraine or not, so we made a box border, which took the furthest

longitude and latitude of the Ukraine border, thus in our map, some of the locations are outside of Ukraine.

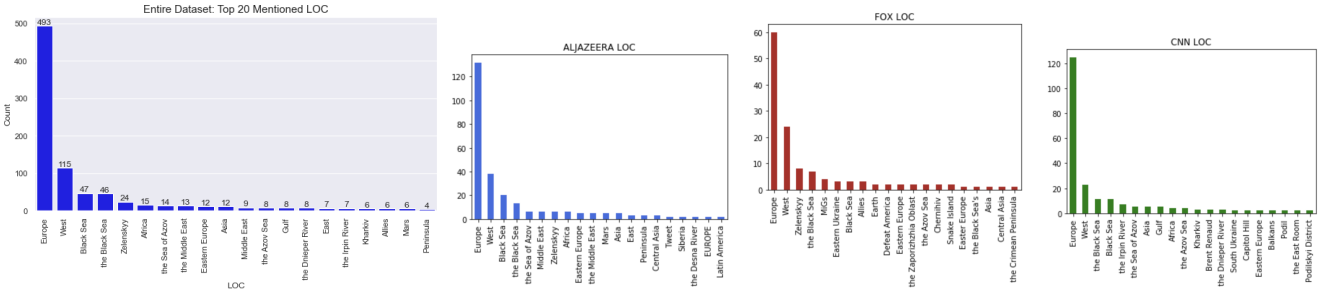
5. **If you identified some interesting differences between the news sources, please describe them in detail and include the bar-charts in the report.**



Observation 1: From the Top 20 PERSON, Volodymyr Zelensky, president of Ukraine, Vladimir Putin, president of Russia, and Joe Biden, president of the US are the top 3 persons being mentioned. It's understandable why Zelensky, president of Ukraine and Putin, president of Russia are mentioned so many times. On the other hand, we also noticed a high occurrence of Biden. After analyzing the owners and locations of three news sources, Aljazeera, CNN and Fox, we knew one of the reasons for the high occurrence of Joe Biden, president of the U.S. Aljazeera is English's standard news owned by the government of Qatar. From the chart "ALJAZEERA PERSON", we could see that the number of Biden and Joe Biden is much less than Putin. However, FOX and CNN are new sources from the US, so the occurrence of Biden or Joe Biden is almost the same with Putin and Zelensky.

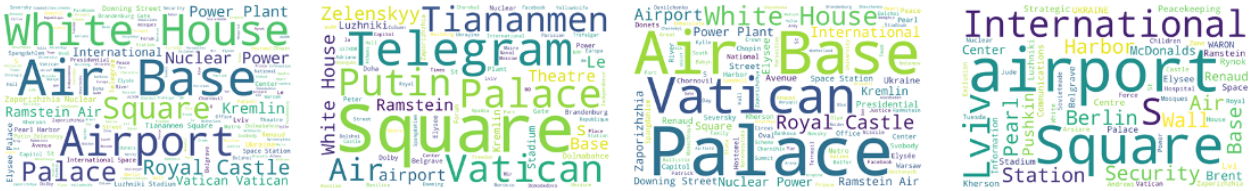
Observation 2: We noticed there were a lot of names that we were not familiar with in the FOX and CNN's bar charts, but most of the person names in Aljazeera were the world's politicians that we knew. Driven by curiosity, we searched for unfamiliar names and found out most of them are CNN's or FOX's writers. We realized this situation might be caused by the fact that the news images from FOX and CNN were including their writers, but the news images from Aljazeera were not containing the writer's names. So, it could have been better if we found a way to remove the writer's names to get a more accurate result.

Observation 3: From the Person's bar charts of the entire database, Aljazeera and CNN, except Fox, some names do not belong to a specific person, such as Kyiv which is the capital name of Ukraine, and Kherson, a city in Ukraine. In fact, for people who don't know Ukraine much, it is hard to determine whether Kyiv and Kherson are city names or not. Also, spaCy, the name recognition tech we used, wrongly regarded these names as person's names when it analyzed Aljazeera and CNN sources but not FOX. Considering the news' readability, FOX news may explain in their content that the name, Kyiv, is a city name rather than a person's name. Thus, by parsing near words or the same paragraph, spaCy could know that the name is a city name.



Observation 4: From location bar charts, Europe, West, Black Sea/ the black sea, Zelenskyy, Africa are Top 5, we all know that two of them, West and Zelenskyy are not locations. Besides the wrong locations, the top 3 of all data are Europe, Black Sea/ the black sea, and Africa; the top 3 of FOX are Europe, the Black Sea, and Eastern Ukraine; the top 3 of CNN are Europe, the Black Sea/Black Sea and the Irpin River. Three news sources extremely focus on Europe because this is related to the main topic of the Ukraine-Russia conflict. The black sea is also an important location near both Ukraine and Russia. CNN also focused on mentioning the Irpin River where Ukraine's military successfully blocked Russia's from entering Ukraine's capital.

Observation 5: After using spaCy to perform the Named Entity Recognition in the text we extracted, we divided all content into 10 categories. In terms of FAC, from the word-cloud we can see that "white house", "air base" and "airport" are cared about most, from all three news sources. However, for the individual news factory, Aljazeera mentioned telegram and Square for most of the time, CNN focused on palace and the Vatican while Fox cared about airport and Square.



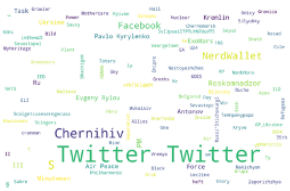
Total Aljazeera CNN Fox

About the EVENT category, there are more similarities than that in the FAC part. "World War " is the most popular phrase in the news, and all three news factories were reminded of World War II and the Cold War.

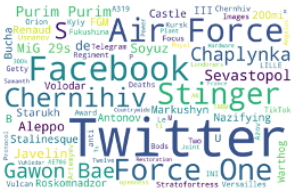
Total Aljazeera CNN Fox

As for the PRODUCT category, we can find that Twitter is the most popular social media when people over the world discuss matters about Ukraine and Facebook is an important information resource for the news companies as well.

Total



Aljazeera



CNN

Fox

6. Conclusions

In conclusion, we had successfully implemented Optical Character Recognition, Named Entities Recognition, as well as geoparsing and generating dynamic maps of the geolocated entities over time. All of the packages are way more easy and convenient to use than the previous projects. They are well-documented and bug-free when we are using them. The whole project was interesting and we found it useful for our future careers. Other than the techniques, we've also learned more about geopolitics and news reports. Overall, it was a joyful learning experience.