Group: Seize the Data
Building Visual Apps to Explore Current World Events from News Sources using Data Science:

Introduction

While taking on individual roles as strategic analysts, our group was tasked with examining current geopolitical events surrounding the war in Ukraine. Due to limited resources provided at hand, our group depended heavily on a variety of analytical tools such as Optical Character Recognition, Named Entity Recognition, and Geo-parsing to uncover details within each media file. The OCR tool aided in-text discovery, Named Entity Recognition assisted in locating text patterns, and Geo-parsing supported in pin-pointing geolocation information of each news event.

Optical Character Recognition

In order to extract text from images using Google Tessaract, the images had to be split as Tessaract supports a maximum height of 32,767 pixels. Using the OpenCV Python library, we split the image into several image parts of 32,500 pixels or less. Once this was done, extracting texts from the image files was a straightforward process using Tesseract. Installing Tesseract was straightforward both on a local machine using Homebrew and in Google colab. Using it to extract text from the images was also simple enough using a "pytesseract.image_to_string()" function that takes the path to the folder that contains the images as an argument, and with some further string manipulations, join the extracted words via " " and split the text into lines wherever a "/n" character was found after which we were able to print the text from each image into a csv file. This HW submission has a .py file named "Tesseract.py" that can run on a local machine if the code is updated with the correct path to a folder on a local machine that has .png images that require text extraction. However the actual code that was used to extract text from the images for this HW is in a Colab notebook named "Tesseract_Text Extract to CSV.ipynb". The notebook mounts the 3 pre-processed image folders which were uploaded to Google drive. For each folder there is a corresponding cell in the notebook that loops through the images in the folder and extracts text from each image into a csv file named after the image name with a .csv extension. However, we discovered that it was not perfectly accurate. Although Tesseract was able to mostly make out what each line of text read, there were special characters and misspellings in the text. This made it difficult to parse through each individual article as we used dates and times as tags for when an article begins and ends. Since the news sources were all from the month of March, we were able to transform each misspelling of "March" to be spelled out correctly. In order to remove advertisements, the script looks for lines that read "Advertisement" or something similar depending on the news source. Once this is read, the subsequent lines are skipped until another date tag is read to indicate the beginning of a new article. Using this approach, menu items at the top of a page were also able to be skipped.

Named Entities Recognition

Excluding March 1st through March 11th, the Named Entities Recognition resulted in the graph distributions towards the bottom of the report. There, the graphs and their analyses are broken down in more detail specific to each category and news source. Overall, however, what is interesting to note about the three distributions on a larger scale is that they are fairly similar when it comes to the shape of the distributions themselves. Most of them are heavily skewed trailing off to the right, which is no surprise due to the fact that certain topics are very popular in the news while other issues may be reported on less frequently.

The differences primarily come down to the content, but even there, the differences are fairly small as much of the news and the output is centered around the Ukraine-Russia conflict. Many of them tend to focus on Biden, Putin, Zelensky, Russians, Ukrainians, etc., and they are mentioned in a relatively similar number of occurrences. Additionally, Aljazeera reported on U.S. topics and issues less frequently than CNN and Fox, which is explained by the fact that Aljazeera is the only international news channel while the other two are both U.S. channels. There are differences in number of occurrences on some level, such as Fox mentioning fewer NORP related entities overall, but by just looking at the distributions it is hard to tell whether one news source reports drastically differently from another. They are fairly neck and neck.

Geolocation

Geoparsing: Difficulty And Issues While Performing This Step

Geopy was not difficult to set up or use, we decided that the best way to map the locations was to output a csv file with the dates the location was mentioned and the corresponding coordinates. In order to retrieve coordinates, geopy needs an input which can be as specific as an address or as generic as a city name. The NER step produced 3 csv files with approximately 25000 locations combined. We had to create a list of each unique location, later this was used to create a dictionary with each location as the key and the coordinate as the value. This successfully allowed us to avoid being blocked by geopy's API. Nonetheless, we did experience some issues with geoparsing due to the location names produced by the NER step and we experienced issues with date outputs due to the date format produced by the OCR step.

Out of the 25000 locations most worked without issue, however, some included names of presidents, journalists, hashtags, misspellings, multiple locations in one output which produced errors when trying to run them through geopy. A few examples include "America Biden", "Baltics Poland Romania", and "El Salvador's". Some of these issues were very specific and multiple lines of code were written to specify exactly how to fix the errors which was the most tedious step. Additionally, we had an issue with the date format that was produced from the OCR step. This resulted in dates that were not consistent across all sources making it difficult to clean. In the end we had about 300 rows of data with the date "March 40" and we excluded those from our map plot. Ultimately, it was difficult to perform meaningful quality control testing because certain locations could have also been parsed by NER from names and there is no way to know from context such as "Wallace" and "Michelle". Another example was "TWEET," were they actually talking about the location or was it in the context of Twitter? However, there are several instances of this such as "Again" and "Spotify."

Plotting Geoparsing output and creating a World and Ukraine Map

The world and Ukraine maps can be found in a colab notebook named "World and Ukraine maps_HW3.ipynb". The maps are already rendered using Plotly Express and Mapbox and can be readily viewed and manipulated. **The notebook cells should NOT be run as this will damage the animation since it relies on data from a Team member's mounted Google drive**. The cells titled "World Map" and “Ukraine Map” use the Geoparsing output that is mounted on a Team Member’s Google drive and loaded into a pandas dataframe in the notebook.

The learning curve for Plotly and Mapbox was a bit challenging. First we experimented with Plotly alone and plotting a single instance (locations from a single day) on top of a Geojson file that contains a map of the world, but we quickly realized that Mapbox’s animation functionality combined with Plotly Express can result in an animated map for the all the locations throughout the entire month that presents a more elegant solution than taking snapshots of the plot of each day and combining them into a gif. In order to use Mapbox one of our Team Members had to register and create a Mapbox account to receive a token that allows the user access to Mapbox’s maps.

The code to create an animated map that uses the date from the Geoparsing output files as the animated frame in the Plotly and Mapbox function was simple enough. To create 2 maps, one for the world and the other for the Ukraine, we plotted the extracted locations at Country level in the world map, and in the Ukraine map we plotted a filtered dataset of longitudes and latitudes within Ukraine at the GPE and LOC level. Centering the Ukraine map on Ukraine required some research until we finally found that using “fig.update_mapboxes(center_lon=31.2718321)” and “fig.update_mapboxes(center_lat=49.4871968)” functions (Ukraine’s longitude and latitude), and adjusting the zoom level of the map ensures the Ukraine map centers on Ukraine.

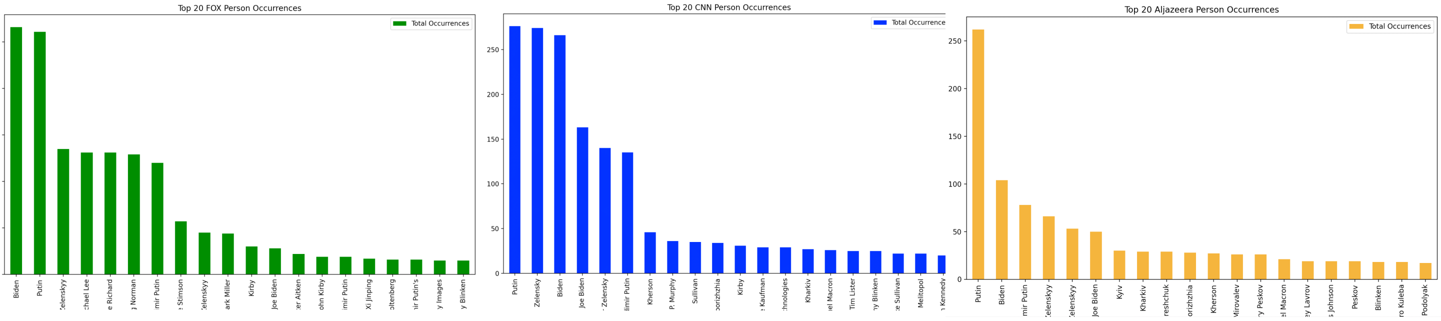
We discovered a couple of bugs in Mapbox as we realized in our first trial that not all the locations were being plotted beyond the first day as it seems to use the locations of the first day as a reference and truncates the locations in subsequent days so that it only plots a location if it was included on the first day. Through our research we found that this is a documented bug in Plotly https://github.com/plotly/plotly.py/issues/2259 and the workaround to resolve this issue was to create a day 0, with all the possible locations assigned a dummy lat long and add it to the Geoparsing output dataset. That way in all subsequent days starting day 1 the locations were plotted correctly.

The second bug we found was that on some days the animation would show previous days in addition to the day on which the slider had been set (for example on day 5 you could see markers from day 4). In addition pressing the play button resulted in a choppy animation as there was a delay in the map visualization as the animation transitioned from one day (frame) to the next, and the only way to view the animation was to manually move the slider from one day to the next. Our insight was that since the documented bug describes an issue related to the colors of the markers we decided to attempt to remove the colors of the markers from the animation and make the marker color uniform for all locations (blue) since we realized the marker colors are redundant (you can see the location of each marker on the map, and hovering on a marker shows you its associated data). This proved to be successful in removing the error and the map finally displayed the locations each day exactly as per the Geoparsing output dataset. It also resolved the issue of the choppiness of the animation and finally pressing the play button resulted in a smooth animation especially when we controlled the speed of the animation using the layout.updatemenus() funcion.

Additional questions:

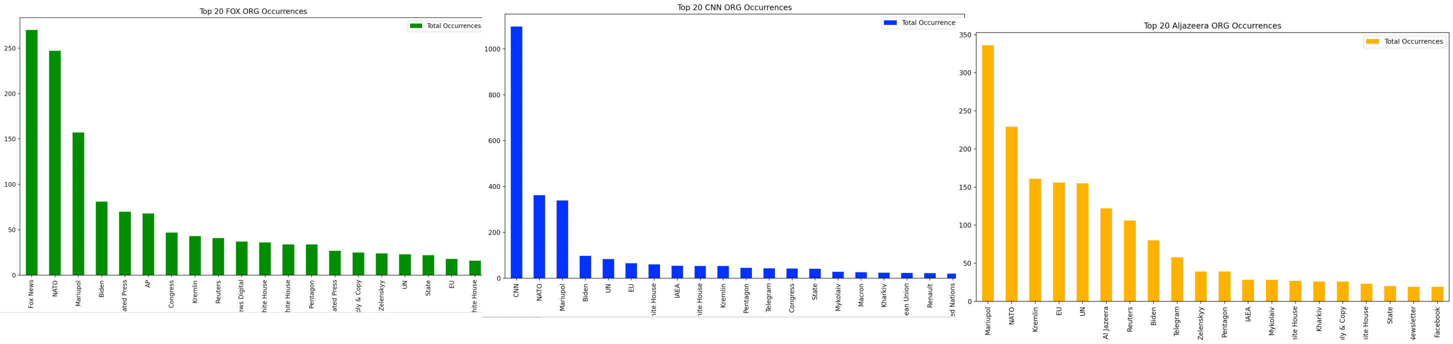
Below are the graphs from Step 4. The font is small, so green represents FOX, blue represents CNN, and orange represents Aljazeera. Finally, in red, are all of the five entities and their respective top 20s across all categories.

Top 20 Person per News Source



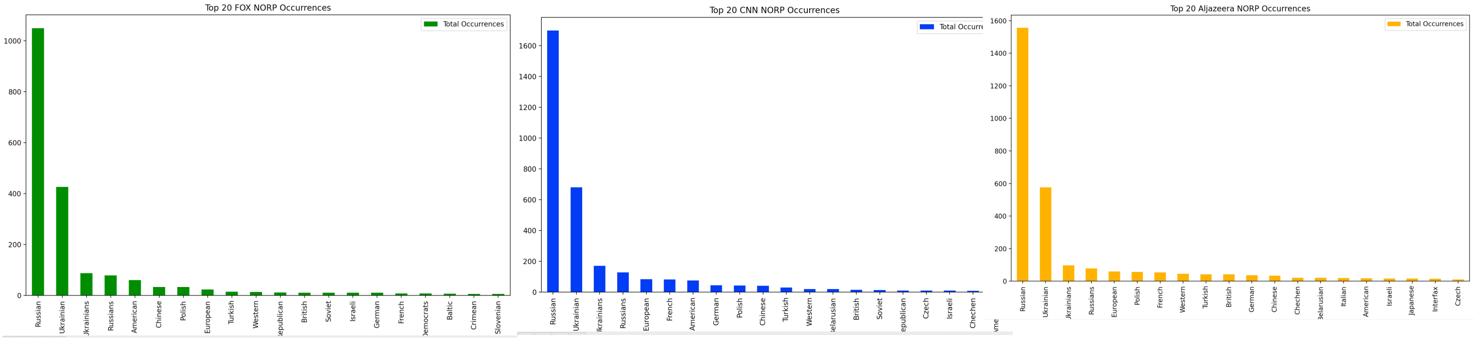
When looking at the PERSON category specifically across all three news sources, it appears that almost all three news sources heavily focus on the key players involved in the current Ukraine-Russia conflict. Namely, Biden, Putin, and Zelensky. Those are the top reported amongst each news source. It is interesting to note, however, that Aljazeera is really heavily focused on only one person, Putin, while the others have at least two to three other people that are mentioned with relatively the same amount of frequency or still fairly high compared to the rest. Knowing that Aljazeera is the only channel of the three that is international, this could explain why Aljazeera mentions many people less frequently: they likely are less focused on U.S. politicians. Based on a visual analysis of the five entity categories, the PERSON entity has the most variable distributions among the three news sources.

Top 20 Org per News Source



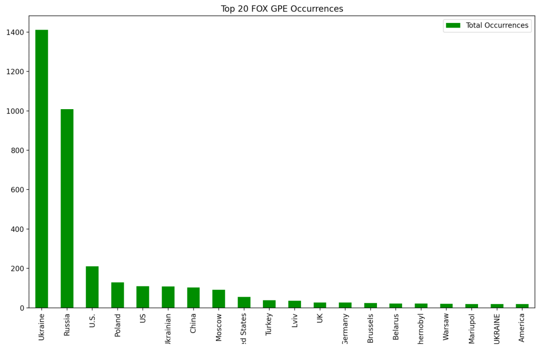
When looking at the ORG category across all three news sources, specifically, it appears that almost all three news sources here also heavily focus on the key players involved in the current Ukraine-Russia conflict. However, here, what is interesting is that Fox and CNN mention their own organization as the top most mentioned organization while Aljazeera does not mention it’s own name the most. All three mention NATO and Mariupol as one of their highest mentions. It is interesting to note, however, that CNN is really heavily focused on one organization, itself, while the others have at least two to three other organizations that are mentioned with relatively the same amount of frequency or still fairly high compared to the rest.

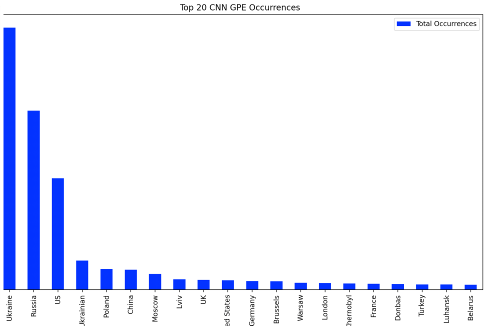
Top 20 NORP per News Source

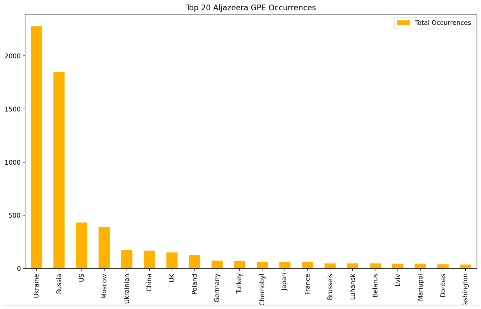


When looking at the NORP category across all three news sources, specifically, it appears that almost all three news sources here have very similar distributions on which NORP groups occur most frequently within their content. They are also heavily focused on the key players involved in the current Ukraine-Russia conflict. All three mention Russian and Ukrainians several times and are the highest mentions. It is interesting to note, however, that the rest of the distribution for each of the three news sources tails off and is relatively small across the other categories when compared to the mentions of Russia and Ukraine.

Top 20 GPE per News Source

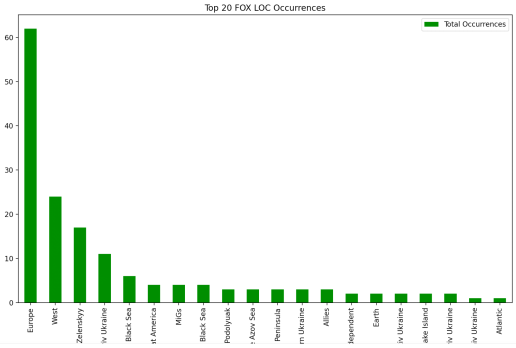


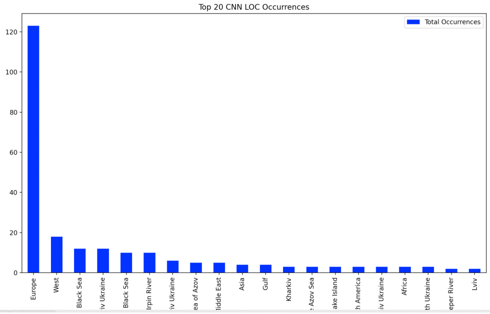


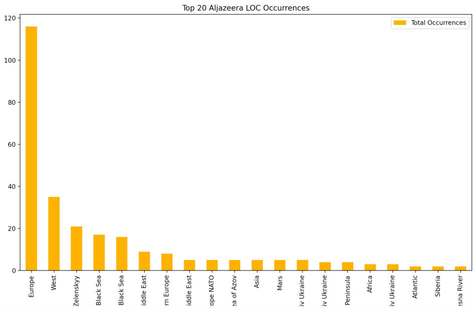


The most-mentioned GPE entities have fairly similar distributions across the three news sources. All three sources mention Ukraine and Russia with a much higher frequency than other GPE entities, which would not surprise anyone due to the intensity of the Russia-Ukraine conflict over the past several months. While all three distributions are fairly similar, the most notable difference is that CNN has a much higher frequency of U.S. mentions than Fox and Aljazeera. This would suggest that CNN discusses domestic news more often than the other two news sources. Aljazeera is actually an international news channel, so it makes sense that their content would focus less on the U.S. As for Fox, the NER process actually found some mentions of “U.S.” and some mentions of “US,” and these were categorized as separate entities. Therefore, if we combined the two bars together, then Fox’s mentions of the U.S. would be slightly more comparable to CNN’s.

Top 20 LOC per News Source

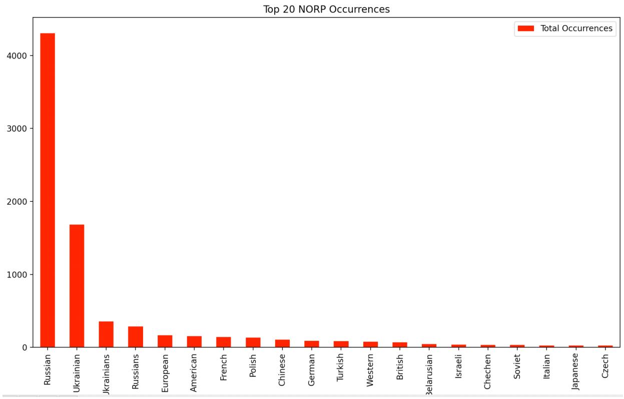


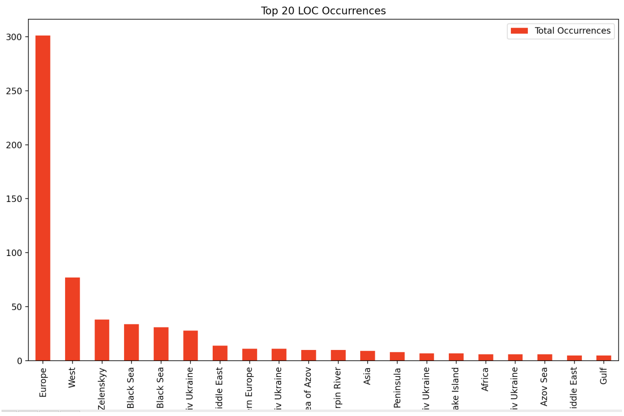


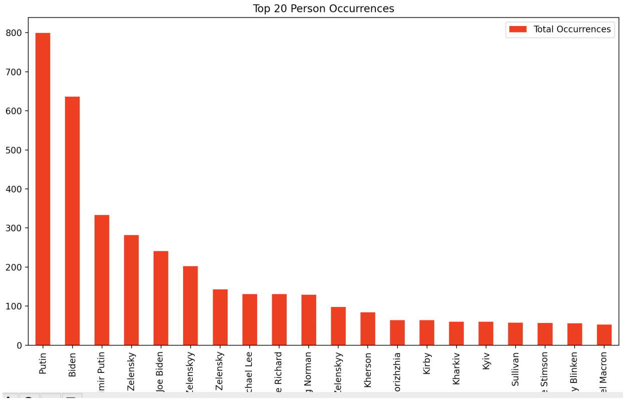


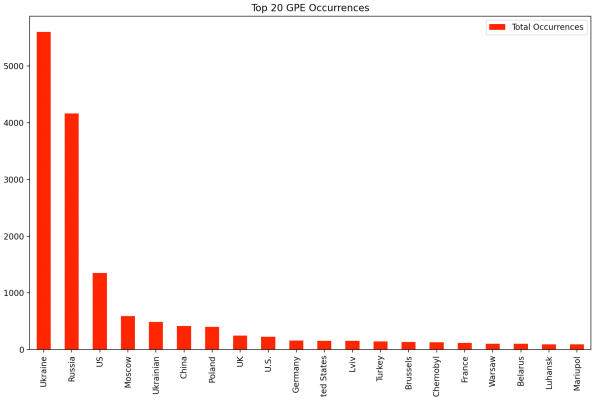
The top-mentioned LOC entity across all news sources was Europe, which was mentioned much more often than the others. CNN has the most drastic drop in frequencies after Europe, while Fox and Aljazeera trail off more gradually to the right. Interestingly, The third most mentioned LOC entity for Fox is Zelensky, which appears to be an error in the NER process since Zelensky is in fact a person. In general, the distribution of top LOC entity mentions is fairly similar across all three news sources with few big differences.

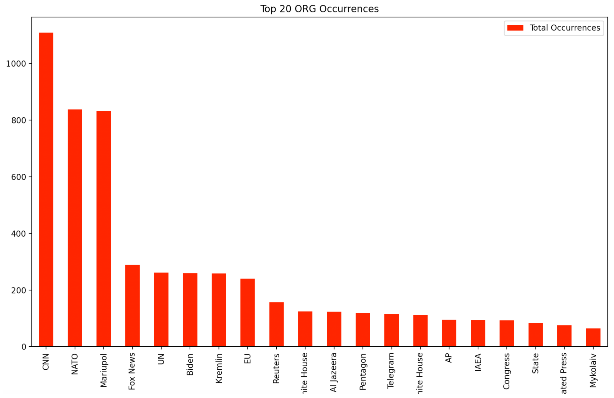
Top 20 Across All 5 Categories











Across the five entities that were analyzed for the purpose of this project, we see huge variability in the distributions of top-mentioned entities. The PERSON entities have the highest frequencies of mentions, with the 20th top PERSON having around the same frequency as the 3rd top NORP. The ORG entities are not far behind the PERSON entities, so it seems that in general news sources report most often on people and organizations, and they discuss locations, geopolitical entities, and nationalities/religious/political groups less frequently. Clearly this data also demonstrates the lack of variety that we consumers sometimes feel when we turn on the news and the same issues are being discussed over and over again.