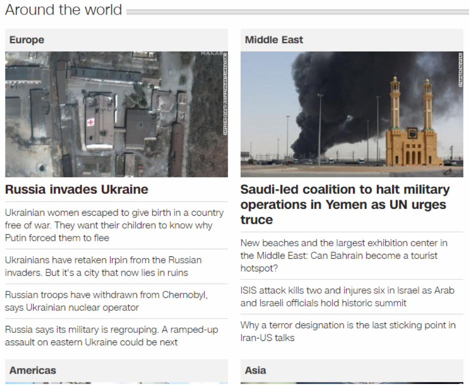
Homework: Building Visual Apps to Explore Current World Events from News Sources using Data Science
Due: Friday, April 29, 2022, 11:59:59pm PT

1. Overview



Around the world

Europe

Russia invades Ukraine

Ukrainian women escaped to give birth in a country free of war. They want their children to know why Putin forced them to flee

Ukrainians have retaken Irpin from the Russian invaders. But it's a city that now lies in ruins

Russian troops have withdrawn from Chernobyl, says Ukrainian nuclear operator

Russia says its military is regrouping. A ramped-up assault on eastern Ukraine could be next

Middle East

Saudi-led coalition to halt military operations in Yemen as UN urges truce

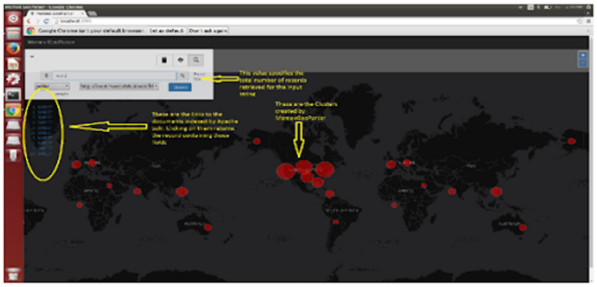
New beaches and the largest exhibition center in the Middle East: Can Bahrain become a tourist hotspot?

ISIS attack kills two and injures six in Israel as Arab and Israeli officials hold historic summit

Why a terror designation is the last sticking point in Iran-US talks

Americas

Asia



Google Chrome isn't your default browser. Set as default Don't ask again
Search

This value specifies the total number of records retrieved for the input string

These are the links to the documents indexed by Apache Solr. Clicking on them returns the record containing those fields

These are the Clusters created by MemexGeoParser

In the third assignment, you will create a set of visualizations and plots to illustrate some of the current World events. This may include maps of the specific locations of interest, illustrations of international relations, etc. You will explore and analyze the entities appearing in the news, as well as their relation. Additionally, you will compare and contrast various news sources and identify the key similarities and differences between them.

This time, you take the role of a strategic analyst, and your task is to analyze the current events in Ukraine. It is a war time; the internet is down, and you don't have the proper access to the news sources. However, you have been provided the screenshots of the news coverage by a foreign intelligence agency. Your superiors are asking you to extract the data from the images and to analyze the events as they unfold day by day. You are one of the few people who can do it successfully. World peace depends on you.

2. Objective

The objective of this assignment is to build upon the great analytic work and data science work you did exploring semantic forensics and scientific literature. This time, we will shift our focus from scientific literature to current events. You will use the technologies such as OCR, Named Entity Recognition, GeoParsing and the visualization techniques to explore and interact with your data.

You have been provided the screenshots of the news coming from three sources: CNN, Fox News and Aljazeera. Each screenshot contains all the news from a single day from March 2022, sorted chronologically and stacked on top of each other. The images are fairly "tall", with a standardized width of 1000px and height ranging from 20000px to 75000px. Before doing any OCR, you should consider preprocessing the images first.

Some of the OCR packages might not be able to read such large images and splitting them in a few smaller chunks might be necessary.

To analyze the news, you should perform the OCR on the provided images. Please use Python and you can use any OCR software/package you prefer. There are many OCR tools out there, such as Tesseract with a nice Python wrapper at https://pypi.org/project/pytesseract/. Make sure to keep the extracted text organized by date and the news source either in a json object, pandas DataFrame, or the text files. You will notice that there is some extra text you probably don't need, such as menus, ads… You should remove those. It is up to you how to approach this: you can compare the text from multiple days and detect what sentences are repeating, or you can cut the top/bottom part of each image where the menus are usually located, or you can do this manually by observing what parts of the image are likely menus/ads and remove that text from the extracted text… If you come up with some other method, feel free to use it.

With all text extracted and organized, you should move on to the Named Entity Recognition phase. For this, you can use any package you prefer. SpaCy works very well and it is nicely documented: https://spacy.io/. If you are familiar with some other NER tools, feel free to use it. The following entity types are important:

PERSON:	People, including fictional.
NORP:	Nationalities or religious or political groups.
FAC:	Buildings, airports, highways, bridges, etc.
ORG:	Companies, agencies, institutions, etc.
GPE:	Countries, cities, states.
LOC:	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT:	Objects, vehicles, foods, etc. (Not services.)
EVENT:	Named hurricanes, battles, wars, sports events, etc.
DATE:	Absolute or relative dates or periods.
TIME:	Times smaller than a day.

When you extract the entities, you should start analyzing them. You want to know what PERSON is mentioned the most. What about the second most mentioned? Here, you should plot a distribution of the top 20 mentioned people in the entire dataset. A simple bar plot illustrating the number of mentions is the best way to do it. The same should be done for the: NORP, ORG, GPE and LOC. And what about the differences in reporting between the news outlets? Do they equally often mention various people, organizations, or countries? Are the distributions different? Here you should plot the same distributions for each news outlet side-by-side. You should get three distributions (bar plots) for each named entity type (one for each news outlet). Look at the distributions. Are they different? Can they tell us something about the differences in reporting between the three outlets? Note: The Fox News dataset is missing days between March 1st and March 11th. When comparing the news sources, you should consider this and exclude those dates from CNN and Aljazeera as well.

Next, you want to plot the entities that were mentioned in the text on the map. Here, we are interested only in GPE and LOC entities, as they could be geographically referenced. To find a location of an entity, you need to perform the geoparsing. For this purpose, you can choose any tool you like. For example, Python GeoPy is a well-documented package for geoparsing: https://geopy.readthedocs.io/en/stable/# It uses external APIs such as OpenStreetMap. Some of those external APIs have rate limits, so be careful when using them: https://geopy.readthedocs.io/en/stable/#geopy-is-not-a-service. To reduce the number of queries, you should query the unique GPE entities only once and save them somewhere.

When you have all the relevant entities geolocated, you should locate them on the map. The maps should be dynamic, showing the entities as they appear in the news day-by-day. You can use packages such as Plotly: https://plotly.com/python/maps/, or any other service you like (e.g. D3js: https://d3js.org/). You should make maps with two “zoom” settings. In the first setting, you should show the entire World, where we can observe the activities/mentions of the larger entities, such as countries. In the second setting, you should zoom onto Ukraine and locate the smaller entities that appear in the news such as cities.

Belarus
Russia
Warsaw
Media reports of Russian troops crossing from Belarus into Ukraine
Poland
Separatist-claimed areas
Krakow
Kyiv
Lviv
Kharkiv
Ivano-Frankivsk
Ukraine
Slovakia
Luhansk
Dnipro
Donetsk
Moldova
Hungary
Chisinau
Rostov
Mariupol
Romania
Odessa
Separatist-held areas
Bucharest
Krasnodar
Key
◉: Media reports of missile strikes
Russian troops arrive in port cities of Odessa and Mariupol, according to Interfax Ukraine

The assignment specific tasks will be specified in the following section.

3. Tasks

1. **Getting the data:**
   Download the screenshots from https://drive.google.com/file/d/1XhV7EUPRAKlzKHYpjR_LN1W3IjP8l-5U/view?usp=sharing

   Split and organize the images if needed.

2. **Optical Character Recognition:**
   Use the OCR software/package to perform the OCR on the extracted images. You can use Tesseract (https://pypi.org/project/pytesseract/) or any other package you prefer.
3. **Save the extracted text:**
   Save the extracted text into a preferable format that will allow you to fetch the text again when needed. It can be saved as a JSON, or as a DataFrame and then pickled, or in the set of textual files properly named and organized in the folders

so you know the date and the news source of each text. Make sure to showcase that your script can remove the unnecessary extra text from the menus or the ads.

4. **Named Entities Recognition and Analysis:**
   - Perform the Named Entity Recognition on the text you extracted. You can use https://spacy.io/ or any other NER package you are familiar with. Get all the: PERSON, NORP, FAC, ORG, GPE, LOC, PRODUCT, EVENT, DATE and TIME from the text. If you don't use SpaCy, and the NER software you use does not have the exact same types of entities, feel free to replace them with the appropriate types from the software you use.
   - Plot a distribution of top 20 mentioned PERSON, NORP, ORG, GPE and LOC in the entire dataset. A simple bar plot illustrating the number of mentions is the best way to do it.
   - Compare the distributions of the entities between the news sources. Note: The Fox News dataset is missing days between March 1st and March 11th. When comparing the news sources, you should consider this and exclude those dates from CNN and Aljazeera. At the end, you should get 15 bar charts, one for each required entity type and each news source.
   - Look at the distributions. Are they different? Can they tell us something about the differences in reporting between the three outlets? Please write this in your report
5. **Geolocation:**
   - Perform the geoparsing on GPE and LOC entities. You can use Python GeoPy: https://geopy.readthedocs.io/en/stable/# or any other geoparsing tool you are familiar with. Be cognizant of the potential API limits and make sure to optimize for the number of queries.
   - Make two dynamic maps that display the geolocated entities over time (day-by-day): a) The World map, where we can observe the activities/mentions of the larger entities, such as countries; b) Ukraine map, to show the smaller entities that appear in the news such as cities.
     You can use packages such as Plotly: https://plotly.com/python/maps/, or any other service you like (e.g. D3js: https://d3js.org/). A simple approach would be to take the snapshots of a map from each day and compile them in a GIF. A more advanced approach would be to use tools such as mapbox that integrate with plotly. Here are some ideas:
     -https://towardsdatascience.com/simple-plotly-tutorials-868bd0890b8b
     -https://towardsdatascience.com/how-to-create-animated-scatter-maps-with-plotly-and-dash-f10bb82d357a
     -https://plotly.com/python/maps/
     Any approach you choose will be considered valid.

6. **Extra Credit:**
   - For the extra credit, you should make and host a webpage (e.g. on GitHub pages or some other service) that can showcase your work: entities distribution, dynamical maps, as well as the additional analysis and explanations.

4. Assignment Setup

4.1 Group Formation

You should keep the same group from your previous assignments. There is no need to send any emails for this step, unless there are changes in the groups.

5. Report

Write a short 4-page report describing your observations. Please answer the below questions:

1. How difficult was performing the OCR on provided images? Was there any step in the OCR process that was particularly challenging?
2. What can we say from observing the distributions of the entities in the text? Are there any differences between the news outlets?
3. How difficult was the geoparsing? Did you encounter any issues while performing this step?
4. If you identified some interesting differences between the news sources, please describe them in detail and include the bar-charts in the report.

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 11:59:59 pm PT*** on the specified due date, via D2L > My Tools > Assignments (https://courses.uscden.net/d2l/home/22303). A team can submit multiple times, but only the last submission counts. Anyone from a team can submit. However, we suggest designating one person to submit.

- All source code is expected to be commented, to compile, and to run.
- Also prepare a readme.txt containing any notes you'd like to submit.
- If you used external libraries, you should include those necessary files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.
- Save your report as a PDF file (TEAM_NAME_DATAVIZ.pdf) and include it in your submission.
- Compress all of the above into a single zip archive and name it according to the following filename convention:

  TEAM_NAME_DSCI550_HW_DATAVIZ.zip

  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.
- If your data is too big and exceeds the D2L file limit of 2GB: 1) upload your data to Google drive, 2) include the links to the data in a README file, 3) compress the report, README file and the code and upload it to D2L. You do not need to upload the images we already provided. A pointer in the readme.txt or a comment in the code that explains how to ingest them, should be sufficient.

Important Note:

- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.

- Again, please note, a team can submit multiple times, but only the last submission counts. **To avoid confusion: designate someone to submit.**

6.1 Late Assignment Policy

- -10% for every day or part thereof