BALANCE OF POWER

An Effective tool to evaluate and contrast the balance of power in countries

Project team 2

Alex Harrell(atharrel)

Sachin Ahuja(sahuja3)

Bharat Mukheja(bkmukhej)

Pranav Firake(ppfirake)

Sneha Shah(smshah4)

Project Guide

Michael Kowolenko

Computer Science Department,

North Carolina State University, Raleigh NC

Table of Contents

Sr.	Name	Page Number
1	**Abstract**	Page 3
2	**Problem Statement** (2.1) Mind Map	Page 3
3	**Approach** (3.1) Generating the Truth Table (3.2) NLP Aggregate Scores (3.3) Weighted Order Decision Making to Evaluate Scores (3.4) Front End User Control	Page 5
4	**Tech Stack**	Page 9
5	**Web application Development**	Page 10
6	**Discussion**	Page 15
7	**Conclusion and future work**	Page 16
8	**Appendix**	Page 17
9	**References**	Page 17

1. Abstract

The assignment given to us was to somehow analyze some kind of balance of power. Upon receiving this task we decided to analyze the relative balance of power in the middle east, using middle-eastern countries, several surrounding countries, the United States, and Russia as our subjects. We also decided that we would have five primary metrics by which we would measure power; Demographics, Economy, Geography, Infrastructure, and Society. Using these five categories, the twenty-two countries that we selected, our truth table generated from World Factbook data, and our Newspaper Articles we were able to successfully create a system which allows the user to dynamically evaluate the balance of power of any combination of the twenty-two countries using a balance of the five primary metric groupings that they determine.

2. Problem Statement

Our aim was to assess balance of power in middle eastern countries. The middle-east has been a volatile region, not just recently, but throughout its past. The entire region has been plagued by various religious conflicts for hundreds of years making an interesting subject for our study on the balance of power. This region also has a high dependency on exporting goods (primarily oil but also some other goods depending on the country) which means that its economy is tied to many international economies where oil is highly valued, hence why we have included the United States and Russia in our study. Given the regions volatility, foreign codependence, and unique geography we created the mind map (Figure 1) to get our project started.

(2.1) Mind Map

The major factors that we researched and believed would affect the decision of finding out the balance in power were

1. *Geography:* This included factors like the location of the country, its land and water , the natural resources it has.
2. *Government:* That is the type of government in the country whether it is autocracy , democracy , military power.
3. *Military:* On how strong the country is from the standpoint of military, which could be measured by the countries military expenditure and investments, factors like service requirements, etc.

4. *Economics*: This is one of the important criteria in answering the questions regarding the balance of power and mainly includes factors like GDP of the country, GDP growth rate, the unemployment rate, Inflation rate , the amount of exports and import the country does.
5. *People and Society:* The important factor in this category is the population of the country , other good factors that we felt that do affect the decision were the age structure, population growth rate, net migration rate, infant mortality rate, literacy rate as well as the major religion being followed by the people of the country and the ethnic groups as well.

So as we can see from the factors listed above some of these can be obtained and considered as structured data which can be directly measured in the form of numbers, like population, military expenditure, GDP, export, import etc while for some of the factors the data is unstructured for instance the government type, religion, ethnic groups and so on. So we have processed the data accordingly and have done analysis on the structured and unstructured data separately by generating the truth table(structured) and also calculating the NLP aggregate scores(unstructured) which is explained in detail in the later sections.

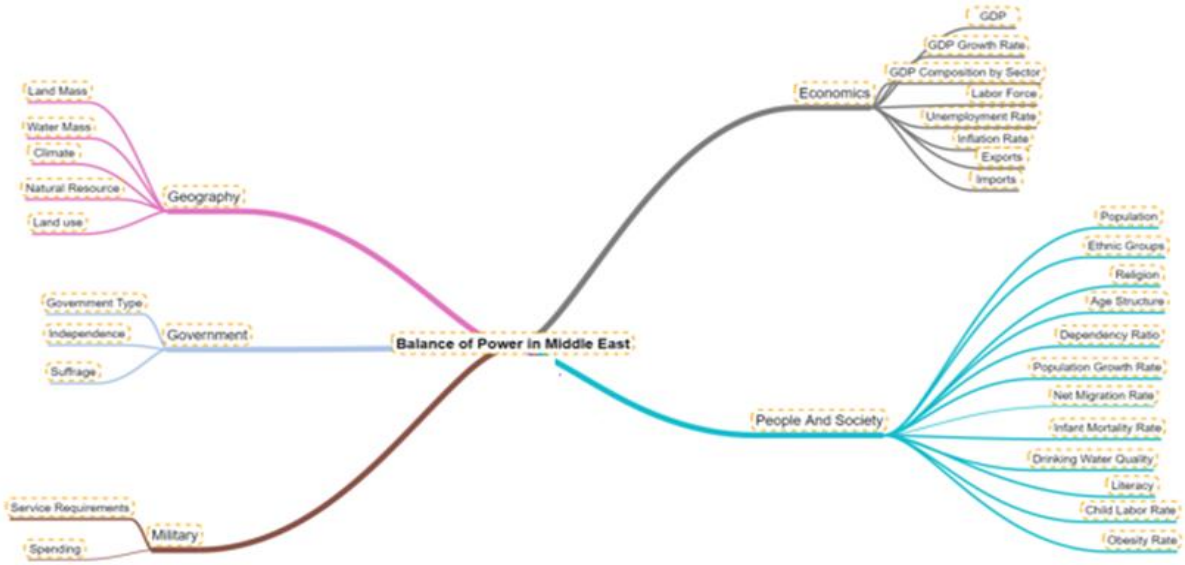


Figure 1: Mind Map

To provide specificity and context to an otherwise ambiguous statement, in this report we answer some of the core questions and build on them to reach an objective verifiable solution. A few of these core questions could be summarized as follows:

(a) What are the countries that are part of middle east?

(b) What factors are going to affect the power?

(c) Is a country dependent on another country for its resources?

(d) Where can we get the data from?

(e) How can we evaluate power using different attributes in the mind map?

(f) How to compare one country with another country?

The very first assumption that we make is to narrow down on middle eastern countries rather than taking all the countries. Hence, for the scope of this project, we will consider assessing Balance of Power for middle eastern countries and using weighted order decision making, will determine scores for these countries which will be an overall indicator of the assets that a country has. We will then compare these scores for decision making.

3. Approach

(3.1) Generating the Truth Table

To create our truth table, we designed a web-scraper using Python 3, Beautiful Soup and Regex to parse data from the CIA World Factbook. The World Factbook was a nearly ideal source for creating a truth table for our project as it has many metrics for every country which we selected, it is publicly available, and it is updated yearly (while this is not very frequently, it is very up to date considering many datasets that could be used for truth tables are several years old. Initially we scraped about 120 values (as seen in the scraper code) but we ended up deciding to take the most important attributes from each of our five attribute groups to make the decision process smoother, resulting in a total of thirty variables in the final model (not including NLP based variables. Along with the 30 numeric variables, we also scraped 6 non-numeric attributes from each country (text lists such as natural resources, potential natural hazards, important industries, etc) to use later in the process for NLP analysis. The resulting truth table was not actually one table, but rather six separate tables (one for each of our five metrics as well as one for the lists to be used for NLP processing later) which were exported into csv files to be used for further analysis.

(3.2) NLP Aggregate Scores

To include sentiment from newspapers, as well as, a variety of Lexis Nexis documents we created three aggregate scores. The three scores were natural resource strength, industrial strength, and instability. We decided on these because they covered several gaps in our existing truth table and made sense as far as the lists that we could extract from the initial scraping of the CIA Factbook as shown in the figure given below:

ba	['oil', ' associated and nonassoc	['periodic droughts', ' dust storms']	['petroleum processing and refining', ' aluminum smelting', ' ir
cy	['copper', ' pyrites', ' asbestos',	['moderate earthquake activity', ' d	['tourism', ' food and beverage processing', ' cement and gypsu
eg	['petroleum', ' natural gas', ' iro	['periodic droughts', ' frequent eart	['textiles', ' food processing', ' tourism', ' chemicals', ' pharmace
ir	['petroleum', ' natural gas', ' coa	['periodic droughts', ' floods', ' dust	['petroleum', ' petrochemicals', ' gas', ' fertilizers', ' caustic soda
iz	['petroleum', ' natural gas', ' phc	['dust storms', ' sandstorms', ' flood	['petroleum', ' chemicals', ' textiles', ' leather', ' construction m
is	['timber', ' potash', ' copper ore	['sandstorms may occur during spri	['high-technology products including aviation', ' communicatio
jo	['phosphates', ' potash', ' shale	['droughts', ' periodic earthquakes'	['tourism', ' information technology', ' clothing', ' fertilizers', ' p
ku	['petroleum', ' fish', ' shrimp', ' r	['sudden cloudbursts are common f	['petroleum', ' petrochemicals', ' cement', ' shipbuilding and re
le	['limestone', ' iron ore', ' salt', '	['dust storms', ' sandstorms']	['banking', ' tourism', ' food processing', ' wine', ' jewelry', ' cen
ly	['petroleum', ' natural gas', ' gyp	['hot', ' dry', ' dust-laden ghibli is a s	['petroleum', ' petrochemicals', ' aluminum', ' iron and steel', '
mu	['petroleum', ' copper', ' asbesto	['summer winds often raise large sa	['crude oil production and refining', ' natural and liquefied natu
pk	['arable land', ' extensive natur	['frequent earthquakes', ' occasiona	['textiles and apparel', ' food processing', ' pharmaceuticals', ' c
qa	['petroleum', ' natural gas', ' fish	['haze', ' dust storms', ' sandstorms	['liquefied natural gas', ' crude oil production and refining', ' ar
rs	['wide natural resource base in	['permafrost over much of Siberia i	['complete range of mining and extractive industries producin
sa	['petroleum', ' natural gas', ' iro	['frequent sand and dust storms', 'v	['crude oil production', ' petroleum refining', ' basic petrochem
su	['petroleum', ' small reserves o	['dust storms and periodic persiste	['oil', ' cotton ginning', ' textiles', ' cement', ' edible oils', ' sugar
sy	['petroleum', ' phosphates', ' ch	['dust storms', ' sandstorms', "volca	['petroleum', ' textiles', ' food processing', ' beverages', ' tobac
tu	['coal', ' iron ore', ' copper', ' chr	['severe earthquakes', ' especially i	['textiles', ' food processing', ' automobiles', ' electronics', ' mir
ae	['petroleum', ' natural gas']	['frequent sand and dust storms']	['petroleum and petrochemicals', ' fishing', ' aluminum', ' ceme
us	['coal', ' copper', ' lead', ' molyb	['tsunamis', ' volcanoes', ' earthqua	['highly diversified', ' world leading', ' high-technology innovat
ym	['petroleum', ' fish', ' rock salt',	['sandstorms and dust storms in sur	['crude oil production and petroleum refining', ' small-scale pr

Figure 2: NLP Lists

To generate these scores, we started by extracting the individual documents (in the case of Lexis Nexis) and extracting the free text from the newspaper articles scraped from various international news organizations websites. Once all these documents (we used ~25k documents) were extracted into a single list in Python we then extracted the year in which the article was written from each article. This was to establish what we call a recency bias. When pulling documents from Lexis Nexis , some of the documents were dated as old as 1909 which presents an obvious issue in terms of analyzing how useful a document like that might be in modern day. Even though older documents may not be as relevant in modern day, we also wanted to consider them since certain countries have been historically dominant or volatile which often gives them

greater or lesser power in the present. To fix this we normalized the extracted years and scaled them logarithmically such that documents in years up until about 2000 (most of the documents

were more recent than this so the distribution was highly skewed) were treated with some, but significantly less importance than the newer documents. Once this recency bias was established we then began actually analyzing the documents.

To begin the analysis we used a dictionary containing our twenty-two country names (as well as some synonyms) to search each document individually and extract the countries which accounted for 15% or more of all mentions of countries in the document. To determine the number 15% we looked at the resulting documents at increments of 5% from 5-30% and the resulting countries in order to evaluate if the countries that were being captured were indeed the countries which the article was discussing.

Once we had established which countries the article was discussing we began to generate our aggregate scores. The first score we generated was the instability score. To do this we used a dictionary of many terms which generally indicated instability (terrorism, rioting, killing, etc) and searched through each document to find occurrences of such words. Using the number of occurrences of these words, the primary countries for the document, and the recency score for the document, a score is then calculated and funneled into a total score for the related countries.

The process for evaluating natural resources strength was similar, but slightly different in that we used dictionaries of natural resources and natural hazards scraped which were individualized for each country from the CIA Factbook data to create this score. In a similar process, we looked at the primary countries and the word frequencies of the natural resources and hazards, using each country's individual dictionaries. However, we then combined the two scores (positive for the resources and negative for the hazards) along with the recency bias numbers for each article to create a single score for natural resource strength.

For the final aggregate value, industrial strength, we used dictionaries generated from our scraper, similar to the natural resource score, in order to analyze the industrial strength of each country.

Overall this method showed much promise but could probably be improved with better dictionaries and by including some sort of sentiment analysis to be able to better determine if the article is talking about such industries, resources, or stability in a positive or negative light.

(3.3) Weighted Order Decision Making to Evaluate Scores

For our overall scoring system for each category we used a simple scoring system, scoring each attribute from -5 (a very important negative attribute) to 5 (a very important positive attribute):

Attribute	Weight
Exports	4
Imports	2
GDP	5
GDP Per Capita	5
Oil Production	4
Refined Petroleum Production	4
Total Area	1
Water Area	3
Population	5
Dependency Ratio	3
Unemployment	2
Literacy	2
Poverty	-3
Population Growth	2
Net Migration	3
Urban Population	1
Obesity Rate	1
Life Expectancy	3
Fertility	2
Internet Access Rate	2
Infant Mortality	-4
Military Expenditures	5
Health Expenditures	3
Education Expenditures	3
Industrial Growth	4
Electricity Production	3
Natural Resource Index	4
Industrial Index	2
Instability Index	-2

Here a negative weight means that the attribute contributes negatively to power. For example, if poverty is high then it should contribute more negative effect to power. Therefore, its weight was taken as negative. On the other hand, GDP has a positive contribution towards power. So we

have kept it value to maximum value of 5. After creating the truth tables and fixing these weights, we calculated the scores for each country by multiplying the attributes of these value by their weights and adding them up. In our final table in User Interface each row represents the country with columns as these attributes one last column for the scores. The table was sorted in decreasing order of scores. In summary, higher the score, more is the power.

(3.4) Front End User Control

In addition to providing a robust backend analysis system. We also gave the user the option to control much of how they perceived power from the front end of our application. The user is given the option to select between the twenty-two countries which we selected to determine which ones he/she would like to do analysis on and he/she is also then able to determine the attributes, out of our five primary attribute categories, which they value the most. They can either decided to exclude categories of attributes or give each category a weight to determine its importance in calculating a score. We felt that this was a necessary step given that many users may have different perceptions of power or be interested in different aspects of what is influencing power, and this gives them the power to further investigate these ideas.

4. Tech-stack

Architecture

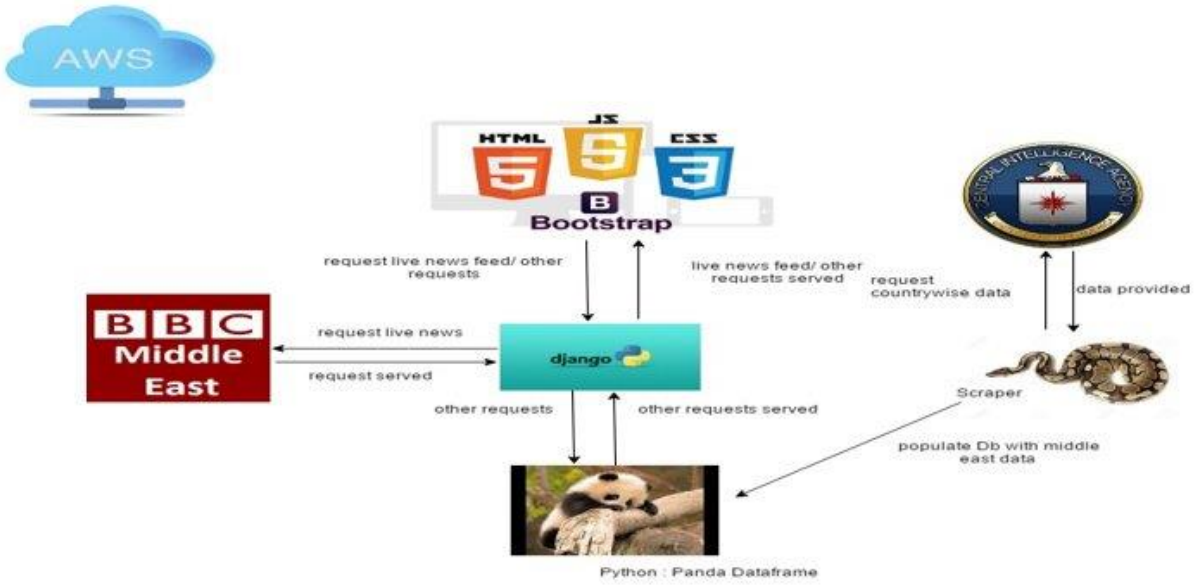


Figure 3

So, the flow of our application is described in the figure (3) above.

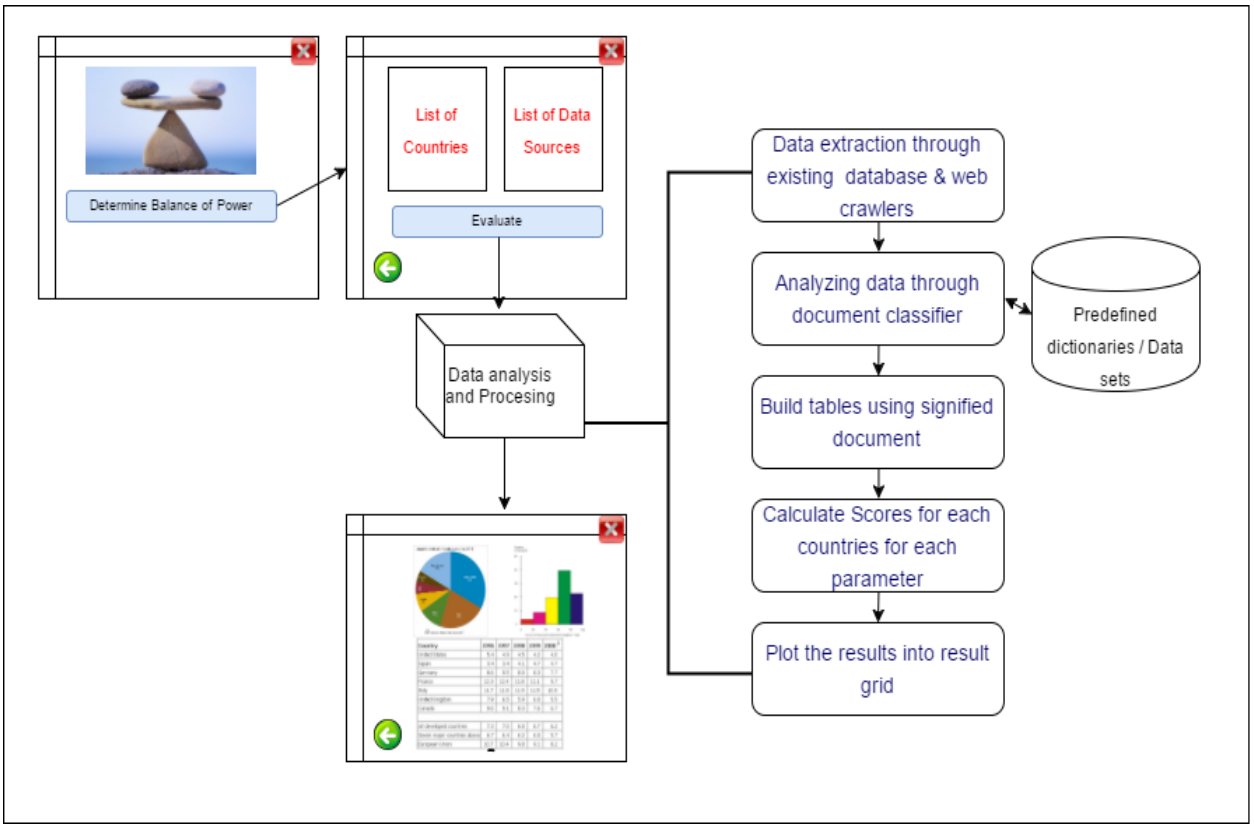
The complete technology stack that we used for the application is :

1. HTML, JavaScript, CSS and BootStrap for the front-end, basically validations and UI
2. Django Framework (Python) for the middleware and business logic
3. Python Panda Dataframe and CSV for the database

5. Web Application Development

Web application for Balance of Power can be thought of evaluator based analyzer where we would give the inputs as we want and can get results with scores for each country accordingly.

Design Diagram:

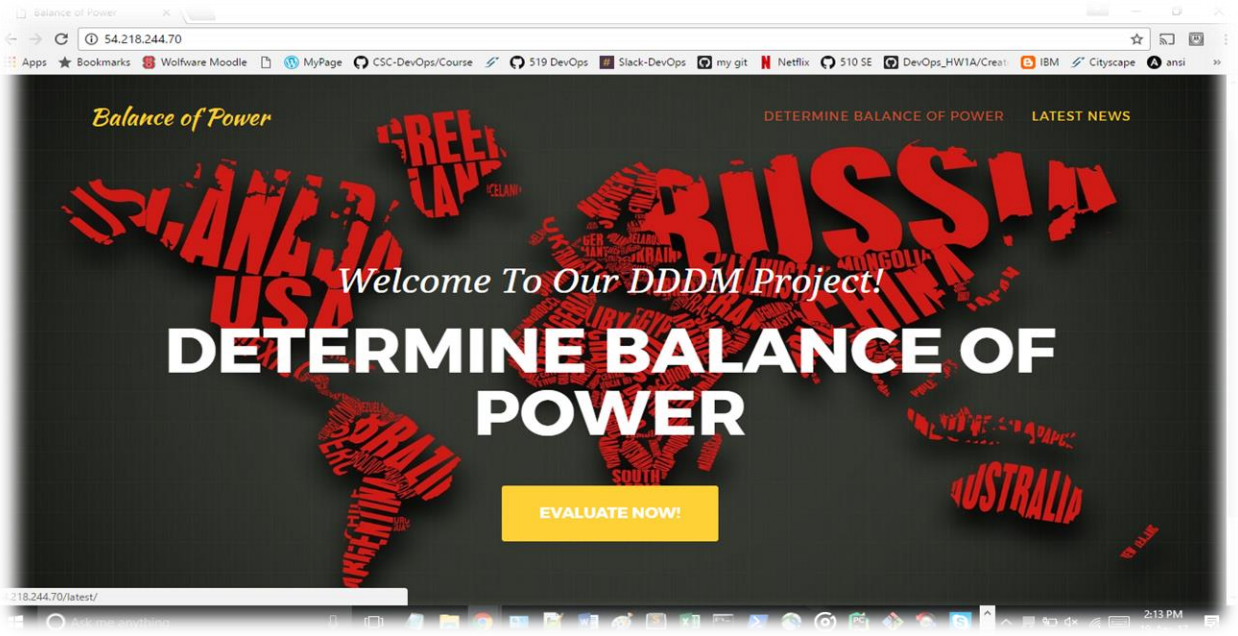


- Web application is developed in Django framework and we deployed using Amazon EC2 instance by AWS

- For UI, 4 pages we chose are
    - Welcome Screen
    - Latest News Screen
    - Form filling for power evaluation
    - Results and display of scores

1. Welcome Screen:

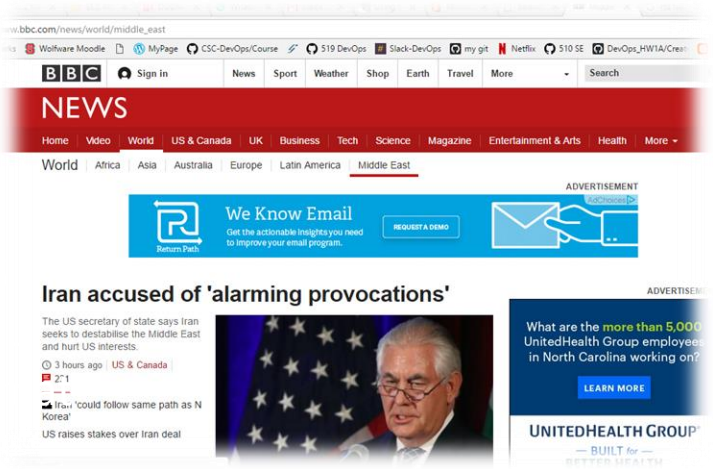
This would have important navigations including the inputs page and to see latest news.

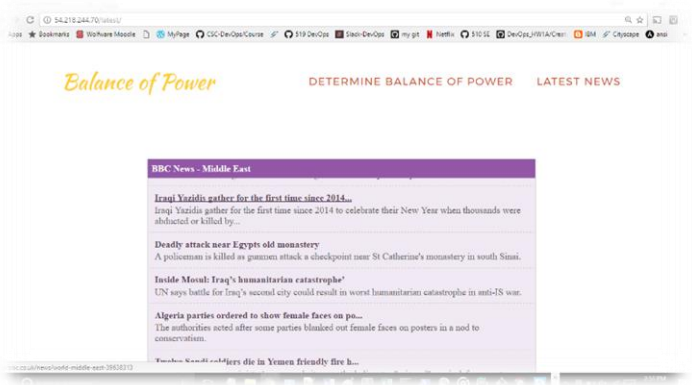


2. Latest News Screen

Here one can see the latest happenings in Middle east countries.
We have used news scraper for this and latest news from middle east countries will be aggregated here.

Tools used: BBC news, Javascript for scraping and aggregating data through XML.

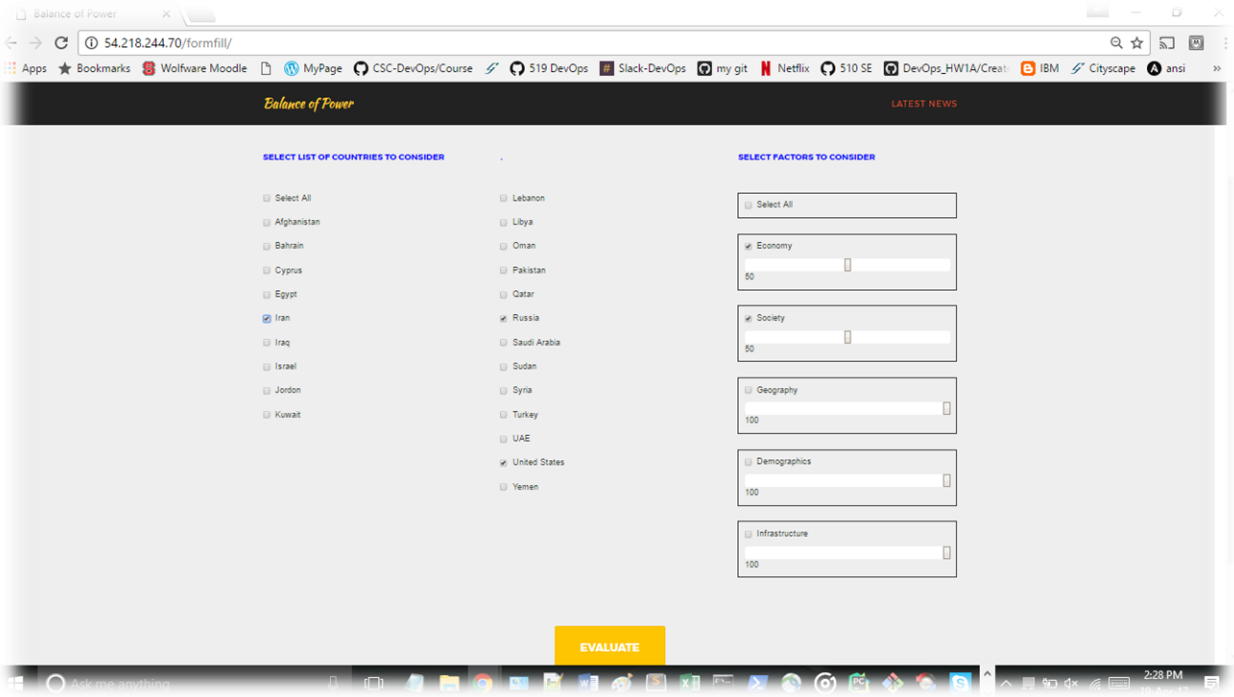




3. Form filling for power evaluation

Here we specified various categories one might choose for consideration

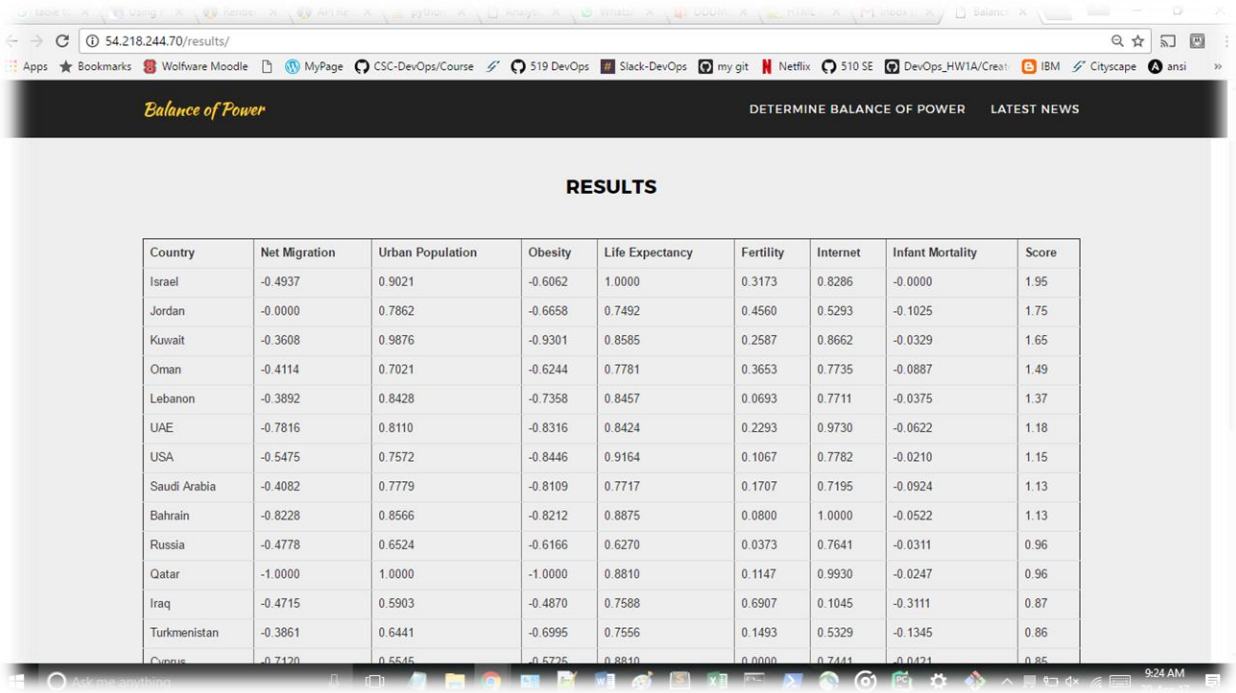
- Countries: One can select one/few/all of the countries to consider for balance of power evaluation
- Factors: One can select few/all of the factors including Demographic, Military, Economic



4. Results and display of scores

This screen will display the results of the evaluation.

This will be in tabular format with the scores which would represent the power of that particular country.



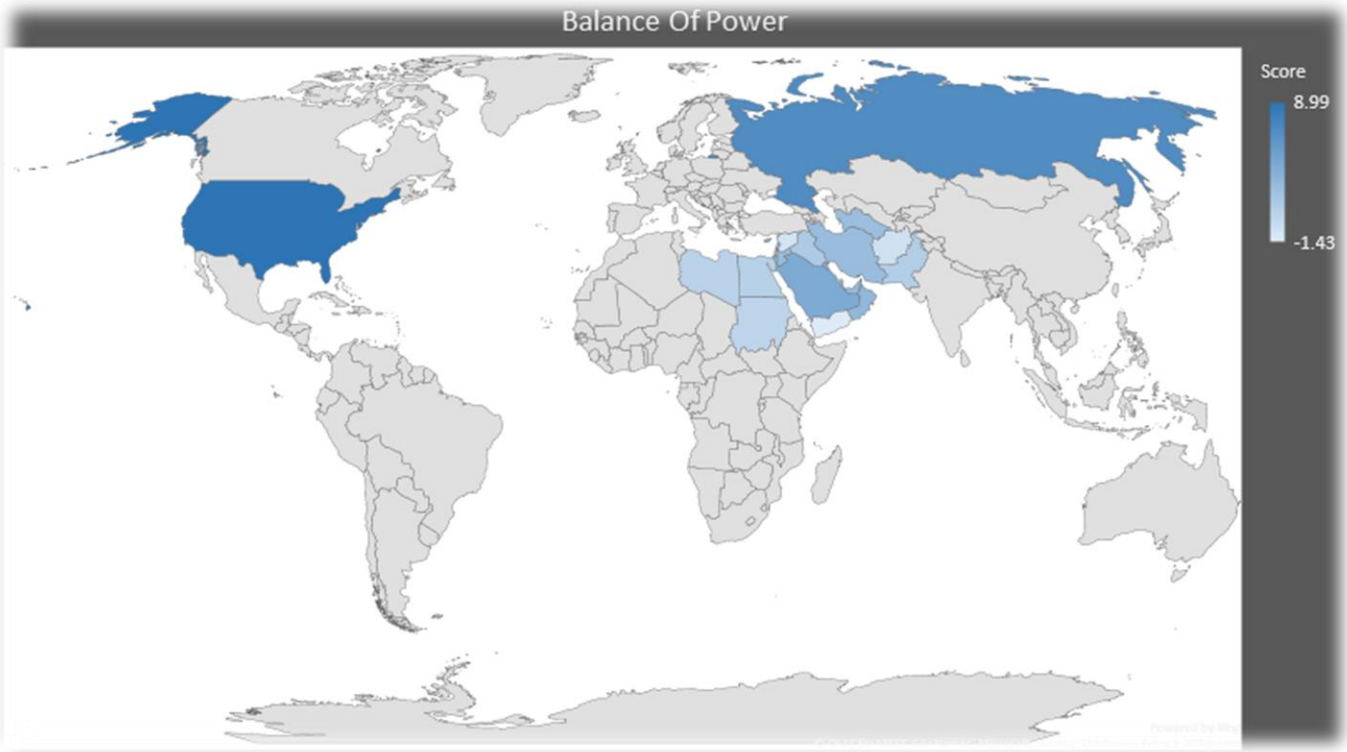
54.218.244.70/results/

Balance of Power DETERMINE BALANCE OF POWER LATEST NEWS

RESULTS

Country	Net Migration	Urban Population	Obesity	Life Expectancy	Fertility	Internet	Infant Mortality	Score
Israel	-0.4937	0.9021	-0.6062	1.0000	0.3173	0.8286	-0.0000	1.95
Jordan	-0.0000	0.7862	-0.6658	0.7492	0.4560	0.5293	-0.1025	1.75
Kuwait	-0.3608	0.9876	-0.9301	0.8585	0.2587	0.8662	-0.0329	1.65
Oman	-0.4114	0.7021	-0.6244	0.7781	0.3653	0.7735	-0.0887	1.49
Lebanon	-0.3892	0.8428	-0.7358	0.8457	0.0693	0.7711	-0.0375	1.37
UAE	-0.7816	0.8110	-0.8316	0.8424	0.2293	0.9730	-0.0622	1.18
USA	-0.5475	0.7572	-0.8446	0.9164	0.1067	0.7782	-0.0210	1.15
Saudi Arabia	-0.4082	0.7779	-0.8109	0.7717	0.1707	0.7195	-0.0924	1.13
Bahrain	-0.8228	0.8566	-0.8212	0.8875	0.0800	1.0000	-0.0522	1.13
Russia	-0.4778	0.6524	-0.6166	0.6270	0.0373	0.7641	-0.0311	0.96
Qatar	-1.0000	1.0000	-1.0000	0.8810	0.1147	0.9930	-0.0247	0.96
Iraq	-0.4715	0.5903	-0.4870	0.7588	0.6907	0.1045	-0.3111	0.87
Turkmenistan	-0.3861	0.6441	-0.6995	0.7556	0.1493	0.5329	-0.1345	0.86

5. Graphical results:



6. Discussion

Challenges Faced

There is hardly any project without challenges, but we made sure we tackled ours with consistent work. Below are some of the challenges we faced which were either in implementation(i.e. non-resolvable) or unable to fit in current implementation(i.e. we would love to resolve for the next iteration of the project)

a) Data Integration from multiple resources

As discussed above, the nature of data we are using is dynamic(i.e the GDP population keep on changing, as well as wars keep happening). There is no single source of this huge amount of data. CIA website has the most data, but other news sources don't. Also each source provides data with its own variable names which may either match to other source or not. Then the data reporting metrics are different, some have metrics in days/months, others have in years. News sources have altogether different format of test reports.

b) Inconsistencies in data

Even within a single source, there are a lot of inconsistencies in the data, for example, cia data has a lot of misses for literacy rate i.e. US literacy rate isn't mentioned. This data may not be freely available elsewhere. Different countries have different periods of national surveys, hence the data doesn't match or is missing.

c) Indirect correlations

Consider irrigation rate. This data doesn't have much effect on international relations, but is directly affected by the climate and directly affects the future crop production. Hence it is a latent factor. Similarly there are a lot of latent factors which have indirect correlation on the present variables. A simple model with the predictor variables has too much autocorrelation to be useful for a prediction model. Hence latent factor identification becomes important.

d) News to quantitative data

This is a portion we still have in store for future implementation. We needed to implement a news processing(sentiment analysis, integer mining, and nlp conversion) to convert news sources into quantitative data of decent useful quality. Sadly that couldn't be implemented within the given time frame. We have that planned for future.

7. Conclusion and future work

(7.1) Future Work

Using news data for current affair extraction is a very good idea. Similarly, using social media and better modelling. Below are our future prospects which we thing bring more to the table regarding this project.

a) Using news data

We haven't used news data to its fullest extent. We have implemented a rudimentary algorithm for the extraction of data from news articles but we can This would require some NLP, sentiment analysis, relevant news filtering, and a data processing pipeline to process. The backend can be implemented in high end machines to make it more robust and real-time.

b) Using Social Media

Like news feed, social media can be used to identify mass sentiment regarding current affairs. In present times, social media has become a huge influencer of mass opinion and naturally wherever mass opinion flows towards, the power flows. But it would require a lot of latent factor processing which is currently out of scope of this project. With time, a model can be made and fine-tuned for such purposes.

(7.2) Conclusion

Through this project, the intent was always to understand the business challenges and implement an end-to-end solution. In this iteration of our solution we were able to cover a significant portion of our challenges and implement a basic end-to-end design which covers the necessary portions. The underlying program can b=now be focused upon and made more intelligent, and frankly there is no limit to how smart a program is, the intention is always to make a good end-to-end deliverable. With that in mind we have been able to cover the following things –

a) Defining the power of a country in terms of its economic, demographic and military strengths gives promising results.
b) The critical thinking process is a well defined approach to solve real world problems.
c) Critical thinking is not only required to solve the problem, but intermediate steps of critical thinking can also be used to understand the problem better.

d) This understanding can further be used to establish approach to solve the problem and gain more promising results.

Finally, a huge token of gratitude towards **Dr. Kowolenko** for guiding us throughout the project – it was really informative and enjoyable experience.

8. Appendix

Important links:

a. Code repository:

Github repository: - https://github.com/bmukheja/DDDM

b. Data scraping from:

CIA site: - https://www.cia.gov/library/publications/the-world-factbook/

c. BBC news:

News scraping from: http://www.bbc.com/news/world/middle_east

References

1. Critical thinking kickstart guide: CSC 591 DDDM course slides by Dr. **Kowolenko**
2. **Country** facts: cia : https://www.cia.gov/library/publications/the-world-factbook/
3. Deploying application to AWS: http://docs.aws.amazon.com/elasticbeanstalk/latest/dg/create-deploy-python-django.html
4. Great powers @ **Wikipedia** https://en.wikipedia.org/wiki/Great_power
5. **AWS** https://aws.amazon.com/
6. **NLTK** framework http://www.nltk.org/