Fundamentals of Data Science: Assignment 1

Global Response to the Needs of the Planet: Through UNGD Corpus Analysis

Team C5
Adam Eljasiak, Andesh Haribhajan, Antonis Klironomos, Roopa Sudhakar, and Sehra Elahi

University of Amsterdam
Faculty of Science
MSc. Information Systems: Data Science Track
04 October, 2021

Abstract. In this report, we analyze the data sets from the United Nations General Debate corpus (UNGDC from 1970 to 2020) to draw insights into which countries are interested in climate-related issues. Moreover, we scrutinize this further by collectively inspecting the relationship between the country and its total forest areas over the years, to comprehend the relation between the happiness index of the country and the number of times these countries have climate change in the UN speeches. Our research question and hypothesis is therefore based on the following proposal: “Is there a correlation between the number of times a country mentions climate change, environmental policies, and the collective happiness of the country?” We take the forest area percentage into account as an indicator for the action taken by countries. The answer to this question is not conclusive, but some interesting relationships give the further investigation a reason.

Keywords: UN General Debate · Natural Language Processing · Topic Modelling · Climate Change · Deforestation · Machine Learning · Classification · Regression · Data Science· Happiness

1 Introduction

Each year the members of the UN hold a General Debate (GD) to discuss a variety of international issues. One of the main global concerns disputed at the GD is that of climate change [3]. Over the past century, there has been a rapid negative change of the climate and the environment, mainly due to human activities [4], one such action being deforestation. Forests and trees are essential to the global climate, and clearing them threatens the biodiversity of the planet, which in turn destroys wildlife [5]. However, the rapid change in climatic variation is not just limited to the natural environment. As it has been found in the research, climate change may have an impact on the happiness of the population [6]. Given that, we would like to know what is the impact on the general happiness of the countries that often raise the topic of climate change, and if it can be explained by deforestation in the region. In this report, we will explore datasets containing information about deforestation and happiness in different countries alongside the log of UN speeches to answer the hypothesis statement.

Hypothesis:

> *“Is there a correlation between the number of times a country mentions climate change, environmental policies and the collective happiness of the country?”*

2 Methodology

To test the hypothesis, additional relevant data had to be collected. All these data sets were inspected carefully and prepared for further pre-processing. After the pre-processing the data was modeled to apply the most applicable machine learning methods to test the hypothesis.

2.1 Data Sets: Description & Preparation

We will be considering a total of four datasets to answer the research question; The UN General Debate Corpus (UNGDC) [2], UN Standard Country Methodology (UNSCM) [7], Forest area (% of land area) [1], and the Happiness Report Dataset [8]. The UNGDC and the UNSCM are considered as the base datasets. These will provide an apprehension into what countries and regions take part in the UN debate, simultaneously allowing us to analyze the speech text that is available for the sessions from 1970 to 2020. The Forest area dataset consists of the percentage of forest land of 266 countries from the years 1990 till 2020. The forest % value in the dataset is derived by dividing the total area covered by forests by the country's total land area, multiplied by 100 to receive the total percentage of forested land in the country. The Happiness report dataset contains variables associated with national happiness and the well-being of citizens within the countries, from 2005 to 2020. The factors of consideration with respect to climate change and happiness, across nations that persistently speak out about the environment, will be debated by merging these datasets and identifying variables of interest. The parameters can be found in Table 1.

Table 1. Variables of interest and join keys of the datasets

Dataset	UNGDC	UNSCM	Forest Area	World Happiness
Joined by	Country Code, Country, Year			
Variable(s) of Interest	Session Speech	Region Sub Region Developed/Undeveloped	Forest area %	Life Ladder

The datasets considered in Table 1 were joined on the common parameters. To further understand the properties of the data, each of the raw datasets were pre-processed, cleaned and prepared, prior to merging.

UNGDC Dataset This data set has 8481 rows and the following four columns counting object type data.

1. *Session*: the UN session. There is one session per year, and the data in this dataset ranges from session 25 to session 75, for a total of 51 unique sessions.
2. *Year*: The year of the session, from 1970 to 2021, for a total of 51 unique years.
3. *CountryCode*: The representative's country, as an ISO 3166 Alpha-3 country code. There are 201 unique country codes.
4. *Text*: The complete text of that country's statement in the general debate from that year. With a total of 8481 unique speeches.

UNSCM Dataset We conducted data manipulation in Excel for the UNSCM dataset where there were extra commas in country names and was replaced with a hyphen. While merging the UNGDC data with UNSCM data we found that the total number of speech text left after merging it was 8384 in comparison to 8481 speech text that was in UNGDC. Upon further analysis on the missing rows, we found that the countries with ISO - CSK, DDR, EU, POR, YDYE, and YUG were omitted. The countries Czechoslovakia (CSK), South Yemen (YDYE), and Yugoslavia (YUG) do not exist later in the 90s decade as they were either divided or dissolved. Due to the Forest Data available only from 1990 and Happiness Index being measured from 2005, it becomes unnecessary to process this data to add in the merged dataset, as it would anyway be left out as these speeches are effective prior to 1990. We replaced DDR (German Democratic Republic) to DEU (Germany). Also, POR was replaced with PRT for Portugal as POR may be a mistake in the dataset. We skip replacing EU (European Union) as there are only 10 speeches of this which is not a significant count.

Forest Area Dataset Initially, the dataset was reshaped to create a single column for 'Year', instead of one column per each year. Next, all records with the years earlier than 2005 were discarded, since they would not join with the happiness dataset. Unnecessary columns were removed, leaving Country, Country Code, Year and Forestation parameters. Finally, the preprocessed forestation data was merged with the rest of our data.

World Happiness Dataset The Happiness dataset contains variables associated with national happiness and the well being of citizens within the countries, from 2005 to 2020. We are interested solely in the happiness index of the countries, located under “Life Ladder” column.

2.2 Data Pre-Processing

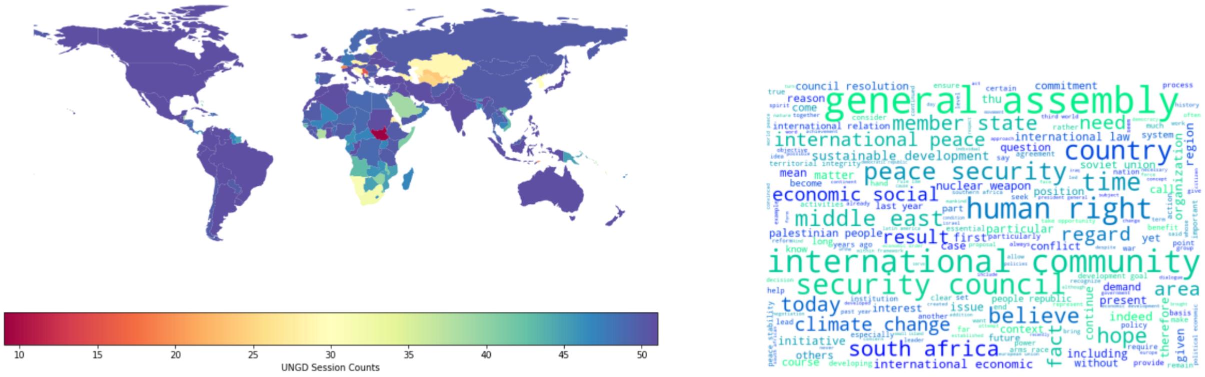
The text files with speeches also contain many imperfections that might severely impact further analysis. They can be editorial remarks, references to legal documents or unwanted symbols. Therefore, each speech undergoes a trimming according to these rules:

- Text between curly and square brackets is removed entirely, since it contains editorial remarks
- Text between normal brackets that is shorter than 150 characters is discarded for the same reason as above. Longer ones usually contain parts of speeches
- Numerical identifiers are removed, such as `135-344/4`
- Finally, all punctuation marks except for comma, dot and hyphen are removed

Natural Language Toolkit library was used to pre-process speech before applying our model in the following ways: (i) tokenization, (ii) punctuation and special characters removal, (iii) lower casing all the text, (iv) stopwords removal and (v) lemmatization.

The total number of sessions that each country spoke in after the data analysis and merge of UNGDC with UNSCM is visualized on the world map in Figure 1(a).

Fig. 1. Exploratory analysis of UNGD.



(a) The total number of sessions that each country spoke in UNGD

(b) Cloud of the most frequent words in the corpus

Since the text contains multiple insignificant words, which are not useful for classifying the speech text into different topics, those words are removed by adding as extensions to stopwords. Furthermore, to verify if the pre-processing required further adaptation, a word cloud was created to view the most frequent words for easy identification of more words that might be insignificant for training our classification model as seen in Figure 1(b).

2.3 Data Modelling

Classification The data in the merged dataset has mostly categorical values and requires unstructured text analysis. We employed the generative statistical model, **Latent Dirichlet Allocation (LDA)**, as LDA can deal with multiple-class classification problems and give an appropriate prediction.
The procedure that we followed could be summarised by the following points:

1. Speeches were transformed to **feature vectors** by being converted to a matrix of word frequencies. During this process tokenization, stemming and removal of stop words had been performed for the model to use a clean form of the speeches. Additionally, words with a frequency higher than 70% were ignored because it was observed that there were many words (e.g. general, council) that were common among the majority of speeches but had no significance for our application.
2. LDA model with **10 topics** was fitted to the previously mentioned matrix. The number of topics was provided by **Grid Search**.
3. Visualisation was performed using an Intertopic Distance Map and a bar plot for the most relevant terms of each topic. This helped in interpreting the topics and visually validating the model.
4. **The most dominant topic**, which was assigned by LDA to approximately **65% of the speeches**, appeared to contain the word "clim" (stemmed version of "climate") as the second most relevant term. So, it was obvious that this was the topic that corresponded to climate-related speeches.
5. The dataset was filtered and sorted to contain data about the 5 countries that had the most climate-related speeches.
6. The number of climate-related speeches, the happiness index, and forestation data of these 5 countries was visualised to examine correlation.

Regression To further understand the correlation better between happiness and deforestation, a linear regression model for the variables 'Life Ladder' and 'Forest area %' will be implemented to grasp the data behavior better. To determine the best hyperparameters for our model, we run `GridsearchCV` from the scikit-learn library. This helps to loop through predefined hyperparameters and fits the model to the training set. The results from running grid search, with cross-validation set to 5, are defined in Table 2. From the results obtained, we can determine that a polynomial regression model is needed with degree 0, to expand the feature space and observe non-linear relationships in the data (as determined in Figure 3(a). Furthermore, the y-intercept will have to be determined by the line of best fit, as its set to True. We can now arrange the data in a feature matrix and use the parameters found to fit the model to the dataset.

Fig. 2. Linear Regression line

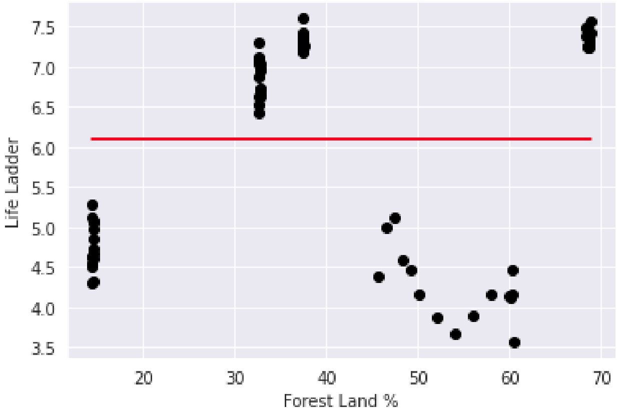


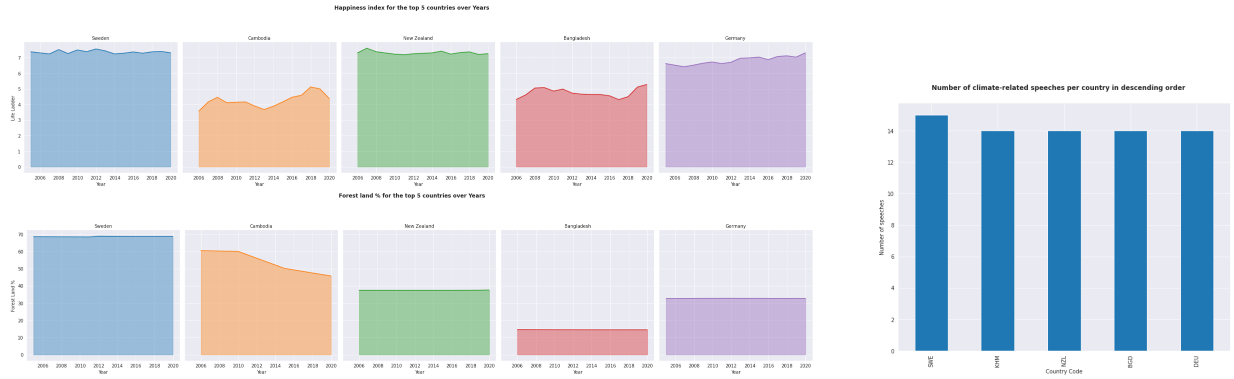
Table 2. Best hyperparameters found during the grid search for the training set

Hyper-parameters	**Result**
linearregression_fit_intercept	True
linearregression_normalize	True
polynomialfeatures_degree	0

3 Results & Discussion

We visualized the relation between the happiness index of the top five countries who spoke the most about climate-related topics and their respective country's forested area percentage. The regression line formed from data modeling can be seen in Figure 2. Due to this, we conclude that there is no correlation between happiness index "Life Ladder" and "Forest area %", as the line appears to be horizontal, across the x-axis with a slope of 0. Hence the changes in deforestation across the top countries that speak out about climate change, have no impact on their happiness index, as there is no regression.

Fig. 3. Visualisations for the top five countries



(a) Percent of land covered by forest and happiness index for the 5 countries that gave the most climate-related speeches

(b) Five countries that gave the highest number of climate-related speeches

4 Conclusion

We conclude that UN state members who speak most frequently about topics related to climate change, do more to improve and/or maintain the environmental conditions of their respective countries. This can been proven by the observation that most of the examined top five countries (see Figure 3(a) and 3(b)), maintained their percentage of forest area throughout the years. Furthermore, there are some exceptions of countries that mentioned the subject of climate changes but for them the percentage decreased significantly. On the other hand, judging by Figure 3(a) we can deduce that the happiness index of a country is independent of the size of its forest land. This means that citizens are not significantly affected by the environmental situation of their country.

References

1. Forest Area Dataset, Food and Agriculture Organization
2. Jankin Mikhaylov, Slava; Baturo, Alexander; Dasandi, Niheer, 2017, “United Nations General Debate Corpus”, Harvard Dataverse, V6
3. Alexander Baturo, Niheer Dasandi, and Slava Mikhaylov, “Understanding State Preferences With Text As Data: Introducing the UN General Debate Corpus” Research & Politics, 2017
4. “Climate Change: How Do We Know?”, Earth Science Communications Team, NASA
5. “What is the Relationship Between Deforestation And Climate Change?”, Rainforest Alliance, 2018
6. Krekel C., Mackerron G.: How Environmental Quality Affects Our Happiness on Proceedings, Chapter 5 Happiness Report (2010)
7. Standard country or area codes for statistical use, United Nations
8. Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. 2021. World Happiness Report 2021. New York: Sustainable Development Solutions Network