Andrew Plum, Leo Bomboy, Sebastian Fedane, Carrie Nickel, Palak Pandey

Professor Ma

CS 489

12/4/2024

Final Project Report: DBpedia Data Analysis

1. “Choose an investigation goal and identify and collect existing sources of linked data that can address the research goal:”
   1. “Choose and state the research goal. Describe the overall design and implementation of the workflows in the whole assignment.”
      1. The research goal was to determine the quality of country related DBpedia data and to use that data to make predictions of the rate of happiness in a country.
      2. To do this, I first had to design a SPARQL query to retrieve the data from the DBpedia stores through a SPARQL DBpedia endpoint and save the data as a CSV file.
      3. I then downloaded the dataset from the Happiness World Report as a CSV file.
      4. I then used python to load the DBpedia data into the web application jupyter notebook and combined it with World Bank data retrieved through its api and then performed analysis on the combined datasets to achieve the research goals I outlined.
      5. The analysis I performed was exploratory data analysis which mostly included basic visualizations like histograms, a correlation matrix, and a heatmap; I mapped the difference in quality between the DBpedia and the World Bank data; and I trained several machine learning models using the DBpedia data to predict the rate of happiness of a country.
   2. “Describe the progress of exploring and identifying classes, properties and individuals from the existing SPARQL end points. After you get familiar with ontologies and instance records on the end points, describe how you design the workflow to query them. Describe the SPARQL queries implemented in order to collect the necessary datasets.”
      1. To identify relevant classes, properties, and individuals I wanted to involve in my SPARQL query, I first familiarized myself with the part of the DBpedia ontology related to my topic of countries.
      2. I did this by selecting potentially relevant classes, properties, or individuals to return in a temporary SPARQL query.
      3. If the results (the selected classes, properties, and individuals) were useful in that they were some final result that I wanted returned from the main query or it helped me to obtain some final result I wanted from the main query, then I proceeded to incorporate it into the main query only if I had conceived of a way to do so which was not to computationally expensive (there were a few times when I wrote a query to select relevant data but it timed out the execution of the DBpedia SPARQL endpoint and thus was not a viable method retrieve the desired data).
      4. Since I wanted my dataset to focus on countries and its relevant topics (like the economy and demography of each country), I started my search by first designing a query which retrieved the relevant world countries I was looking for.
      5. To get the most inclusive list of countries, I wrote the triple “?country rdf:type yago:WikicatCountries, dbo:Country”, and then I made use of the SPARQL keyword UNION to only grab the countries which were related to some relevant country-like resource (eg. dbr:country, dbr:state, dbr:republic, etc.)
      6. From the returned results, either from the temporary query or the main query, if a DBpedia resource page was returned, I typically opened up the resource page and surfed for relevant classes, properties, and individuals that could be incorporated into my main query.
      7. This was how I found the relevant DBpedia resource “Country” pages and their relevant demographic properties and the “Economy\_of\_X” DBpedia pages where X is some country which then also had relevant economic properties of countries which I was looking to return.
      8. Designing the rest of the query to retrieve the relevant data I was looking for entailed repeating the aforementioned steps described.
      9. One thing of note is that the main query is split into two separate queries (they are roughly organized to where one retrieves country demographic data and one retrieves country economic data) even though it conceptually made sense to execute as one combined query because as a single query, it was too much of a load for the DBpedia SPARQL endpoint to execute (a specific error was returned explaining such a limitation).
      10. One other thing of note, the SPARQL keyword OPTIONAL was utilized heavily (almost always when possible without causing problems to the query results) so that the data returned was as inclusive as possible while still being structured as desired since I wanted to observe what data was available and what data was missing and it would be easy to remove rows later when working with the data in python.
2. “Data Analysis”
   1. “Develop and state two particular questions/ hypotheses related to the goal of the investigation and that can be answered using the datasets under consideration. Design an analysis study to answer these questions and document the analysis design.”
      1. The first investigative question I had for the project was: What is the difference in the quality of country data from DBpedia and country data from the World Bank where the World Bank country data (the World Bank has high quality data) is used to benchmark the DBpedia data’s quality?
      2. To attempt to answer this first question, I compared the DBpedia data and World Bank data by first loading and joining the data into a single pandas dataframe using python, then I cleaned the data and prepared it for analysis, afterwards I standardized the values in both datasets to the present values in their columns, then I calculated an element wise difference between the two dataframes (only calculating the resulting element when the two elements involved in the calculation were not missing values), after that I took the mean of each of the rows (again ignoring null values), and then I mapped these values (the average difference of the standardized values) by country using geopandas; I also further analyzed these values by averaging these values (to where we calculated the average of the average difference of the standardized values) and took this value and plotted a vertical line at the x = calculated\_value on a normal distribution to where it could visually be seen on average how many standard deviations off a country’s associated DBpedia values were from the referenced World Bank data; all calculations ignored missing values in the dataframes where appropriate.
      3. The second investigative question I had for the project was: What country related factors are most impactful in predicting country happiness rates?
      4. To attempt to answer this second question, using python I loaded the collected DBpedia and World Happiness Report dataset into their own pandas dataframe, I joined the DBpedia data with the column “life ladder” from the World Happiness Report dataset which represents the country’s happiness rate on a scale from 1 to 10, I then normalized the values of each column, and then using the column “life ladder” as the target variable, I trained several regression machine learning models on this combined dataset where R^2 was used to evaluate the model’s performance.
      5. The machine learning models trained included a linear regression, a decision tree regressor, a random forest regressor, and a neural network regressor where the linear regression and decision tree were chosen for their interpretability and the random forest and neural network were chosen to contrast their predictive power from the more interpretable models; surprisingly, the decision tree had the best performance with an R^2 of 0.8130 which was ideal because not only did it have high predictive power but it was also the most interpretable model of the four trained models; the decision tree regressor model explicitly shows how the model makes its predictions and tells the importance of each feature in the model based on its frequency of being used in the model in conjunction with the depth it is used at each of its times; the trained decision tree regressor says that “humanDevelopmentIndex” was by some margin the most important feature and the following three most important features for the model were “grossExternalDebt”, “unemploymentRate”, and “yearsExisted”.
      6. Further experimentation with what combination of features are fed to the model during training could help verify the decision tree regressor’s results in regards to feature importance, and hyperparameter fine tuning with each of the models could be done to possibly increase their individual performance.
   2. “Provide a description of the choice of programming languages/tools/ methods used for the analysis. Perform the analysis in a form that can be validated and describe the steps and results you took to ensure this validation.”
      1. For the data analysis, the programming language I used was python.
      2. For the coding, I utilized a combination of the text editor Notepad++ and the web based integrated development environment jupyter lab; note the output of the code can be observed in Google Collab and Visual Studio Code if you don’t have jupyter lab or jupyter notebook.
      3. I utilized many of the various packages that python has to offer like numpy, pandas, matplotlib, seaborn, sklearn, tensorflow, and geopandas.
      4. For the design of the analysis, it is explained in more detail in the previous section.
      5. The results of the analysis can be validated by downloading the data and code files and executing the code from scratch; personal testing consistently produced the same results.
      6. The conclusions of the analysis can also be tested through further experimentation with the analysis design with steps like modifying the previous models or training new models utilizing different algorithms altogether or by normalizing the values instead of standardizing the values and observing the difference in the results or any other reasonable implementation of statistical methods and then comparing the results to see if similar conclusions can be made.
3. “Visualization”
   1. “Prepare visualizations of both the data and the results of the analysis and describe the meanings revealed by them. Describe the format of the visualization products (e.g. images, interactive visualizations on the Web, etc.).”
      1. All of the visualizations are images produced by methods from some of python’s visualization-based packages.
      2. A graph with a number of missing values

         Description automatically generatedI created a heatmap which showcases the missing values in the DBpedia dataset (yellow means the value is missing and black means the value is present); it can be seen that the dataset contains many null values in each of the columns and rows; this led me to not drop entire rows or columns because then there would be to few entries for analysis; it can also be observed from the heatmap that “yearsExisted” is a column with many missing values.
      3. A screenshot of a graph

         Description automatically generatedFor the exploratory data analysis, I also created a correlation matrix of the DBpedia dataset which does a pairwise matching of each of the features in the dataset (including a pairwise matching with the same feature - this is a diagonal with a perfect correlation between itself); the correlations are in a range of potential values between 1 which is a perfect positive correlation and -1 which is a perfect negative correlation, the correlation matrix is useful in exploratory data analysis to get clues of which variable might be influential in predicting other features because the numerical relationship between the changes in the features are displayed.
      4. I also plotted histograms of each of the columns in the DBpedia dataset; this was useful to do in the exploratory data analysis stage because it visualizes the distribution of values in the columns potentially giving insights into which statistical analyses work well and what conclusions can visually be gleaned from the data; note there are more histograms displayed in the code’s output.

A graph with blue lines

Description automatically generated

* + 1. For analysis 1, I mapped the average difference of the standardized values by country to see if there was a geographic trend in the difference in quality between the DBpedia and World Bank data - there did not seem to be; I also plotted the average of those previous values on a normal distribution to concisely visualize on average how much those DBpedia derived values deviated from the World Bank data; I also did a barplot of the column-wise averages (instead of the row-wise averages) of the difference of the standardized values to illustrate which columns tended to have values with the largest difference on average from the World Bank data; each of these visualizations helped lead to various conclusions regarding the quality of the DBpedia data.

A graph with a red line and a blue line

Description automatically generatedA map of the world

Description automatically generated

A graph with blue rectangles and text

Description automatically generated

* + 1. For analysis 2, I plotted the training and validation loss curves of the neural network regressor to illustrate how the model was learning which was helpful in fine tuning the model and for ensuring there were no problems during the model’s training; printing the decision tree regressor was also useful because it gave visual insight into precisely how the model was making its predictions which is useful for model interpretability and concluding feature importance.

A graph of a graph

Description automatically generatedA screenshot of a computer program

Description automatically generated Note: This is just a snippet of the decision tree above – continues on for a while longer.

* 1. “Describe how your visualization meets the goal of the investigation and highlight any value that was gained.”
     1. The visualizations done during the exploratory data analysis were helpful for gaining an initial feel of the dataset I was working with.
     2. The visualizations made for the analysis of the first investigative question were useful because they helped to make the reasonable conclusions present in the data more apparent and comprehensible.
     3. The conclusions arrived at for the first investigative question was that there was no observable trend between the difference in the quality between the DBpedia and World Bank data; and overall, the difference on average between the DBpedia and World Bank datasets is very minor.
     4. The visualizations made for the analysis of the second investigative question were useful because they helped to determine the structure of the decision tree and the importance of each of its features and to also illustrate how the neural network was learning to ensure it was learning correctly.
     5. The second investigative question’s conclusions were that “humanDevelopmentIndex” was the most influential feature in predicting “Life Ladder” (country happiness rate) with the following three most important features for the model in its prediction of the target variable being “grossExternalDebt”, “unemploymentRate”, and “yearsExisted”; another conclusion was that the decision tree regressor was the best performing model with it explaining about 81.30% of the variability in the dependent variable’s data with the independent variables.
     6. Further testing and experimentation should be done to further verify these conclusions.