**Name: Muneeb Ur Rehman**

**Roll No: BCSF17M017**

**Section: Morning**

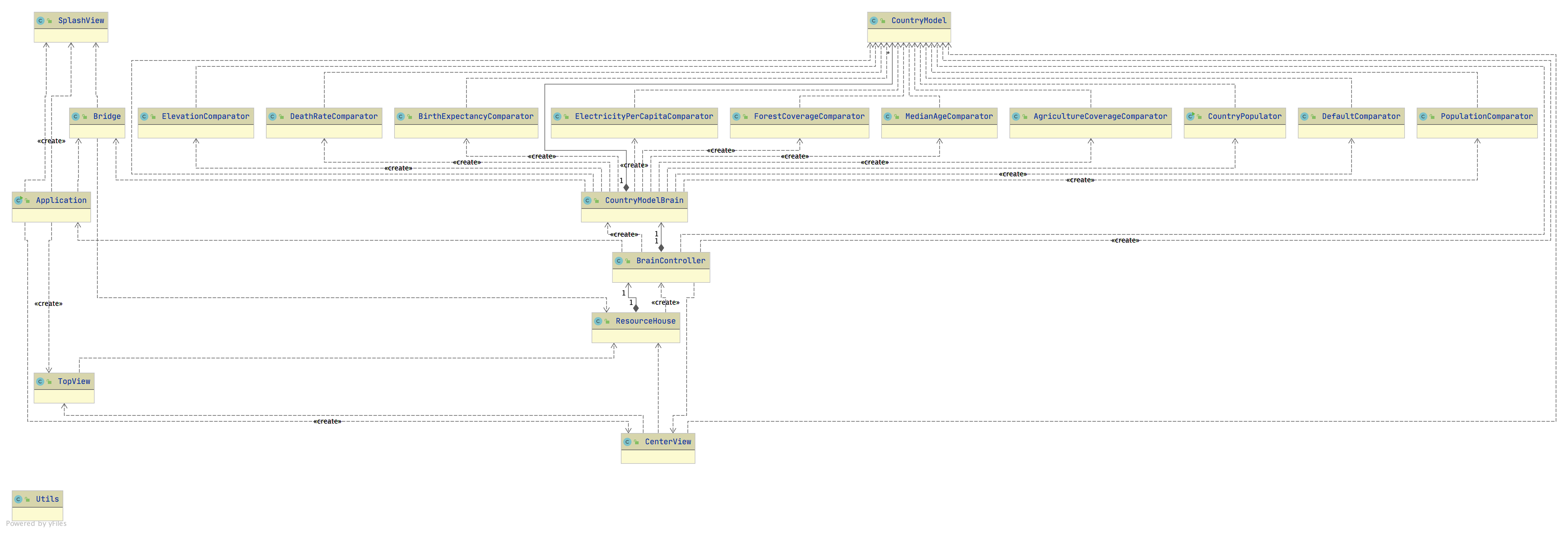
**Project Title:**

A crawler for scrapping, parsing and displaying data of all countries from CIA website.

**Project description:** <https://www.cia.gov/library/publications/the-world-factbook/print/textversion.html>

Given link contains information of all countries which is publicly available. Upon successful connection this crawler will start from this URL and get data of every country in the list in the order they are listed on website. It will then parse data of every country which are nodes technically and extract meaningful data. After that I will construct some regular expressions to throw away unnecessary information to convert data in a form that is savable and presentable. App will also give options to the user to arrange data according to some parameters which will update and rearrange list in real time.

**Project Design:**



**Design features:**

My Web crawler starts with a list of URLs to visit, which are called seeds. As the crawler visits these URLs (GET requests), it identifies all the hyperlinks in the pages and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. It copies and saves the information as it goes. The archives are usually stored in such a way they can be viewed, read and navigated as they were on the live web, but are preserved as snapshot. Also, I have to use many design patterns to update the data on GUI in real-time and to contain a single source of truth with the high complexity of automatic spidering.

**Drawbacks:**

* + It relies heavily on internet as It have to fetch parse extract modify data each time because it relies on live data, not static stored data.
  + I was unable to use database because database doesn’t fit with concept of web spider as it deals with live data all the time.