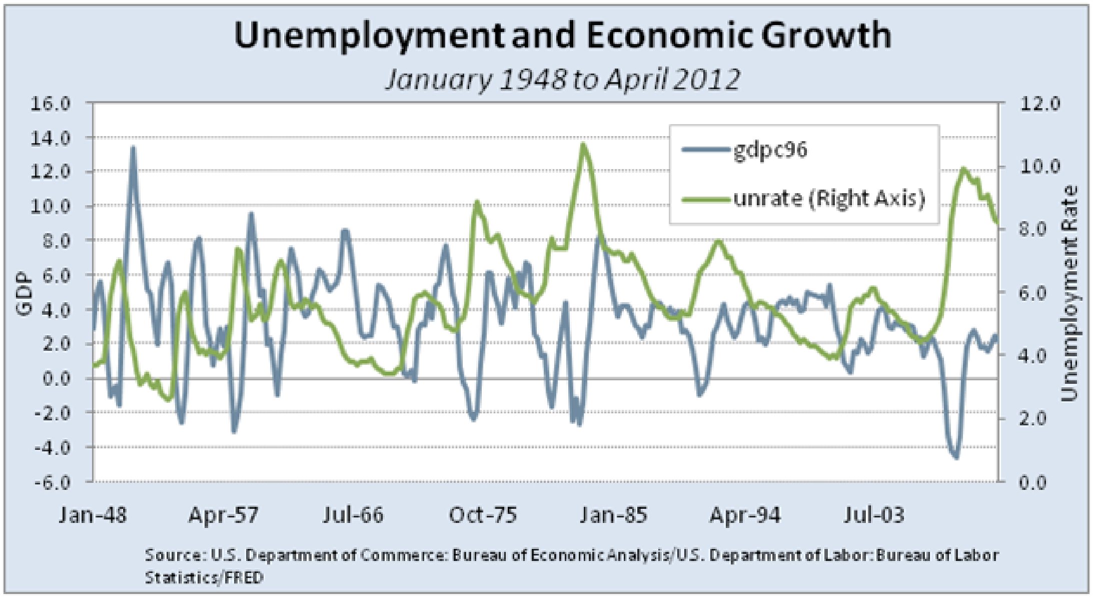
Final Project Proposal for CAP5771

Xi Yang 16216573

Proposed work

One of the most important big data source is economy. Everyday millions of data are produced. Data such as GDP, inflation, stock marketing etc. are every interesting for analysis to extract the useful information for future. All the macro-economy data sets are also closely related with the data sets collected locally. Combining macro-data with local data, a clear picture of distribution of economy based on geo location can be drawn. One of the interesting and easy to obtain data is Local Area Unemployment Statistics data set available from Bureau of Labor Statistics of United States. The unemployment index is always a good indicator of the economy.

Intuitively, the local unemployment rate should be reflected in the local economic data such as local area GDP and local area per capital income (Fig.1). The ideal relationship would be expected as negatively associated which means the lower the unemployment rate, the higher the GDP and per capital income will be. However, the results must be supported by data analysis. Such types of analysis include data integration, plotting, comparison and model training and predicting. Considering the amount of data that can be collected, scaling analysis method required to process these incoming data especially when we start to talk about data at national or global level.



Source: U.S. Department of Commerce: Bureau of Economic Analysis/U.S. Department of Labor: Bureau of Labor Statistics/FRED

Fig.1 The relationship between unemployment rate and GDP growth for US.

This project will focus on two major tasks. The first one is focused on data integration. All the data sets will be preprocessed to remove dirty data and scaled if necessary.

Related data sets will be joined and aggregated for obtaining unemployment rate associated with the detailed location at county or city level. The second project is focused on creating a model to link economic data with preprocessed unemployment rate data such as a linear regression model for example. Along the way, graphical expression will be created to visually reflect the data.

The very important point of this proposal is that I want to study the location related data (city, county, area, state etc.) which is more challenge than only deal with national level data because it would be interesting if some unintuitively results yielded.

Data

The data sources include Bureau of Labor Statistics and US Bureau of Economic analysis. For the BLS data, monthly labor force and unemployment series are available for approximately 7,500 geographic areas, including cities over 25,000 population, counties, metropolitan areas, States, and other areas provided at Bureau of Labor Statistics website https://download.bls.gov/pub/time.series/la. The city based GDP and per capita income data sets can be downloaded from https://www.bea.gov/. The largest data set I am going to use is a 280MB plain text data which includes more than 4 million records.

Execution plan

The data integration stage will be performed with Spark framework using Scala as programming language. The second stage creating linear regression model will be performed with python and the sciki-learn package. If the data process took too long, I will move to Amazon AWS clusters. But for now, I will use my home server with Spark set up already.