# Final Project Proposal

Yibo Zhao 001401747

## Introduction

Cars play important roles in daily life in the United States. However, buying a car is not an easy task, especially for new residents. Though we have many websites to search for an expected cars and even dealers have their own websites, useful information is scattered all over the Internet. Could we have a better option to find the cars wanted?

## Problem

1. Car information are distributed among different websites. For website like cargurus.com, they contain most part of car selling information. Some big car dealers have their own websites. Usually they post some car information only on their websites. This makes buyers hard to find cars. In most cases, buyers have to go through several websites to search for expected cars.
2. Like car information, the ratings of dealers are also distributed among websites. Users are difficult to gather all rating and comments about a dealer with a convenient approach.

## Solution

My solution is building a search engine gathering car information, comments and ratings from different car trading websites and a website to display search result. Here's the steps:

1. **Dataset**. Building a web crawler to collect car information, dealer information, comments and ratings from big car trading websites, like cargurus.com, carfax.com and so on. For dealers who have their own websites, the crawler also collects car information from there.
    a) Considering the duplication of car information, I decide to use the VIN number of a car as its unique id to distinct the dataset.
    b) The amount of data is quite large, so the crawler only captures car information in the Bay Area.
2. **Search Engine**. Using Solr to store data. There are two types of documents. One is car document. The other is dealer document. A car document records the specs, price and any other additional information of a car. A dealer document stores information of a dealer, like name, address, contact, comments and rating. Users can search cars or dealers.
3. **Web Interface**. Building a website as the interface. If users want to search cars, they

need to fill a form of the specs of cars they want. If users want to search dealers, they can search dealer by names.

4. **Extra feature**. If I have more time, I want to build a feature where users can search cars with free text query. This feature may use NLTK to analyze text query and convert it into solr query.