Julia Eng
Alishba Khawaja
Safiyah Sadiq

Data Science Final Report

Link to live demo: http://julia-eng.github.io/498FinalProject/demo/index.html

Objective: To look at the stackoverflow data from the years 2007 to 2014 and analyse the trends seen in the usage of different languages over the years.

Data Sets:
We used the following links to gather our data:

- Unanswered questions and their tags: (Where from date and to date are a span of 1 year)
  https://api.stackexchange.com/docs/unanswered-questions#fromdate=2013-01-01&todate=2014-01-01&order=desc&sort=activity&tagged=coffeescript&filter=!*l7giggq.E5bVmXVQA8MjQMY&site=stackoverflow&run=true

- Tag count for all posts: (Where fromdate and todate are a span of 1 year)
  https://api.stackexchange.com/docs/questions#fromdate=2013-01-01&todate=2014-01-01&order=desc&sort=activity&tagged=coffeescript&filter=!*l7giggq.E5bVmXVQA8MjQMY&site=stackoverflow&run=true

- Freebase Query:
```
[{
 "id": null,
 "name": null,
 "introduced": null,
 "introduced>=": "2007-01-01",
 "type": "/computer/programming_language"
}]
```

This gives us the languages created since 2007. We will use this along with the number of stackoverflow tags to determine if these new languages gain popularity or die out over the past few years.

- We used Wikipedia to get list of all the languages. http://en.wikipedia.org/wiki/List_of_programming_languages

- Created an API using Kimono to extract all the languages from the Wikipedia web page and downloaded that as JSON. https://www.kimonolabs.com/apis/7c7nwyly

Tools, languages and API:

- Data Cleaning Freebase API
- D3
- Freebase API
- Kimono Labs
- Python (and Pandas and Numpy)
- SQLite 3
- Stack Exchange API

Results and Findings:

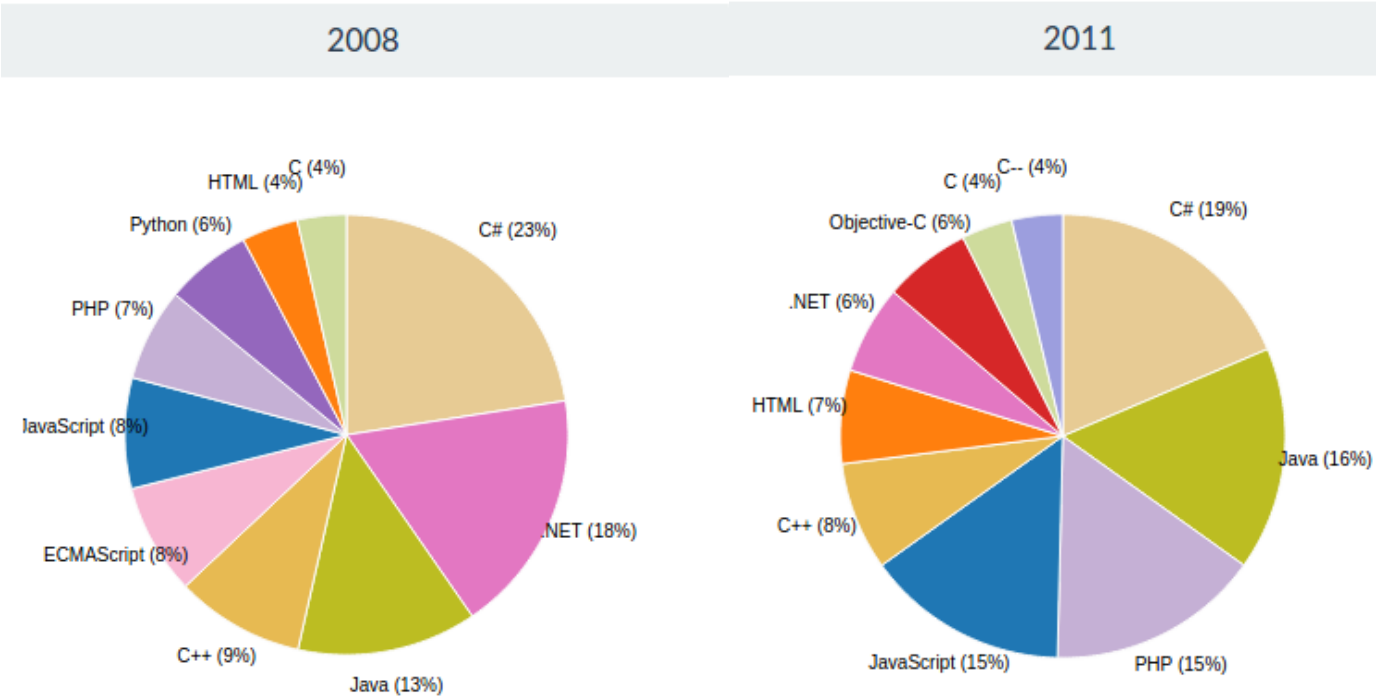
For our first task, from the years 2007 to our present year 2014, we calculated the top ten trending languages of each year. We based our calculation on the total number of questions asked about the language in each year. The total number of questions were derived from the number of times a tag was associated with a question.

In the first task, we first retrieved a list of **six-hundred ninety seven** programming languages from Wikipedia. We created an API to extract all the programming language names from the wikipedia page using Kimino Labs. With the API, we were able to extract all the languages to JSON format. After this, we had to do some manual data cleaning on the file to get rid of some unnecessary words associated with some of the language names.

Now using the cleaned JSON file, we wrote a python script that would retrieve information for each language; tag-name, year, total questions, unanswered question. All the data retrieved using the python script was exported to a CSV file.

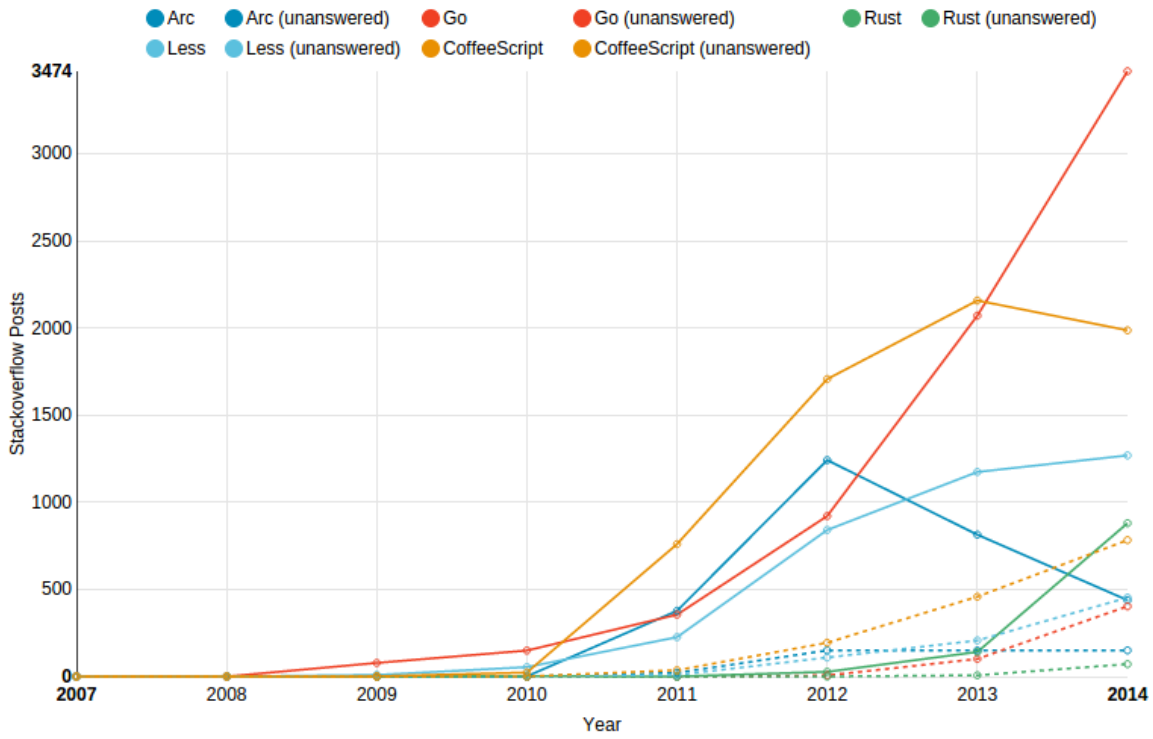
With the use of another python script, we used SQLite3 and inserted all the data into a database with the columns: tag-name, year, unanswered, and total, with the primary keys being the tag-name and year. Using the database we were able to query and order the data by the total and limit the results to ten languages per year, and then export the

results in each year to csv format. We displayed our findings using D3. The following is a preview of the results we found for each year and graphed using D3:



We found that most of the same languages were always in the top 10, but their ranking within the ten changed over the years. For example, C# (23%) and .NET (18%) were very popular in 2008, but as many programmers switched to client side web development in the last few years, we saw these two languages declined and Javascript usage doubled from 8% in 2008 to 19% in 2014.

In our second task, instead of looking at the trends for all the languages we wanted to look at the trends between new languages that had come out in the years 2007 to 2014. With the help of the Freebase API we got all the languages created between 2007 and 2014 into a JSON format. And then for each language we called the stack exchange API to find out the total number of questions asked and total number of unanswered questions for each year since 2007. Following is a preview of the results we found for each year and graphed using D3:



Language	Created	Influenced by	Description
Arc	2008	Lisp	Arc is a dialect of the Lisp programming language now under development by Paul Graham and Robert Morris.
Go	2009-	Limbo,	Go, also commonly referred to as golang, is a programming language initially