What StackOverflow Tells Us About Programming Languages

Rahul Thankachan Nada Aldarrab
Prathmesh Gat

University of Southern California

USC Viterbi
School of Engineering

Agenda

USC Viterbi
School of Engineering

Agenda

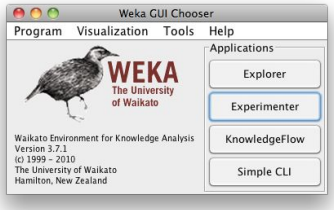
1. Introduction
2. Problem Statement
3. Temporal Based Trend Analysis
4. Topic Analysis
5. Predicting Time To Answer
6. Summary and Q&A

Introduction

USC Viterbi
School of Engineering

Introduction

- **Dataset Used** : Stackoverflow - Internet Archive.
- **Why?** It is one of the largest developer focused open collaborative platform currently.
- Through our study we intend to answers some interesting questions
- Study the rise and fall of popular programming languages
- Can be used to predict future enhancements
- Study effectiveness of Stack Overflow model



USC Viterbi
School of Engineering

Problem Definition

- Basic Analysis: What are the most popular programming languages?
- What are the trends in programming languages?
- What are the most popular topics discussed in a programming language?
- Can we accurately predict the time it takes until a questioner gets an answer?

Related Work

Miltiadis Allamanis and Charles Sutton. 2013. **Why, When and What: Analyzing Stack Overflow Questions by Topic, Type & Code** In *10th Working Conference on Mining Software Repositories. Mining Challenge*. IEEE, pages 53-56.

- Topic modeling analysis
- Used Latent Dirichlet Allocation (LDA)
- Modeled Java Topics of Questions
- Can *evaluate the orthogonality* of different languages
- Stack Overflow questions are *about the code and are not application domain specific*

TABLE VI: Percent (%) of questions asked on a specific day for various tags.

	Java	Java EE	Android	JDBC	Python	C#	RoR	SQL Server	C++	Maven	.NET	iPhone	XML	**All**
Mon	15.9	16.5	16.1	15.5	15.2	16.0	15.5	16.7	15.3	15.9	16.0	16.1	16.0	**15.6**
Tue	17.3	18.0	17.1	18.8	16.7	18.0	17.0	19.0	16.5	18.8	18.0	17.5	18.0	**17.4**
Wed	17.5	18.6	17.2	17.5	17.0	18.1	16.8	19.5	16.6	17.8	18.6	17.4	18.3	**17.6**
Thu	17.2	17.2	17.0	18.0	16.5	17.9	16.5	19.3	16.7	18.8	18.2	16.9	18.0	**17.4**
Fri	15.3	15.3	15.2	14.8	14.9	15.7	15.0	16.3	15.0	16.1	16.0	15.3	15.4	**15.7**
Sat	8.3	7.5	9.0	7.5	9.7	7.2	9.5	3.6	9.8	6.3	6.6	8.8	7.0	**8.3**
Sun	8.5	7.0	8.3	7.9	9.9	7.1	9.8	5.7	10.1	6.4	6.6	8.1	7.2	**8.1**

Related Work

V. Bhat, A. Gokhale, R. Jadhav, J. Pudipeddi, and L. Akoglu. Min (e) d your tags: Analysis of question response time in stackoverflow. In *Proceedings of ASONAM 2014*, pages 328–335. IEEE, 2014.

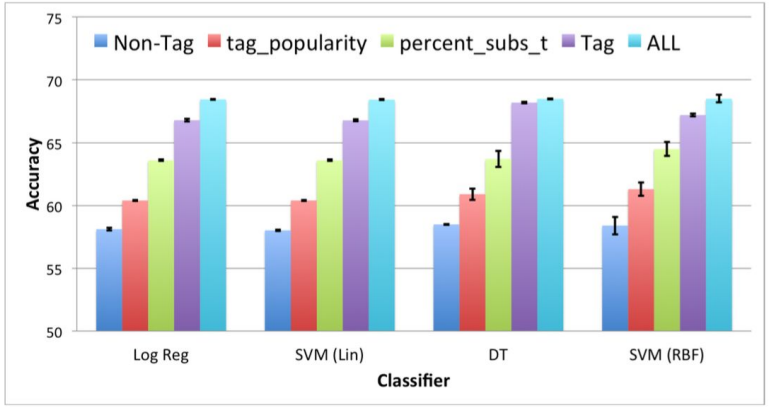
- Two linear classifiers: logistic regression and SVM with linear kernel
- Two non- linear classifiers: decision tree (DT) and SVM with radial basis function kernel

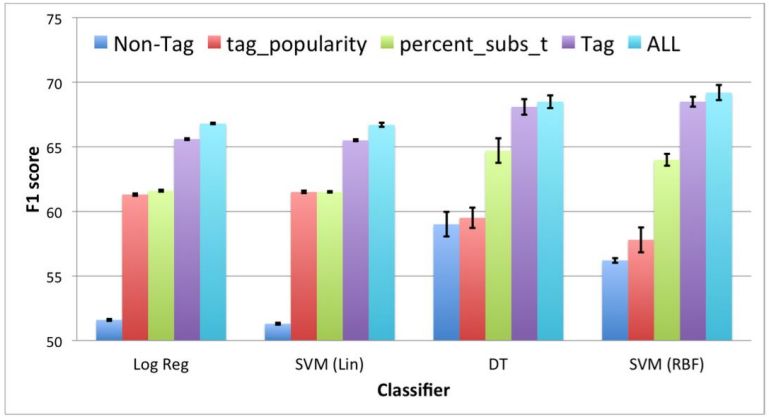
Tag based Question Features
tag_popularity: Average frequency of tags
num_pop_tags: Number of popular tags
tag_specificity: Average co-occurrence rate of tags
num_subs_ans: Number of active subscribers
percent_subs_ans: % of active subscribers
num_subs_t: Number of responsive subscribers
percent_subs_t: % of responsive subscribers

Non-tag based Question Features
num_code_snippet: Number of code segments
code_len: Total code length (in chars)
num_image: Number of images
body_len: Total body length (in chars)
title_len: Title length (in chars)
end_que_mark: Whether title ends with question mark
begin_que_word: Whether title starts with 'wh' word
is_weekend: Whether question posted on weekend
num_active_verb: Number of verbs that indicate action
num_selfref: Number of self references of the asker

Related Work

Prediction Accuracy:



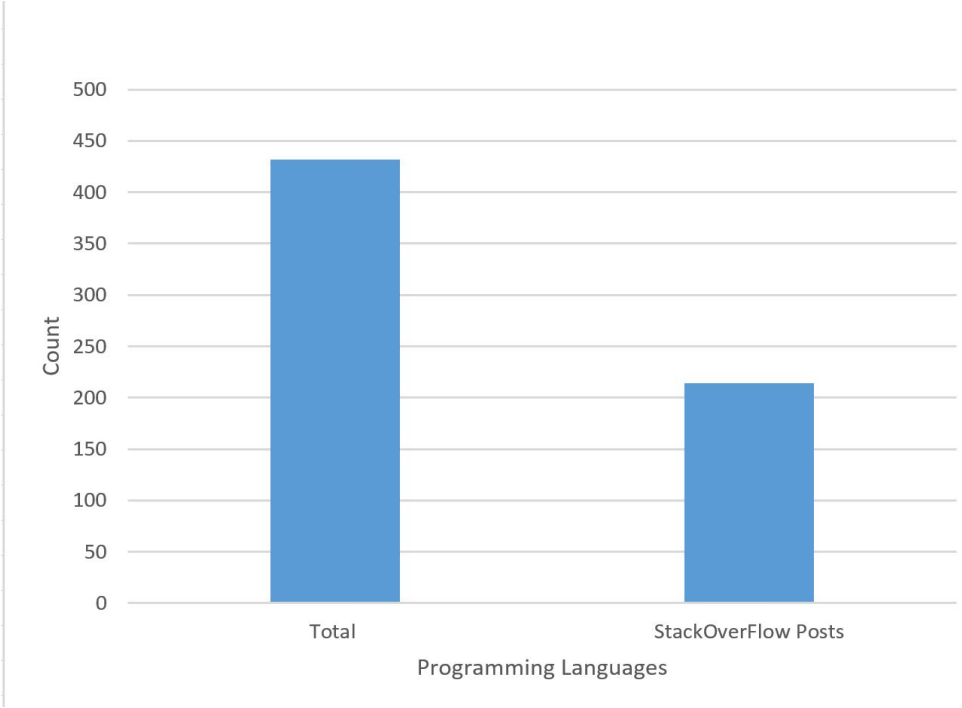


USC Viterbi
School of Engineering

Basic Analysis & Temporal Trends

USC Viterbi
School of Engineering

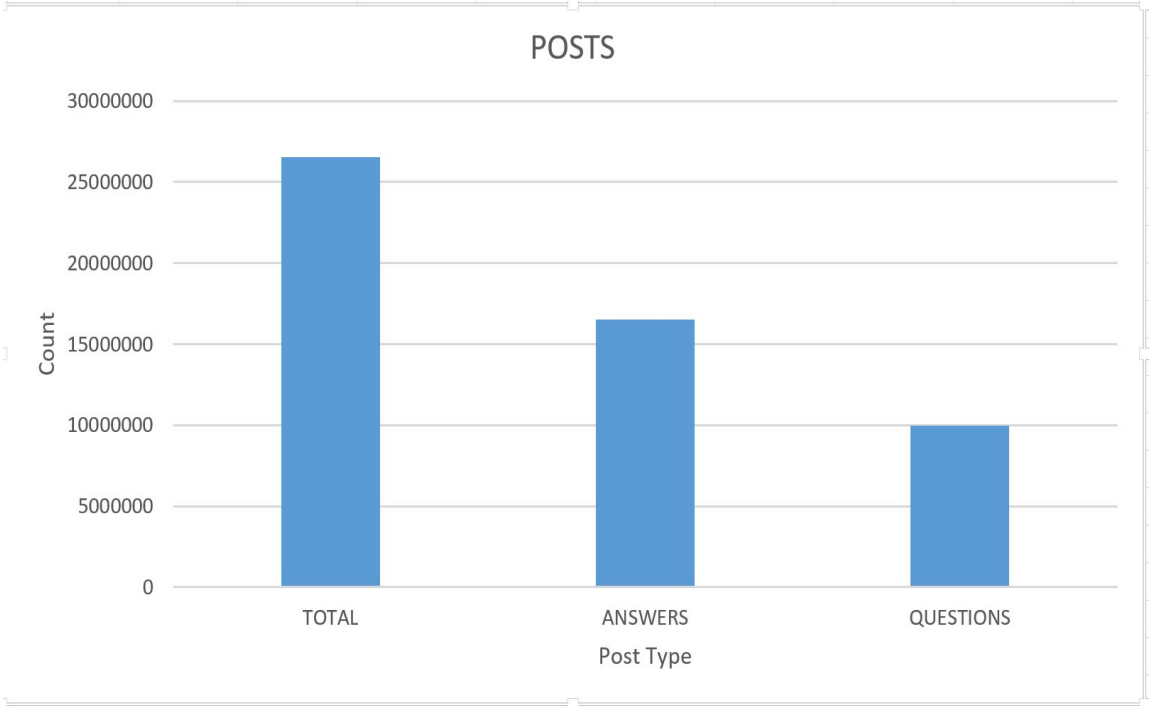
StackOverFlow Activity



USC Viterbi School of Engineering

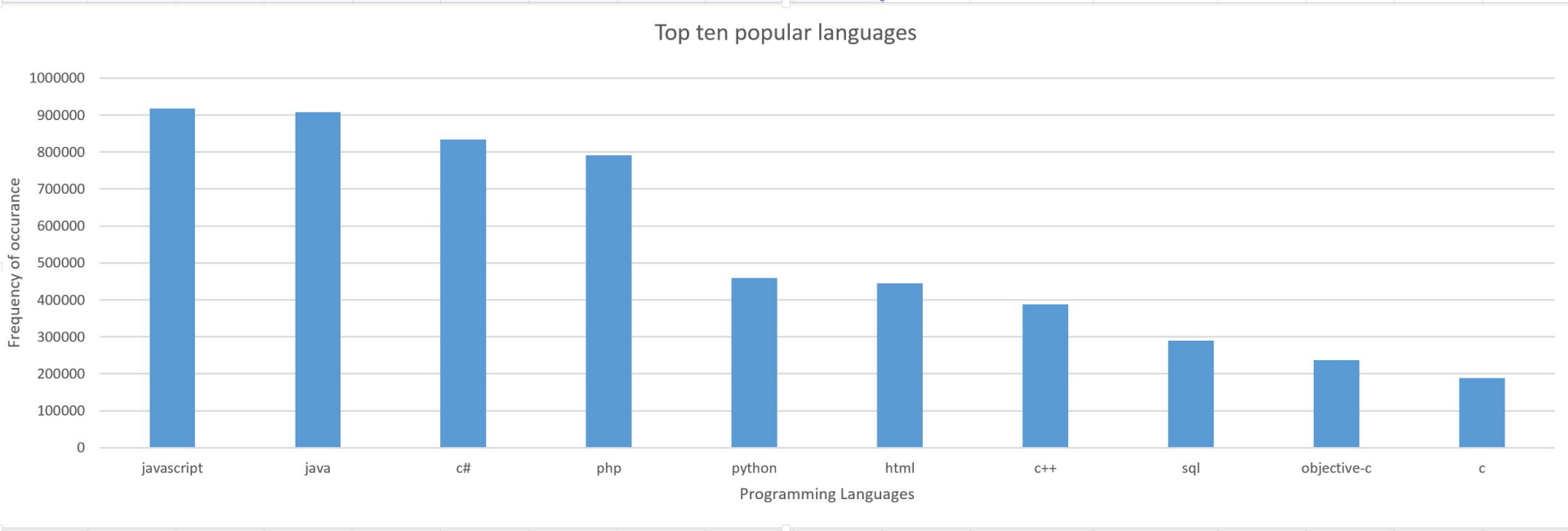
Approx. 45% of all programming languages in world are discussed on StackOverflow!

Post Type



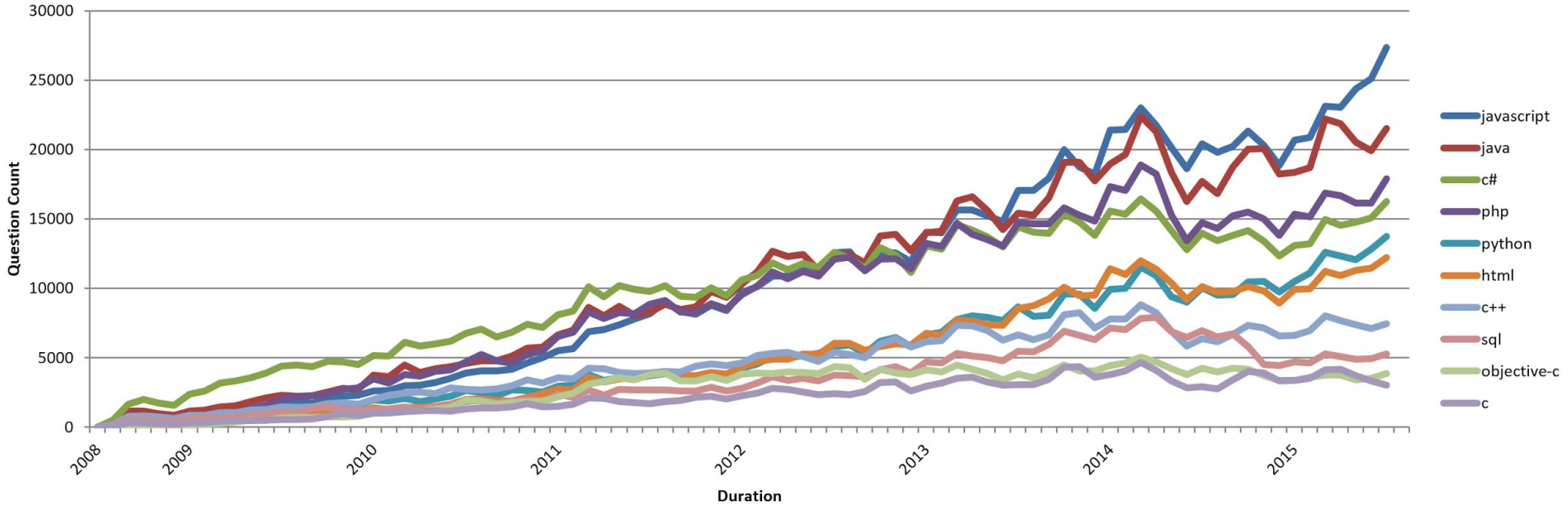
USC Viterbi
School of Engineering

Top Ten Languages



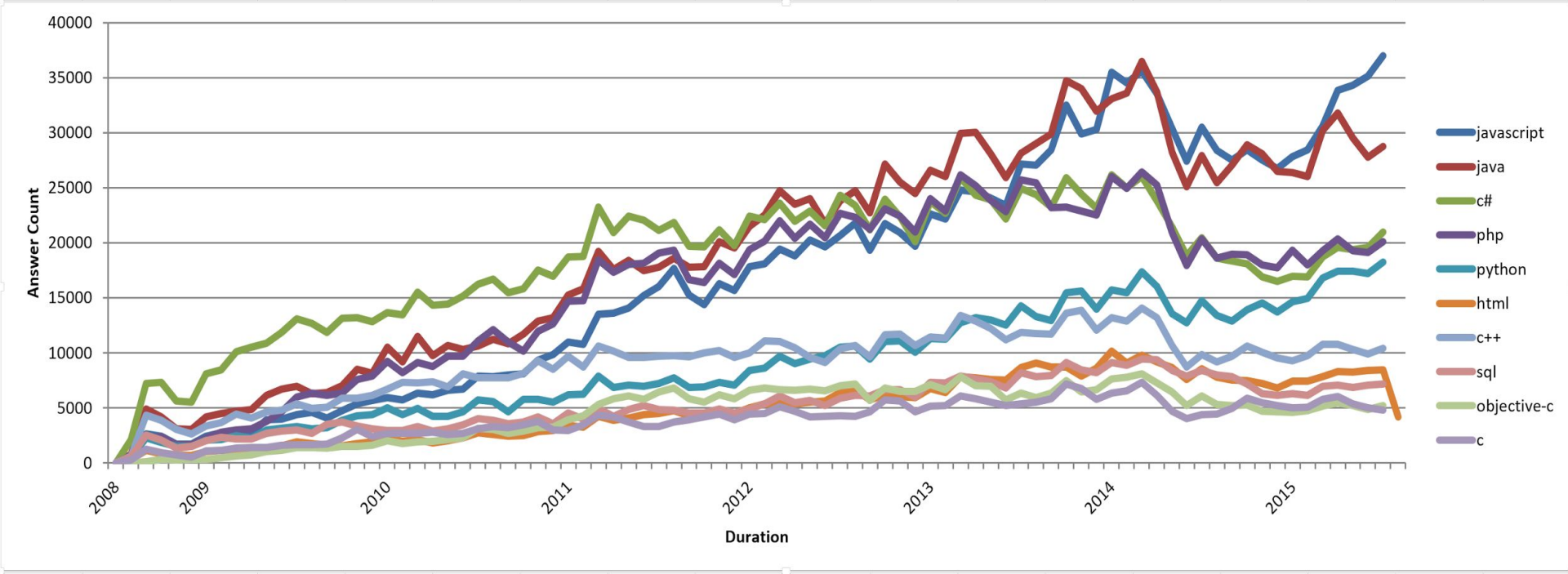
USC Viterbi
School of Engineering

Question Count



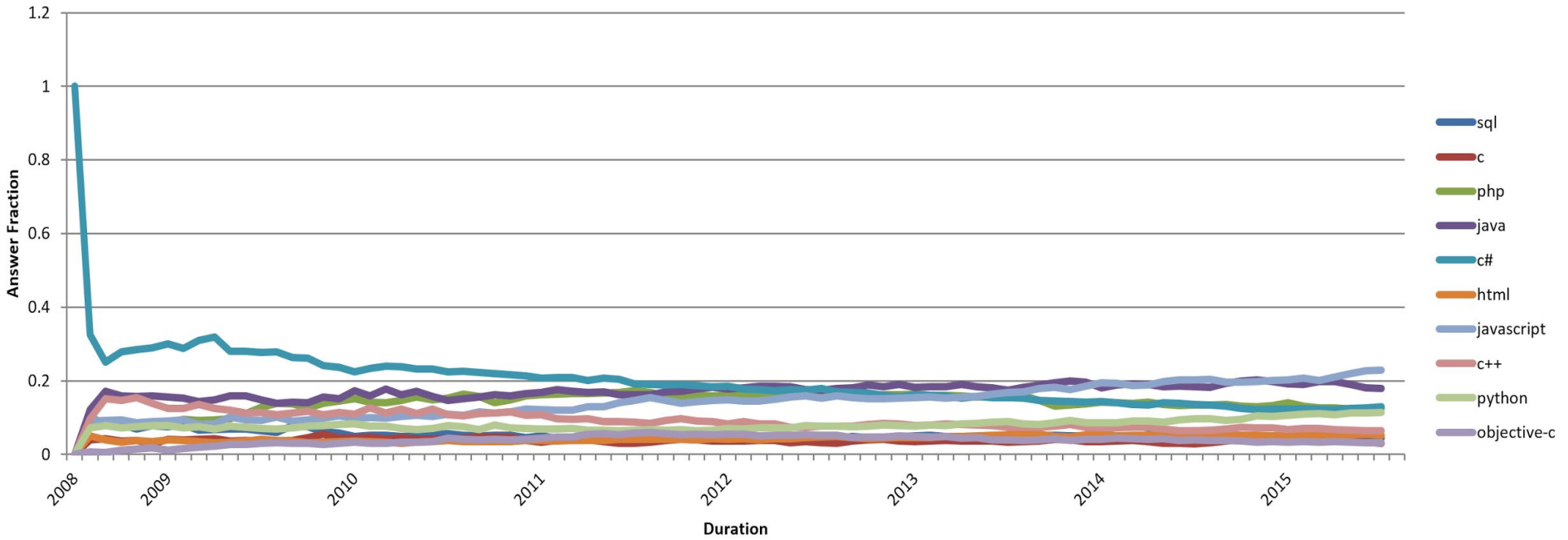
USC Viterbi
School of Engineering

Answer Count



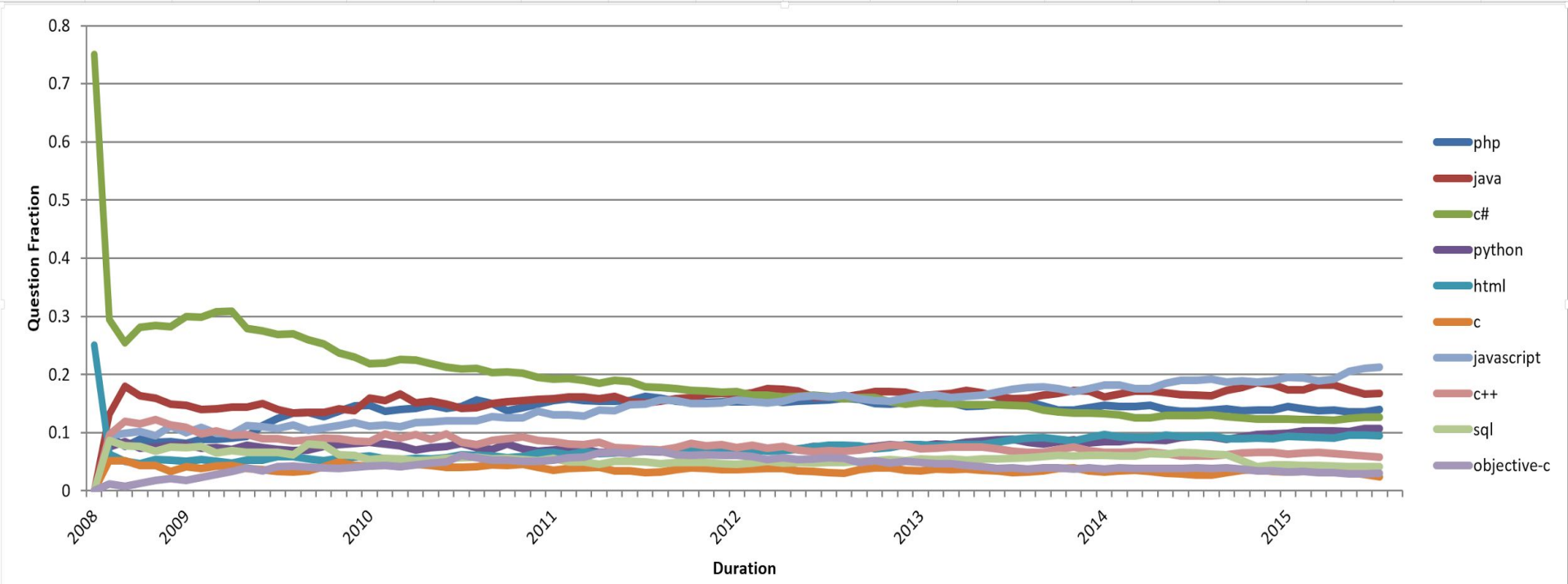
USC Viterbi
School of Engineering

Answer Fraction



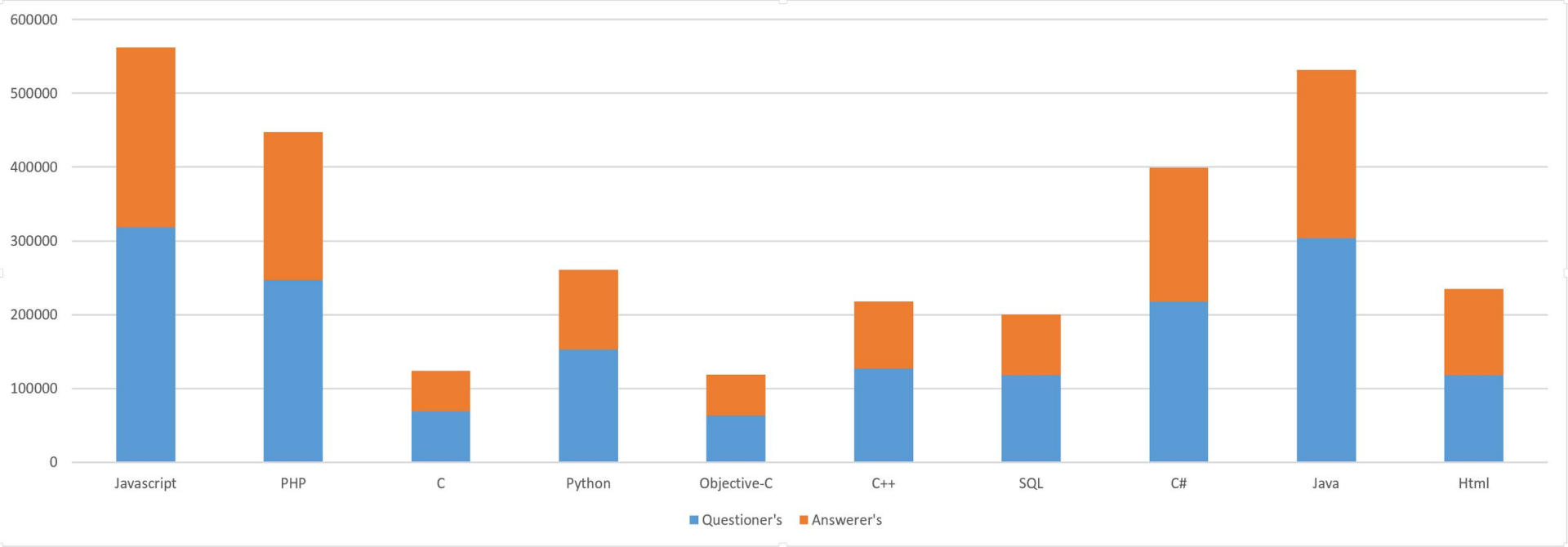
USC Viterbi
School of Engineering

Question Fraction



USC Viterbi
School of Engineering

Questioners/Answerers Distribution



USC Viterbi
School of Engineering

Topic Analysis

USC Viterbi
School of Engineering

Topic Analysis

	TOP TEN TOPICS									
Javascript	jquery	css	ajax	angularjs	html5	node.js	json	asp.net	arrays	regex
Java	android	swing	spring	eclipse	hibernate	arrays	multithreading	jsp	string	maven
C#	.net	asp.net	wpf	winforms	asp.net-mvc	linq	entity-framework	wcf	sql-server	multithreading
PHP	mysql	jquery	arrays	ajax	wordpress	codeigniter	regex	json	forms	apache
Python	django	python-2.7	numpy	python-3.x	list	pandas	regex	matplotlib	dictionary	google-app-engine
HTML	css	jquery	css3	html5	twitter-bootstrap	forms	ajax	asp.net	mysql	image
C++	c++11	qt	templates	boost	windows	arrays	pointers	winapi	visual-c++	opencv
SQL	mysql	sql-server	oracle	database	sql-server-2008	tsql	postgresql	join	sql-server-2005	asp.net
Objective-c	ios	iphone	xcode	cocoa	cocoa-touch	uitableview	ipad	osx	core-data	uiview
C	linux	pointers	arrays	gcc	string	struct	sockets	windows	multithreading	malloc

	Topics marking data structures
	Topics marking some hard aspect of a language

USC Viterbi
School of Engineering

Predicting time until Answer

USC Viterbi
School of Engineering

Approach

Following attributes selected for study:

1. Tag (Only top 10 programming languages)
2. Creation Month
3. Body Length
4. Tag Length
5. Introduced new Nominal class - Time_Answer
6. (less6, bet6and20, 20andmore)

USC Viterbi
School of Engineering

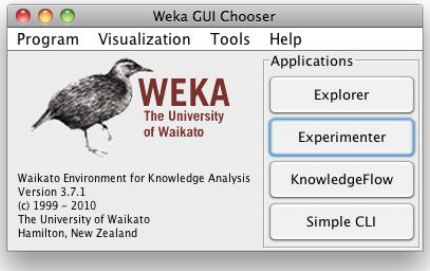
Approach

Tools

Weka - Weka is a collection of machine learning algorithms for data mining tasks.

Data Preprocessing:

Challenging!



Approach(Data Pre -processing)

1. Parse all the answers and link first answer's creation time to creation time of question. We called this field delta-answer.
2. Remove all the Questions which had delta answer negative or zero
3. We developed a Python script which develops .arff file On the fly (Wish to contribute this file)

USC Viterbi
School of Engineering

Evaluation

- Subset Size: 4490947 - Subset - 449000
- Classify response time into 3 types: less than 6 minutes, between 6 and 20 minutes, 20 minutes and more.
- 10-fold cross-validation
- Results are obtained using different feature combinations and different classifiers

Evaluation

Results of classifier J48 (all Attributes)

```
Correctly Classified Instances      212386               47.302  %
Incorrectly Classified Instances    236614               52.698  %
```

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.491     0.262      0.46      0.491     0.475      0.657    less6
                   0.026     0.02       0.343     0.026     0.049      0.522    bet6and20
                   0.77      0.56       0.484     0.77      0.594      0.642    20andmore
Weighted Avg.      0.473     0.314      0.437     0.473     0.403      0.613
```

```
=== Confusion Matrix ===

     a      b      c   <-- classified as
 68982   3210  68296 |      a = less6
 42074   3336  81122 |      b = bet6and20
 38746   3166 140068 |      c = 20andmore
```

USC Viterbi
School of Engineering

Evaluation

Results of classifier (body_length/ tag_length)

Correctly Classified Instances	361948	80.612 %
Incorrectly Classified Instances	87052	19.388 %

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.823	0.091	0.804	0.823	0.813	0.935	less6
	0.713	0.153	0.647	0.713	0.678	0.853	bet6and20
	0.858	0.036	0.942	0.858	0.898	0.971	20andmore
Weighted Avg.	0.806	0.086	0.816	0.806	0.81	0.926	

=== Confusion Matrix ===

a	b	c	<-- classified as
115563	24801	124	a = less6
26897	90194	9441	b = bet6and20
1306	24483	156191	c = 20andmore

Summary

- We were successfully able to find interesting temporal trends for major programming languages
- Using tag based topic analysis we were able to find major discussion topics and to some extent the difficult topics in a programming language
- Using machine learning techniques we were successfully able to predict - time to answer with good accuracy

Future Scope of Work

- Contribute the .arff on the fly generator script.
- Adding Parts of speech as an attribute
- Showcasing the results on a website

Questions?