PACC

PAUL ALLEN COMPUTING CHALLENGE

2013-2014

Data Science to the Rescue!

Welcome to the PACC

Thank you for your participation in the inaugural Paul Allen Computer Challenge. This packet should provide you with everything you need to know to help advertise the challenge to your students, help them form teams, understand the tasks in Part 1, and submit an entry.

Steps to Get Started

Read through the materials enclosed in this packet.

Post the included poster in your classroom.

Tell your students about the program, the amazing skills they can acquire, and cool prizes that can be won.

Help students form three person teams.

Give teams the Part 1 Tasks and Student Resource Document.

Assist with any questions or problems they may have.

Ask if you need more information or help with the competition contact Justin Spielmann, Living Computer Museum Education Coordinator, at: JustinS@livingcomputermuseum.org.

What is Data Science?

Advances in computing machinery have given us the ability to collect and store information at increasing scale. New technologies are also creating ways of *generating* information. For example, while personal and mobile computing devices enable us to have new kinds of interactions with technology and with each other, they also provide mechanisms for recording information about these interactions—e.g. what websites are being accessed online, who is clicking on which links, who is talking to whom within a social media site, what they are saying, etc. In other domains, new instruments are allowing us to collect massive amounts of data about everything from DNA mutations to planetary movements. All of this "big data" presents new challenges as well as exciting new possibilities.

Ideally, all of this data could give us new insight into all sorts of things: from old scientific questions about how the universe works to emerging new questions around human interaction at massive scale—e.g. what are we as humans capable of now that we can interact in social media spaces with millions of people all over the world? But before we can answer these, we need to figure out how to deal with all this information—the millions, billions and in some cases trillions of data points that any given instrument or device generates.

Data science is a growing field focused on finding ways to turn all of this data into knowledge: collecting, storing, processing, visualizing, and otherwise making sense of it all. Data scientists rely on a variety of skills, including computer science, mathematics, visualization, and qualitative skills of interpretation. And, because the problems that can benefit from big data analysis cover everything from astronomy to medicine to sociology to politics to website design, data scientists are increasingly in demand across professions and fields.

Sponsored by:

LIVING COMPUTER MUSEUM

Puget Sound Computer Science Teachers Association

csta Computer Science Teachers Association

University of Washington Computer Science & Engineering

HCDE Human Centered Design & Engineering University of Washington

eScience Institute

tableau SOFTWARE

PACC
PAUL ALLEN COMPUTING CHALLENGE

2013-2014
Data Science to the Rescue!

Information

Overview

The Paul Allen Computing Challenge is an annual challenge that brings together aspiring computer scientists in high school with their peers in order to answer questions about real world scenarios. This year the topic is using data from social media to examine Cyclone Phailin, which hit India in mid-October 2013. This competition is split into two parts:

1. A guided portion where students are provided a dataset and a list of tasks to complete in order to visualize and assess the data provided.

2. After being given a more robust dataset students can form their own hypotheses and provide answers by examining the dataset using the techniques they learned in Part 1.

The Paul Allen Computing Challenge is a collaboration between the Living Computer Museum, The Puget Sound Computer Science Teachers Association, the University of Washington Computer Science & Engineering Department, the University of Washington Human Centered Design and Engineering Department, and eScience Institute.

Prizes

Each student that completes both Parts 1 and 2 of the Paul Allen Computing Challenge will receive an Amazon Kindle pre-loaded with Paul Allen's Memoir *Idea Man: A Memoir by the Co-founder of Microsoft*, a PACC medal, a special edition PACC hoodie, and a Living Computer Museum T-shirt. Selected students will receive an upgrade to a Kindle Fire, also pre-loaded with *Idea Man*.

Format

Part 1. Part 1 provides students with an opportunity to examine a dataset of tweets focusing on Cyclone Phailin. This dataset will be provided by the Paul Allen Computing Challenge. The students will be given a list of tasks to perform on this dataset as well as questions to answer. Students are encouraged to form teams of 2-3 students in order to tackle this challenge. The tasks in Part 1 will allow students to explore and connect several fields including computer science, geography, current events, and data management. Teams may utilize any software package that helps them answers the questions asked, including Excel, Tableau, or custom data analysis software written by the students in the programming language of their choice. Student teams are encouraged to answer as many questions as possible, as completely as they are able. Partial credit will be given. Any team that submits a Part 1 entry will be registered and qualified for Part 2.

Part 2. Part 2 is a more independent project. Student teams will be given a more robust dataset of Twitter data from Cyclone Phailin (or another subsequent disaster event), assigned a University of Washington graduate student mentor, and with his or her guidance, will form a research hypothesis based on the dataset provided. Then, using the skills developed in Part 1 to analyze the data, teams will investigate their hypotheses. The

Sponsored by:

LIVING COMPUTER MUSEUM
Puget Sound Computer Science Teachers Association
csta Computer Science Teachers Association
University of Washington Computer Science & Engineering
HCDE Human Centered Design & Engineering University of Washington
eScience Institute
tableau SOFTWARE

culmination of this project will be the creation of a ~3 ft wide by ~4 ft tall poster summarizing their findings and methods, which the students will present as part of a poster session at Living Computer Museum in the late Spring of 2014.

Calendar

Mid November – December 13, 2013

Distribution of Challenge Packets.

December 14, 2013 - PSCSTA Winter Programming Contest

Part 1 Dataset released on Puget Sound Computer Science Teachers Assoc. website: http://www.pscsta.org/

March 28, 2014

Submission Deadline for Part 1.

April 26, 2014 – PSCSTA Spring Programming Contest

Part 2 Dataset released to qualified teams.

June 7, 2014

Poster session and awards ceremony at Living Computer Museum.

Rules

Answers for Part 1. Answers to Part 1 should have a cover sheet with the following information: each student's name, their high school, their Computer Science teacher's name, and their Computer Science teacher's email address. Please adhere to the Task 1 Answer Template format attached.

Datasets. The datasets provided to students for Parts 1 and 2 will be released according to the calendar above. Students are encouraged to review the Part 1 tasks listed in the accompanying document in order to familiarize themselves with the processes and techniques they will be asked to use.

Participation. Teams must be made up of only currently enrolled high school students in Washington State. Teams should be 2-3 students and all students that participate in a challenge should be identified in their submitted materials by name. All members of a Team from Part 1 do not have to participate in Part 2 for a team to continue to this second part of the competition. All students that participate in Part 2 of the Paul Allen Computer Challenge need to have participated and submitted answers to Part 1 to be eligible.

Software. Students can use any and all commercially available software to assess the datasets in Parts 1 and 2. No preference will be given for students that use more complex software. Most of the questions in Part 1 can be answered using Microsoft Excel or a similar spreadsheet program. Writing computer programs to assess the data may be helpful, but is not an essential component of Part 1 of the competition.

Assistance. High school teachers will be provided with a teacher resource document from which they can assist students with Part 1. This document will provide links to relevant online resources from which they can guide the progress of student teams. Teachers can also contact Justin Spielmann, LCM Education Coordinator, at JustinS@livingcomputermuseum.org with any questions or for general assistance.

For Part 2 students will be paired with a University of Washington Graduate Student mentor. This mentor will meet with the student team and help them frame a hypothesis and help them brainstorm possible ways to examine the dataset to search for answers. The Graduate Student will serve in an advisory capacity and will not actually assist with the creation of the Part 2 culminating poster.

Academic Honesty. Plagiarism of answers, programming code, or techniques will not be tolerated. If a participating Computer Science teacher, University of Washington Graduate Student mentor, or any PACC representative finds evidence of plagiarism or any type of academic dishonesty the student team will be eliminated from the competition.

PAUL ALLEN COMPUTING CHALLENGE

2013-2014
Data Science to the Rescue!

Part One Tasks

In the first part of the 2013-2014 Paul Allen Computing Challenge, your team will examine a data set of social media information about Cyclone Phailin that hit India on October 10-14, 2013. Below are the tasks for Part 1 of the competition. Your team should try to answer as many questions as possible but you don't have to complete every task to win prizes! In particular, Task 10 is an advanced question and should be attempted by teams that feel they have a thorough understanding of the material and processes in Tasks 1-9.

Note that your data set consists of 6,726 tweets. This was created from a larger set of 337,888 tweets. All 1,737 of the tweets with geolocation information were included in your data set. We also randomly selected an additional 4,989 tweets for inclusion.

The Submission Deadline is March 28, 2014

Submissions must be sent to

Living Computer Museum
c/o Justin Spielmann
2245 1st Ave S
Seattle, WA 98134

Task 1 – Visualization. Using the provided data set create a graph or chart that illustrates the volume of Twitter activity about Cyclone Phailin over the four days sampled.

Task 2 - Interpretation. Describe and attempt to explain any interesting observations about Cyclone Phailin Twitter activity (e.g., peaks or valleys, blackouts, etc.).

Task 3 - Evaluation. Here is a list of the top 10 hashtags (besides #phailin) used in the dataset: [#cyclonephailin, #odisha, #india, #cyclone, #news, #andhrapradesh, #phailinfury, #bhubaneswar, #gopalpur, #nari]. What is the frequency of each of these hashtags (i.e. how many tweets in the set contain #keyword)? What do these hashtags likely represent?

Task 4 -Evaluation. Does the frequency of specific hashtags changes over time? What could be the reason for these changes? You may want to create a graph or chart similar to the one in Task 1.

Task 5 - Filtering. Create a subset of the data that contains no retweets. Explain how you identified and eliminated retweets from the dataset. Describe your new dataset. Is it different in any interesting ways?

Task 6 - Filtering. From the subset created in Task 5 identify and remove all data that does not contain geotags (i.e., has no GPS data attached to it). Explain your process.

Task 7 - Interpretation. By completing Tasks 5 and 6 you have filtered the dataset. What sampling biases have been introduced to your new subset based on this filtering?

Task 8 - Application. Using your data subset from Task 6, map all the tweets that contain geolocation information.

Task 9 - Application. The Hurricane came ashore near Gopalpur on the East Coast of India. How many tweets in our set (that have GPS data) were sent from within 100 miles of Gopalpur?

How many tweets in our set (that have GPS data) were sent from within 100 miles of Mumbai (India's largest city - located on the West Coast)? How many tweets in our set (that have GPS data) were sent from within 100 miles of New Delhi (India's 2nd largest city - located in the North)?

Tell us how you calculated each.

What do these numbers tell us about Twitter use and disasters?

Task 10 – Advanced Task. Using information from all previous tasks create a time lapse video for the length of the collected data that shows where on the map tweets were sent from (like Task 8) and when they were sent.

Sponsored by:

LIVING COMPUTER MUSEUM

Puget Sound Computer Science Teachers Association

csta Computer Science Teachers Association

University of Washington Computer Science & Engineering

HCDE Human Centered Design & Engineering University of Washington

eScience Institute

tableau SOFTWARE

PACC
PAUL ALLEN COMPUTING CHALLENGE

2013-2014
Data Science to the Rescue!

Task 1 Sample Submission

Cover Letter

Please fill out the attached coversheet and include it for all Part 1 submissions.

Coversheets must be filled out in full and be legible to be considered in the competition.

Answers to Tasks

Answers to tasks should be organized logically, with a heading listing the task and then your provided answer. Figures, diagrams or tables should also be used if necessary. Refer to previous tasks if necessary in your answer. Answers should be as concise as possible and should only provide information relevant to the task.

See attached Sample for an example of answers for Tasks 1 and 2. This answer is not based on the dataset and is simply used as an illustrative example.

PAUL ALLEN COMPUTING CHALLENGE

2013-2014

Data Science to the Rescue!

Task 1

Submission

Names of Students on Team:

1. _____

2. _____

3. _____

High School name:

_____

Computer Science Teacher name:

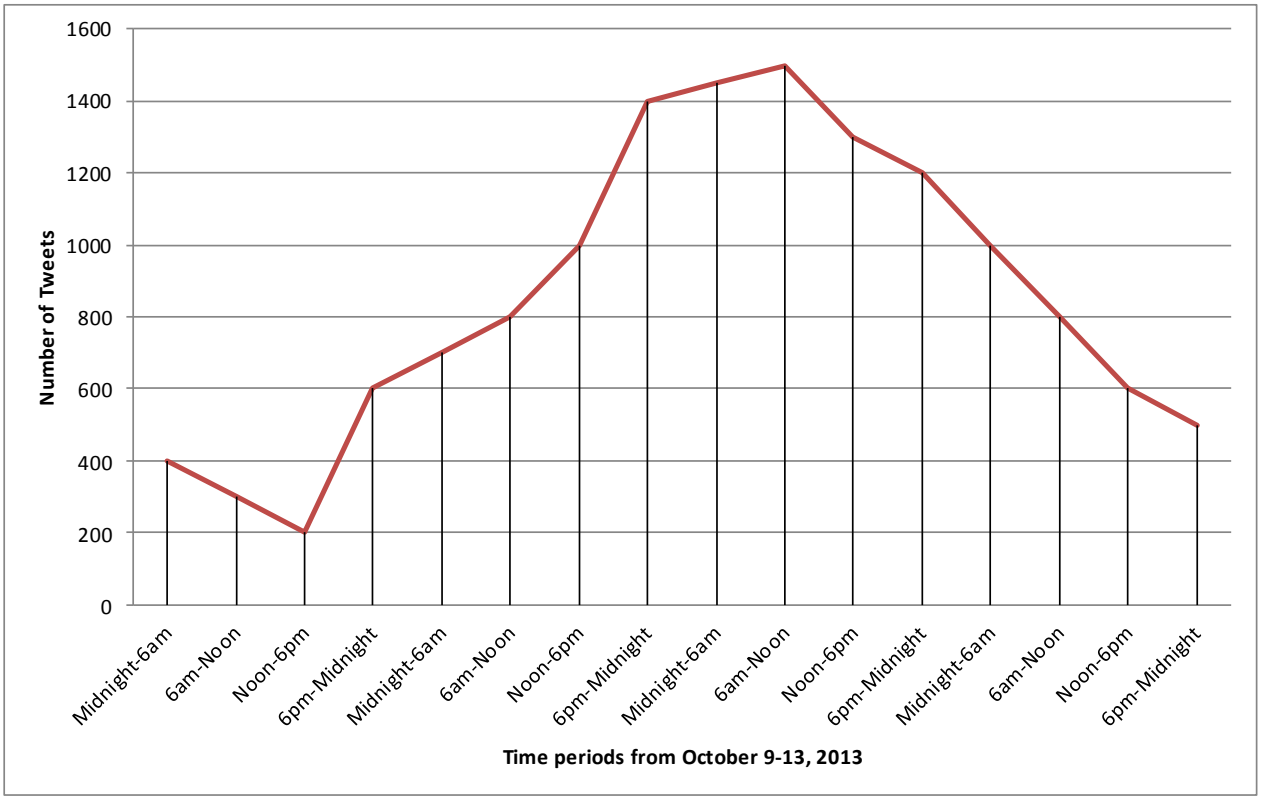
_____

Computer Science Teacher email:

_____

Paul Allen Computing Challenge—Part 1

Task 1



This graph demonstrates the number of tweets collected about Cyclone Phailin during 16 time periods (6 hour intervals over 4 days). The 6 hour interval was chosen because it accurately illustrates large scale trends in the data at a reasonable sample interval.

Task 2

Based on our visualization in Task 1 we did not see any blackouts during the four day period. This may be due to our 6 hour sampling interval, which would not allow us to see small scale (under 6 hour) periods of no information.

There is a clear valley, and relatively low Twitter commentary about Cyclone Phailin on October 9, as exemplified with the valley occurring on the evening of October 9. Discussion increases throughout October 10 and peaks at midday on October 11 and begins a slow tapering off throughout the rest of October 11 into October 12.

PACC
PAUL ALLEN COMPUTING CHALLENGE

2013-2014
Data Science to the Rescue!

Teacher and Student Resources

If you have additional questions about the competition or how to introduce your students to the concepts presented in the PACC contact Justin Spielmann, Education Coordinator at the Living Computer Museum, via email (JustinS@livingcomputermuseum.org).

Querying Data

SQLSHARE

SQLShare is an in-browser data storage and sharing tool, with premade queries. This will provide students with the ability to parse and examine data sets.
https://sqlshare.escience.washington.edu

Visualization

SQLShare visualization tool

http://sqlshare-graphs.herokuapp.com

Scott Murray

http://alignedleft.com/tutorials/d3/ provides basic tutorials on data visualization

Further expanded in his book *Interactive Data Visualization for the Web.*

Text Processing

Natural Language Processing with Python

Structured as a tutorial, that introduces Python, NLTK, and text processing. It is an excellent place to begin learning about text processing. http://nltk.org/book/

Python NLTK (Natural Language Toolkit)

Very easy to use with lots of tutorials. http://nltk.org/

An example Tweet analysis article using NLTK article
http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/

Advanced Text processing

Stanford CoreNLP Toolkit

http://nlp.stanford.edu/downloads/corenlp.shtml.

Online demo available at http://nlp.stanford.edu:8080/corenlp/

Sponsored by:

LIVING COMPUTER MUSEUM

Puget Sound Computer Science Teachers Association

csta Computer Science Teachers Association

University of Washington Computer Science & Engineering

HCDE Human Centered Design & Engineering University of Washington

eScience Institute

tableau SOFTWARE