

Text Classification via Supervised Machine Learning for an Issue Tracking System

Marty Kandes, Ph.D.

HPC User Services Group
San Diego Supercomputer Center
University of California, San Diego

HEPiX Spring 2019
Monday, March 25th, 2019
2:00 - 2:25PM PT

About Me

- ▶ Applied math, computational physics, HPC
- ▶ Distributed High-Throughput Computing Group @ SDSC

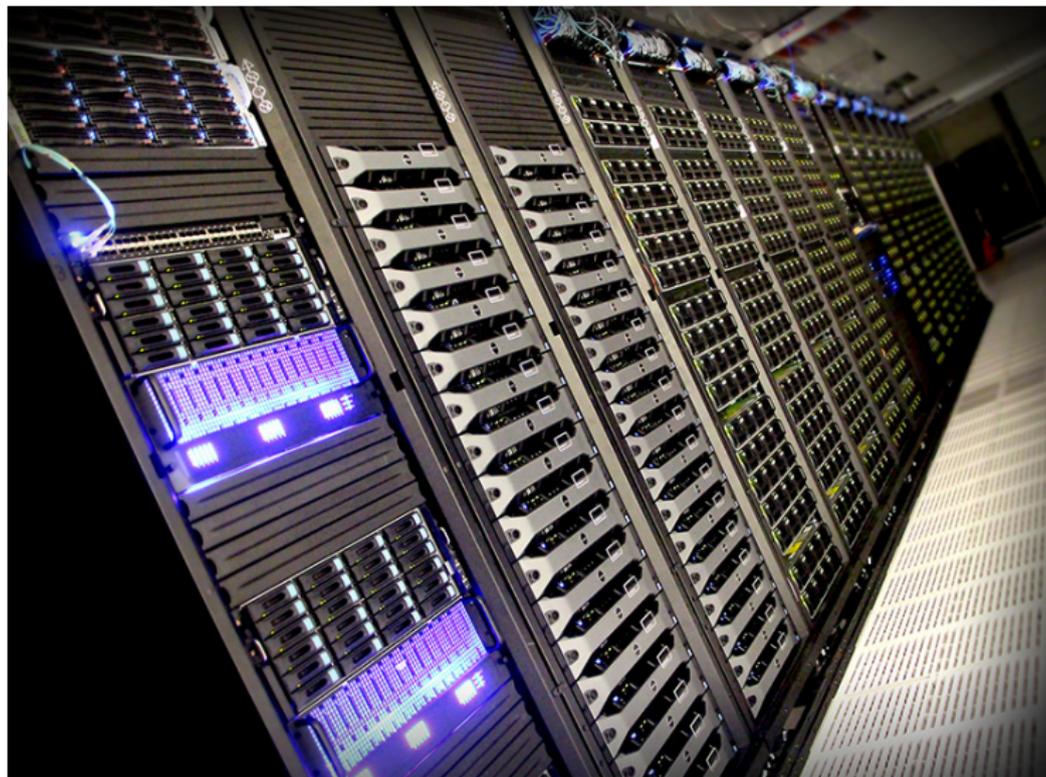
OSG / CMS GlideinWMS Factory Operations
OSG Software Team - Software Dev & Testing

- ▶ HPC User Services Group @ SDSC

A not-so long time ago in a data center not that far,
far away ...

- ▶ In 2012, 99% of all computational jobs run on NSF-funded HPC resources utilized fewer than 2048 CPU-cores, while accounting for approximately 50% of the total core-hours consumed across these resources.
- ▶ Nearly 70% of all jobs actually ran on only a single compute node (16 CPU-cores) or less.

Comet: A Supercomputer Built to Serve the 99%



Design Goals for Comet

- ▶ **Flexibility** - Ability to run a wide range of scientific and engineering applications to support complex, dynamic and multidisciplinary computational workflows.
- ▶ **Scalability** - Ability to support a large, diverse community of modestly-sized research projects that in aggregate represent a significant amount of research activity.

Comet By the Numbers

- ▶ **1944 compute nodes:** Dual-socket; 2.5 GHz Intel Xeon E5-2680v3 processors; 12 cores per processor; 128 GB DDR4 DRAM; 120 GB/s memory bandwidth; 320 GB SSD (210 GB Avail)
- ▶ **4 large-shared memory nodes:** Quad-socket; 2.2 GHz Intel Xeon E7-8860v3 processors; 16 cores per processor; 1.5 TB DDR4 DRAM; 400 GB SSD (260 GB Avail)
- ▶ **36 k80 gpu nodes:** Same as standard *compute* node, but with 2 PCIe-based NVIDIA Tesla K80 dual-GPU accelerators per node
- ▶ **36 p100 gpu nodes:** Dual-socket; 2.4 GHz Intel Xeon E5-2680v4 processors; 14 cores per processor; 128 GB DDR4 DRAM; 150 GB/s memory bandwidth; 400 GB SSD (260 GB Avail); 4 PCIe-based NVIDIA Tesla P100 GPU accelerators per node

2.76 Pflop/s

Comet By the Numbers

- ▶ **Interconnect:** Mellanox FDR (56Gbps) InfiniBand; hybrid fat-tree topology; rack-level (72 node) full bisection bandwidth; 4:1 oversubscription cross-rack bandwidth
- ▶ **Storage:** NSF-based \$HOME storage (100 GB per user); 6.4 PB 200 GB/s Lustre-based parallel filesystem storage (intermediate-term use: at least 500 GB per group allocation in /oasis/projects short-term use: up to 10 TB per user in /oasis/scratch)
- ▶ **Applications:** More than 173 software applications and libraries maintained and deployed via Rocks (Linux) cluster distribution; accessible to users via software modules; span a wide range of scientific disciplines, including, but not limited to, bioinformatics, chemistry, data analytics, engineering, fluid dynamics, mathematics, molecular dynamics, neuroscience, and statistics
- ▶ **Scientific Impact:** 1755 PIs; 358 institutions; 1144 research allocations; 4709 direct-access users; 33000+ gateway users; 997 publications

My Daily Grind: User Support Tickets



User Support Tickets via XSEDE RT

The screenshot displays the XSEDE RT (Request Tracker) web interface. The browser address bar shows the URL <https://tickets.xsede.org>. The page header includes navigation links for Home, Tickets, Tools, and a user login status for 'rickandes'. A 'New Ticket In' button and a search bar are also visible.

The main content area is divided into two sections: 'My Tickets' and 'Quick search'.

My Tickets

#	Subject Requesters	Status	Queue	Owner	Priority	Created	Last Updated	Time Left
8188	Problems with IntelBAM_Knem Lisa.Hero@sdscresearch.org	open	0-SOSC	rickandes	0	1 year ago	10 months ago	
87826	Memory Corruption Error running R on Comet Lisa.Hero@sdscresearch.org	open	0-SOSC	rickandes	0	11 months ago	3 months ago	
81661	SLURM Job, id=17834516 Failed with ExitCode 58 on Comet Lisa.Hero@sdscresearch.org	open	0-SOSC	rickandes	0	8 months ago	8 months ago	7 months ago
86258	Install Gibbon on the MATLAB on Comet rheo@sdsc.ucd.edu	user_wait	0-SOSC	rickandes	0	6 months ago	2 months ago	
99024	Another SegFault running R with MPI Lisa.Hero@sdscresearch.org	open	0-SOSC	rickandes	0	4 months ago	4 months ago	
99178	New SegFault R with Rfisc under OpenMPI Lisa.Hero@sdscresearch.org	open	0-SOSC	rickandes	0	4 months ago	4 months ago	
99004	Inquiry regarding a problem in submitting concurrently running executables within a single batch script on comet@SDSC ajvk_g@prf.com	user_wait	0-SOSC	rickandes	0	4 months ago	3 months ago	
100693	Can't run singularity on comet.sdsc.edu rheo@sdsc.ucd.edu	open	0-SOSC	rickandes	0	4 months ago	2 months ago	
103328	Need help to make an executable LAMMPS ml6@slu.se	user_wait	0-SOSC	rickandes	0	7 weeks ago	4 days ago	
103388	XUP: Using Ambar, having trouble getting my shell to be csh. mml@prf.com	open	0-SOSC	rickandes	0	7 weeks ago	4 days ago	

Prefined search Unowned Tickets not found

Quick search

Queue	new	open	user_wait	ingest	etc.	tech	etc.	vendor	etc.	internal_wait
0-Help	-	-	-	-	-	-	-	-	-	-
0-UI	-	4	-	-	-	-	-	-	-	-
0-Jetsream	4	13	30	-	-	-	-	-	-	5
0-LSU	-	9	-	-	-	1	-	-	-	-
0-NCSA	-	-	-	-	-	-	-	-	-	-
0-NCS	8	9	-	-	-	-	-	-	-	-
0-OSG	-	1	-	-	-	-	-	-	-	-
0-PRC	4	92	-	-	-	-	-	-	-	-
0-SOSC	-	32	58	-	-	-	-	-	-	17
0-Starford	-	-	-	-	-	-	-	-	-	-
0-TACC	4	160	59	-	-	-	-	-	-	5
0-XAlto	5	15	19	-	-	-	-	-	-	43
0-XOC	-	-	-	-	-	-	-	-	-	-
0-VRAS (non-ROSDC)	-	-	-	-	-	-	-	-	-	-
2.1.1-Community Engagement & Enrichment	-	1	14	-	-	-	-	-	-	-
2.1.2-Workforce Development	-	1	14	-	-	-	-	-	-	-
2.1.3-User Engagement	-	2	3	2	-	-	-	-	-	1
2.1.5-User Information & Interfaces	-	3	-	-	-	-	-	-	-	-
2.1.6-Campus Engagement	-	3	5	-	-	-	-	-	-	-
2.2-ECSS	-	5	14	-	-	-	-	-	-	-
2.2.3-Event & Innovative Projects	-	-	-	-	-	-	-	-	-	-
2.2.5-ECSS Science Gateways	-	8	1	-	-	-	-	-	-	-
2.3-XCI	-	4	-	-	-	-	-	-	-	-
2.3.2-XCIRACD	-	1	10	3	-	-	-	-	-	-
2.3.3-XCI Capability & Resource Integration	-	2	9	-	-	-	-	-	-	1
2.6-KSEDE Operations	-	-	-	-	-	-	-	-	-	-
2.6.1-Operations Director's Office	-	-	-	-	-	-	-	-	-	-
2.6.2-CyberSecurity	-	-	-	-	-	-	-	-	-	-
2.6.3-Data Transfer Services	-	2	3	1	-	-	-	-	-	-
2.6.5-Systems Operations Support	-	1	11	8	-	-	-	-	-	1
2.5.2-Allocations Process & Policies	107	240	1	-	-	-	-	-	-	-
3.5.3-Allocations, Accounting & Account Maint	-	-	-	-	-	-	-	-	-	-

Navigation icons are visible at the bottom of the page.

User Support Tickets via Zendesk

Mail - mkandes@sdsc... RT at a glance San Diego Supercomputing https://sdsc.zendesk.com/agent/dashboard

Dashboard Getting Started

Updates to your tickets

- Lauren Gilbert commented on "time to chat to see if HPC@UC meets our needs?". I won't be around after group meeting next week - I have a 4 PM flight out of L... Mar 13
- Rex Douglas commented on "time to chat to see if HPC@UC meets our needs?". That's ok, why don't we push to next week. I'm happy to help you in your research... Mar 13
- Lauren Gilbert commented on "time to chat to see if HPC@UC meets our needs?". Hello. Are we on for tomorrow? I have a doctor's appointment at 1, so I am sk... Mar 13
- Nicole Wolter commented on "time to chat to see if HPC@UC meets our needs?". Marty-I spoke with her and we went over the basics, SLUs was one of the things... Mar 12

Open Tickets (current) Ticket Statistics (this week)

0	33	0	0	2
YOU	GROUPS	GOOD	BAD	SOLVED

Tickets requiring your attention (14) [what is this?](#) Play

<input type="checkbox"/>	ID	Subject	Requester	Requester updated	Group	Assignee
<input type="checkbox"/>	#5035	Old account	Fernando de Sales	Yesterday 12:38	HPC Consulting	-
<input type="checkbox"/>	#5027	Gordon: FFW module linked to MPI?	Keaton Burns	Thursday 14:28	HPC Consulting	-
<input type="checkbox"/>	#4999	Question about perf version update and job monitoring	Hsin Yu Liu	Monday 16:42	HPC Consulting	-
<input type="checkbox"/>	#4886	New Comet Service Agreement	Max Mellette	Mar 16	HPC Consulting	-
<input type="checkbox"/>	#4762	jobs hanging in popeye	Daniel Angles-Alcazar	Mar 16	HPC Consulting	-
<input type="checkbox"/>	#4981	Prempt not working?	willascusa@fatroninstitute.org	Mar 16	HPC Consulting	-
<input type="checkbox"/>	#4778	FW: Access for lakoucheva's lab webpage in SDSC	Mike Dwyer	Feb 28	HPC Consulting	-
<input type="checkbox"/>	#4665	Re: [SDSC] Re: Re: [scicom] Re: [SDSC] Re: adding new simons remote login hosts	dsimon@fatroninstitute.org	Feb 20	HPC Consulting	-
<input type="checkbox"/>	#4663	Re: [scicom] Re: [SDSC] Re: adding new simons remote login hosts	dsimon@fatroninstitute.org	Feb 20	HPC Consulting	-
<input type="checkbox"/>	#4314	Weird problem with an application	Cesar Gomez	Jan 15	HPC Consulting	-
<input type="checkbox"/>	#3643	COMSOL on comet	Michael Getz	Oct 25, 2018	HPC Consulting	-
<input type="checkbox"/>	#2045	max GPUs per job	captian@predsci.com	Sep 07, 2018	HPC Consulting	-
<input type="checkbox"/>	#2549	LAMMPS simulations	QUENTIN FAURE	Aug 07, 2018	HPC Consulting	-
<input type="checkbox"/>	#2253	Issue on Comet with CMakes FindBoost module	Jamie Smith	Jul 18, 2018	HPC Consulting	-

User Support Tickets via Outlook

The screenshot shows the Outlook web interface for a user named XUP. The left-hand navigation pane is visible, showing folders like 'All folders', 'Inbox', and 'Sent items'. The main content area displays a list of support tickets from XUP. The tickets are listed in a table-like format with columns for subject, status, and time.

Subject	Status	Time
[tickets.usede.org #105687] XUP: Unable to access bin in /home/SUSER/ directory	Sat Mar 23 16:35:16 2019; Request 105687 was acted upon; Transaction: Correspondence added b	2:35 PM
[tickets.usede.org #105604] XUP: WRF run extremely slow after maintenance	Thu Mar 21 22:21:54 2019; Request 105604 was acted upon; Transaction: Correspondence added by shu@	Thu 8:22 PM
[tickets.usede.org #105598] XUP: DMTCP for checkpointing?	Thu Mar 21 19:23:26 2019; Request 105598 was acted upon; Transaction: Correspondence added by mahidhar; Queue: 0-5	Thu 5:24 PM
[tickets.usede.org #105687] XUP: Unable to access bin in /home/SUSER/ directory (6)	Sat Mar 23 16:35:16 2019; Request 105687 was acted upon; Transaction: Correspondence adde	2:35 PM
[tickets.usede.org #104637] XUP: job stuck? shows running, but no outputs from the middle of the simulation (2)	Sat Mar 23 13:33:19 2019; Request 104637 was acted upon; Trans	11:33 AM
[tickets.usede.org #105667] XUP: Software Compilation Help Request (2)	Fri Mar 22 18:36:34 2019; Request 105667 was acted upon; Transaction: Correspondence added by mahidh	Fri 4:37 PM
[tickets.usede.org #105674] XUP: Globus file transfer	Fri Mar 22 17:16:29 2019; Request 105674 was acted upon; Transaction: Correspondence added by mckandes; Queue: 0-SDSC Sub	Fri 3:16 PM
[tickets.usede.org #104930] XUP: issue regarding a submission in compute node and shared node (2)	Fri Mar 22 14:41:01 2019; Request 104930 was acted upon; Transaction: Corre	Fri 12:41 PM
[tickets.usede.org #105661] XUP: getting access to Gaussian	Fri Mar 22 14:38:28 2019; Request 105661 was acted upon; Transaction: Queue changed from 0-Help to 0-SDSC; by jhdell	Fri 12:19 PM
[tickets.usede.org #105476] MATLAB user group (Cornet)	Fri Mar 22 14:17:13 2019; Request 105476 was acted upon; Transaction: Correspondence added by imuller@uak.edu; Queue:	Fri 12:17 PM
[tickets.usede.org #105627] Comet SSH Help (2)	Fri Mar 22 13:29:36 2019; Request 105627 was acted upon; Transaction: Correspondence added by mahidhar; Queue: 0-SDSC Subject	Fri 11:30 AM
[tickets.usede.org #105616] time sensitive -- Best practice? (2)	Fri Mar 22 11:49:10 2019; Request 105616 was acted upon; Transaction: Correspondence added by mckandes; Queue: 0	Fri 9:53 AM
XSDC User News	Summer workshop opportunity for postdocs, doctoral students, and domain/computational/information scientists • A new message has been posted to XSDC User News. Categories	Fri 9:50 AM
[tickets.usede.org #105572] XUP: Lag in terminal window commands	Fri Mar 22 11:43:43 2019; Request 105572 was acted upon; Transaction: Correspondence added by ab6528@uark	Fri 9:47 AM
[tickets.usede.org #105276] XUP: Access VASP on comet	Thu Mar 21 23:50:57 2019; Request 105276 was acted upon; Transaction: Correspondence added by jgg; Queue: 0-SDSC Sub	Thu 5:51 PM
[tickets.usede.org #105004] XUP: WRF run extremely slow after maintenance	Thu Mar 21 23:07:11 2019; Request 105004 was acted upon; Transaction: Correspondence added by mahi	Thu 5:07 PM
[tickets.usede.org #105093] Can not access oasis-dm3.usede.org	Thu Mar 21 22:11:25 2019; Request 105093 was acted upon; Transaction: Correspondence added by mahidhar; Qu	Thu 8:12 PM
[tickets.usede.org #105620] XUP: Globus dffOutiles	Thu Mar 21 19:27:45 2019; Request 105620 was acted upon; Transaction: Correspondence added by mahidhar; Queue: 0-SDSC Subj	Thu 5:28 PM
[tickets.usede.org #105598] XUP: DMTCP for checkpointing?	Thu Mar 21 19:26:22 2019; Request 105598 was acted upon; Transaction: Correspondence added by flumbaca@exchange	Thu 5:27 PM
[tickets.usede.org #105617] XUP: JOBS always in priority status	Thu Mar 21 18:56:17 2019; Request 105617 was acted upon; Transaction: Correspondence added by mahidhar; Queue: 0	Thu 4:56 PM
[tickets.usede.org #105552] Error running on Comet: "error in locking authority file"	Thu Mar 21 18:39:47 2019; Request 105552 was acted upon; Transaction: Correspondence added b	Thu 4:40 PM
[tickets.usede.org #103496] Using Temperature Accelerated Molecular Dynamics (TAMD)	Thu Mar 21 17:42:57 2019; Request 103496 was acted upon; Transaction: Correspondence ad	Thu 4:43 PM
[tickets.usede.org #105611] XUP: Comet Trial Allocations	Thu Mar 21 17:12:35 2019; Request 105611 was acted upon; Transaction: Queue changed from 0-Help to 0-SDSC; by buvakul1	Thu 3:14 PM
[tickets.usede.org #105579] memory error help	Thu Mar 21 16:52:20 2019; Request 105579 was acted upon; Transaction: Correspondence added by mahidhar; Queue: 0-SDSC Subject;	Thu 2:53 PM
[tickets.usede.org #105609] XUP: May I apply for a Comet Trial Account	Thu Mar 21 16:34:31 2019; Request 105609 was acted upon; Transaction: Queue changed from 0-Help to 0-SD	Thu 2:35 PM

User Support Tickets by Example: Matlab Group

“ I am receiving and error message trying to run a matlab code.
Can you help me find a solution?”

User Support Tickets by Example: Gaussian Group

“I want to use gaussian09 on comet. I include the following in my batchscript:

```
module load gaussian/09.D.01  
/opt/gaussian/09.D.01/g09 < scoE_small_opt.com >  
scoE_small_opt.log
```

But i get an error saying permission denied. Could you please guide me on how to use gaussian09 for my calculation?”

User Support Tickets by Example: VASP Licencing

“Our group has a VASP licence with number: 5-321. Could you please give me access to VASP on Comet?”

User Support Tickets by Example: No \$HOME Directory

“When I gsissh to comet using SSO Hub I get an error ” Could not chdir to home directory /home/user008: No such file or directory”. I’m unable to work because of this. Please let me know how this issue can be resolved.”

User Support Tickets by Example: git old SSL libs

“I am writing to ask about an ssl error that occurs when I try to clone git repositories from github into my comet account. The error and context are below:

```
git clone https://github.com/words-sdsc/modular_deep_learning.git
Cloning into 'modular_deep_learning'...
fatal: unable to access
'https://github.com/words-sdsc/modular_deep_learning.git': SSL
connect error
```

Do I need to load some specific module to fix the error?”

User Support Tickets by Example: Slurm Down

“I am trying to submit a job using sbatch and it does not work apparently. It gives me the following:

```
sbatch batch_job.sb
```

```
sbatch: error: slurm_receive_msg: Socket timed out on send/rcv  
operation
```

```
sbatch: error: Batch job submission failed: Socket timed out on  
send/rcv operation
```

queue is not working either. Do you know if anything is wrong over there?”

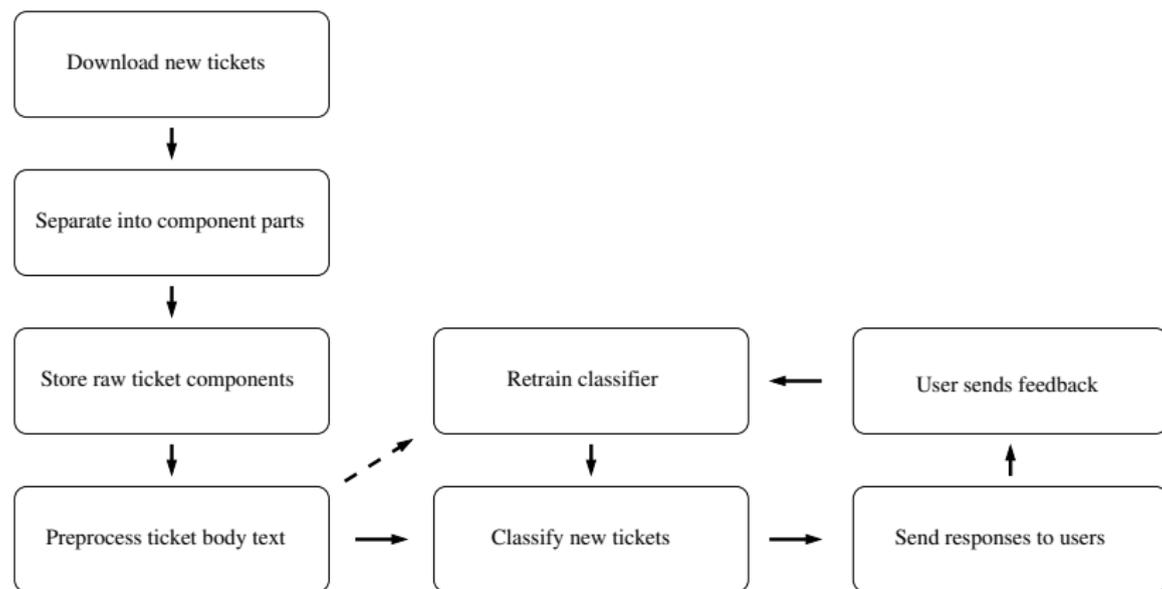
The Fundamental Question(s)

- ▶ What if we could utilize machine learning algorithms to classify and answer common user questions as they arrive?
- ▶ In an efficient classification scheme can be developed, can we provide useful, generic automated email responses to users to help them resolve their issues more quickly?
- ▶ How much time would such a system save the us?

Project Objective(s)

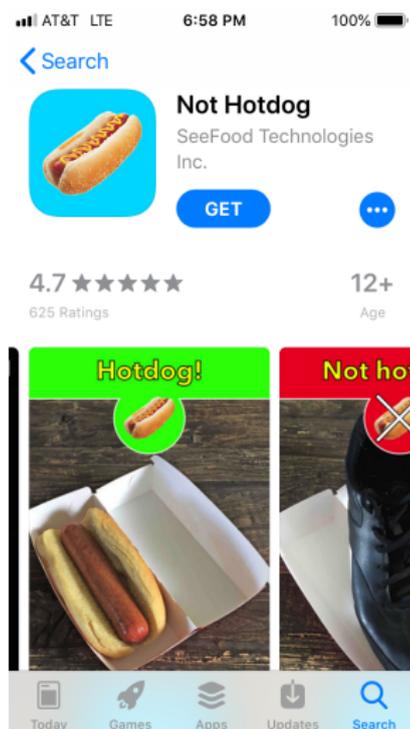
- ▶ Utilize supervised machine learning algorithms to classify common user support tickets with greater than 90% accuracy.
- ▶ For classifiable tickets, provide automated email responses to users (during non-business hours) within minutes of arrival.
- ▶ Focus on only initial ticket responses.
- ▶ Incorporate user feedback mechanism to help improve training dataset labels over time.

Classification and Response Workflow



Naive Bayes Classifier

- ▶ Conditional probability model for text categorization
- ▶ First introduced and studied in early 1960s
- ▶ Classifies documents as belonging to one category or another based on word frequencies
- ▶ Assumes feature independence for each given class
- ▶ Advantage: Only requires small amount of training data



Naive Bayes Classifier

For each word frequency vector $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$ representing a document to be classified, compute the conditional probability that the document belongs to category c_k

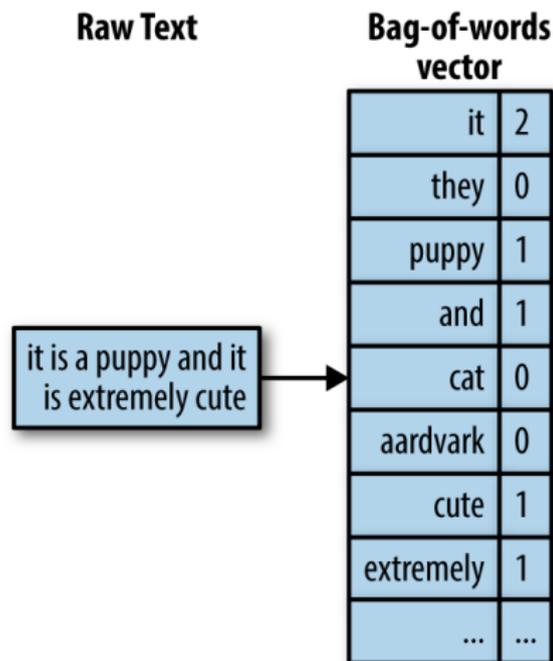
$$p(c_k|\mathbf{x}) = \frac{p(c_k)p(\mathbf{x}|c_k)}{p(\mathbf{x})}$$

for each of the K possible outcomes or classes $\mathbf{c} = \{c_1, c_2, \dots, c_K\}$.

Assign document to category c_k when $p(c_k|\mathbf{x})$ is the highest probability in the set and above some minimum acceptable cutoff.

Bag-of-Words Model

- ▶ Simple representation for natural language processing
- ▶ Any text is represented as a multiset of its words
- ▶ Count frequency of occurrence of each word in a text
- ▶ Ignore grammar and word order



Blacklisting

Remove most commonly used words

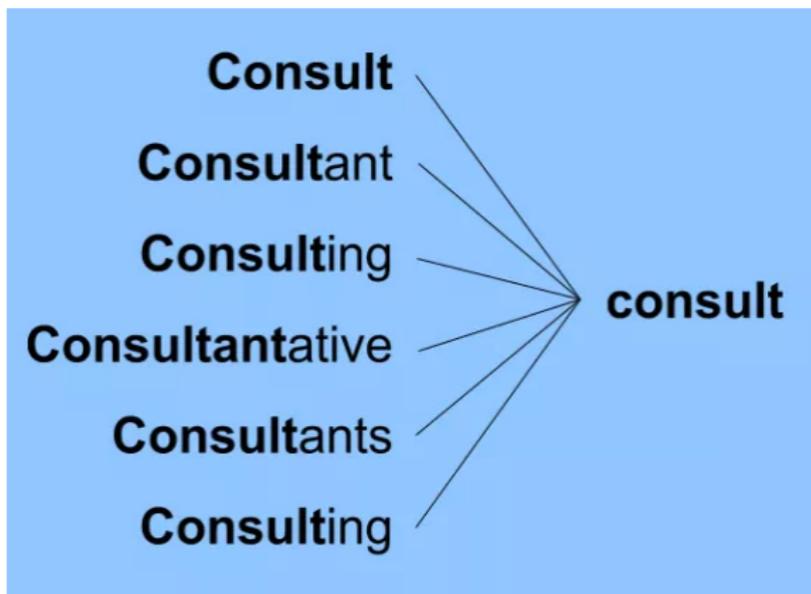
The screenshot shows a spreadsheet with the following data:

	A	B	C	D	E	F	G
1	the	1					
2	be	2 am	m	is	isn	s	are
3	and	3 ands					
4	of	4					
5	to	5					
6	a	6 an					
7	in	7 ins					
8	have	8 haven	ve	has	hasn	had	hadn
9	it	9 its					
10	you	10 y	ye	your	yours		
11	he	11 him	im	his			
12	for	12					
13	they	13 them	em	their	theirs		
14	not	14 t					
15	that	15 those					
16	we	16 us	our	ours			
17	on	17					
18	with	18					
19	this	19 these					
20	i	20 me	my				
21	do	21 does	don	down	did	didn	doinn

New General Service List is a list of approximately 2,800 of the most important high frequency words in the English language

Stemming

Process of removing all modifiers of a keyword including prefix, suffix, and pluralization until only the root of the word remains.



Porter Stemming Algorithm: Removes suffixes

Prototyping: Summer 2018

Text Classification via Supervised Machine Learning for an Issue Tracking System



N. CLARK¹, R. A. MARTIN², D. H. WU³, AND M. C. KANDES⁴

UC San Diego

¹ La Jolla Country Day School ² Canyon Crest Academy ³ Pacific Ridge High School
⁴ San Diego Supercomputer Center

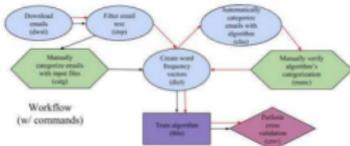
Abstract

The objective of our project was to use a text classification algorithm to categorize support ticket emails into common issues. We implemented our project in python, employing regular expressions and the imap library to download emails. After building a classified database of emails, we researched the process of building and training a text classification algorithm, focusing on the Naive Bayes text classification algorithm, a simple but powerful probabilistic algorithm that makes the naive assumption that words in the sentence are independent from one another. Gradually, we developed our own version of Naive Bayes and a scheme for collecting and processing email data and finally compared other implementations of text classification algorithms to our own. Here, we discuss our process and initial results.

Introduction

For each scientist that innovates in his or her respective field with the help of SDSC's supercomputers, someone has to help navigate him/her through the often confusing and error-prone world of software development. Each day, users encounter technical issues and send support ticket emails, and its up to a few help desk employees to point them in the right direction. Their work is often repetitive as users frequently run into common issues. We studied different text classification and python techniques involving regular expressions, stemming, and file organization, among other things. Our goal was to conduct research on text classification, so that we could design a machine learning algorithm to categorize support tickets pertaining to the 20 most common issues encountered by SDSC users and relieve user services of handling simple errors.

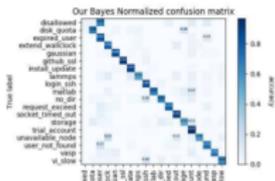
Methods



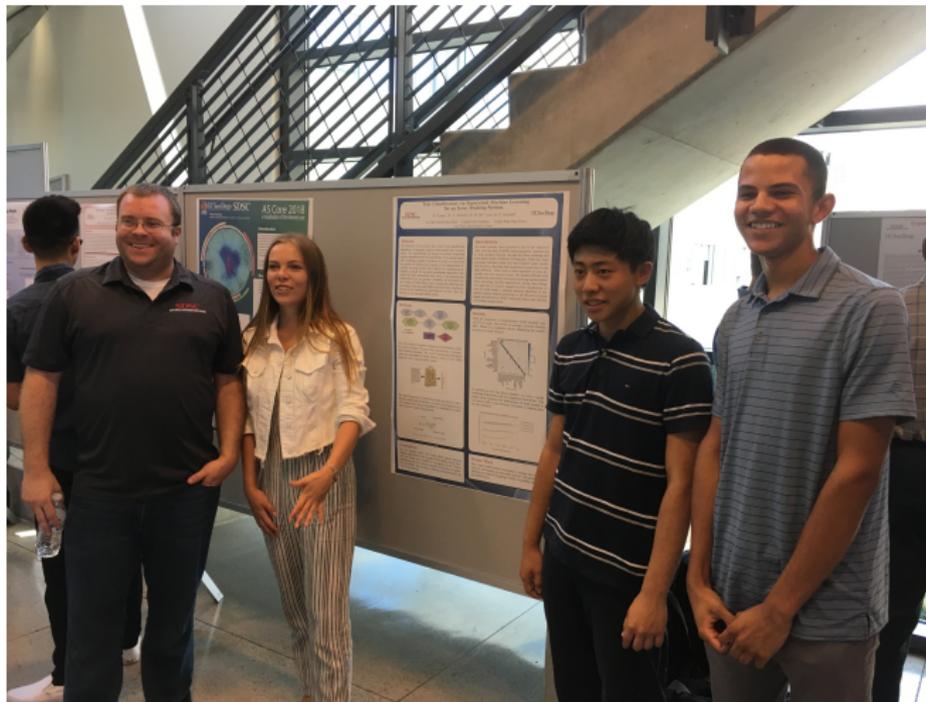
The Naive Bayes Classifier technique is probabilistic classifier based on Bayes' theorem with assumptions of independence between features. The "naive" assumption disregards order and placement of words within a body of text—often referred to as a 'bag of words' model.

Results

With 20 categories, a randomization would correctly classify 5% of emails. Our model, on average, correctly classifies 80%. Below is a confusion matrix, displaying the model's accuracy for each category.

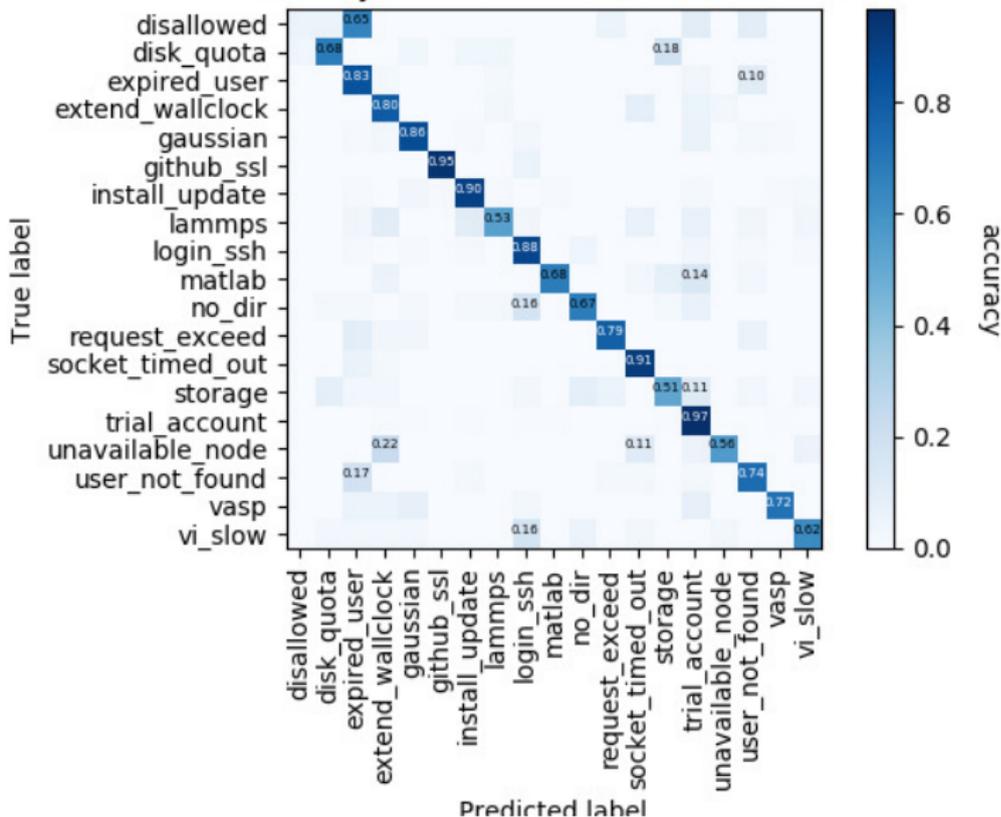


Research Experience for High School Students



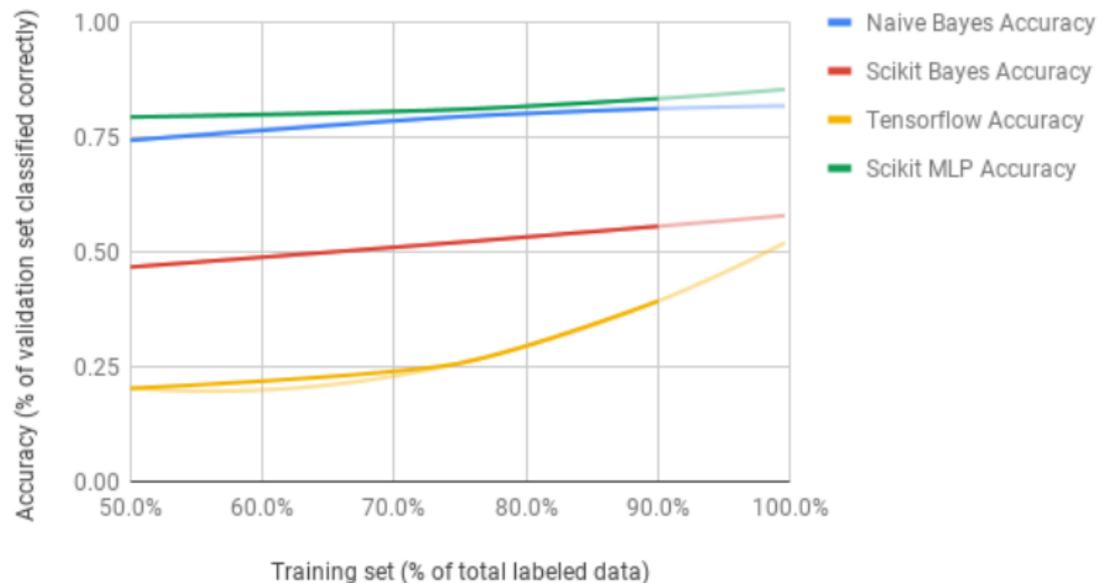
Preliminary Results

Our Bayes Normalized confusion matrix



Preliminary Results

Text Classification Cross Validation Accuracy



Current Status and Future Work

- ▶ Still in prototyping phase; working towards minimum viable prototype for production test
- ▶ Developing sqlite3 database schema; plan to complete and implement prior to REHS 2019
- ▶ Developing generic python class to help simplify sqlite3 database management
- ▶ Continue to label and increase raw training dataset size
- ▶ Integrate XOAuth2 support with Outlook Mail API
- ▶ Explore multilayer perceptron (MLP) and support vector machine (SVM) approaches more completely
- ▶ Explore stemming algorithms; e.g., algorithms for prefix and pluralization removal
- ▶ Explore text segmentation algorithms; e.g., in-body metadata removal and other personally identifiable information

Questions?

