Sentiment Analysis and Data visualization

SLAC NATIONAL ACCELERATOR LABORATORY

Spencer Rodrigues[1], Nikola Stojanovic [2], Elizabeth Goodwin[2], Joe Robinson[2]

[1]Statistics Department, California Polytechnic State University, 1 Grande Ave, San Luis Obispo, CA 93407, USA

[2]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA

contacts: stojanik@slac.Stanford.edu egoodwin@slac.Stanford.edu jsrob@slac.Stanford.edu

Background

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral.

Problem Statement

This project addresses the need for quick and accurate analysis of user's reviews from their time in LCLS. This is necessary because we ask them many questions and when we receive their responses we want to be able to pick out which responses are negative and discuss how we can improve in those areas. By having a program run through the responses that we get we will also be saving a lot of time.

Program

This was all achieved by using the statistical software called R. We had to download several different packages to be able to achieve our goal. The packages that were downloaded were: sentiment, tm, NLP, and Rstem. The function that we used is called Qdap and takes the string as an input. The output that we are given rates the text as positive, negative, or neutral. It also gives a polarity score, which determines the strength of the rating as mentioned above. If we cared about the emotions portrayed in the review we could also apply a different function that has the capability of returning the main emotion portrayed in the review, this function is called classify_emotion.

Question Id	Title	Text Input	Sentiment
9	Overall Beam Performance and Quality (Beam Stability, Noise, Drift)	Shz endstation A was giving problems due to the switching. After discussions with operations, we were able to cancel delivery to endstation A and continue our work with normal b..	positive
		A power outage cost us 16 hours of beamtime.	positive
		After one hour of adjustment we could first use the beam. This should not count against the first user as it is necessary. See comments above.	negative
		Apart from an issue with a two-wavelength structure, the beam stability in terms of energy and stability was excellent	positive
		Apart from several short beam drops, everything was working very well. We asked for a change in pulse length as well as beam energy during the last shift which also went smoot..	neutral
		As always, there is room for improvement in terms of x-ray stability!	positive
		As far as realistic parameters are concerned it was excellent.	neutral
		As I felt through communication with the team.	positive
		As long as these features are well-understood and well-documented, they can be accounted for in post-processing. The beam, to my understanding, worked about as well as exp..	negative
		as said, gun problems two times.	negative
		As stated above, poynting instability caused some challenges.	negative
		Beam at 9.5 keV was the best of all the times we have been to LCLS.	positive
		Beam is as stable as possible (fluctuations) , if circular polarization would be available we would have been able to perform single shot Holographic imaging with circular polarized..	positive
		Beam performance and quality seemed good. However, during this PCS time, we lost the beam for ~3 hours (The beam was down for ~6 hrs total on that date). It seeded that wa..	negative
		Beam performed to within the expected parameters. Beam intensity stability fluctuated from shot to shot (expected) resulted in variation in the x-ray diffraction signals captured. Th..	positive
		Beam was mostly very stable during the experiment, but showed a higher than usual timing jitter with a bimodal distribution separated by 600 fs	negative
		Beam was quite stable except part of the second time. Overall very satisfied withe x-ray beam and beamline support.	positive
		Beam was unexpectedly down for about 5 hours.	negative
		Beam was unstable at times and we lost some time.	negative
		By beam we mean the LCLS. The RPA of the IR laser had some issues, but they were solved during beamtime.	negative
		Due to a fire in the SLAC accelerator short before the LDS7 beamtime (July 17) we have had a very unstable and quite low energy beam. We would have preferred stable and low ..	negative
		Due to a hardware malfunction LCLS was down for most of the last shift. However, we did get a few extra hours to partially make up for that. Thank you!	negative
		Due to problems apparently related to the unusually cold weather (with low temperatures unprecedented during LCLS running, nearly so for the LINAC), the FEL performance was..	negative
		Due to the hot weather (over 100 degree), our day-shift beam ..	positive
		During our run, problems with the L1X klystron caused frequent beam instability. A problem with a PPS system caused 5 hours of lost beamtime. At the end of the run the LCLS st..	negative
		Even though tuning times were extremely long, when we had the beam the delivered pulse energy, photon energy, time separation, energy separation was conform to specs discu..	negative
		Excellent beam stability, we had close to 80% of the shots within the narrow monochromator bandwidth	neutral
		Excellent Linac performance; excellent operation in seeding mode; tuning times (energy changes/seeding) appropriate; shorter tuning times desireable.	positive
		excellent up time. A pity it takes a while to change photon energy.	neutral
		Excellent. However, running during a major holiday is not ideal (see below point 15)	positive
		Experiment was performed during MD shifts, which necessarily required "special conditions". Nonetheless overall operation was smooth.	positive
		Experiments at and around the phosphorous K-edge. For the first time, we exceeded X-ray energies of 2. keV at the AMO station during this run. We collected data at 2.0 keV, 2.1..	positive
		Explored a new machine mode which performed above and beyond expectation due to excellent machine physics support.	positive
		FEL ran beautifully at 11.2 KeV. very little down time. The accelerator team was even able to deliver 2 mJ @ 11.2 KeV, far better than expected.	positive
		For the most part we had good beam. There was one shift where we had short-lived beam-drops which happened about once a minute or so. This was a real problem for us becau..	neutral
		Generally the machine	positive
		High overall beam energy and position stability, as well as high overall percentage of beam availability.	positive
		I can't rate higher due to shot-to-shot noise, but there isn't much that can be done about it.	positive
		I cannot speak hihgly enough about the accelerator physicists during this run, who pushed the linac harder than it had ever gone before to achieve lasing at 11.2 keV, a record hig..	positive
		I put excellent here in spite of only having about 22 hours of useful beamtime. The reason is that the experiment pushed the boundaries of what the LCLS could do. The lowest p..	positive
		In general beam quality was good, however, for our type of experimenta it was important not to have beam dumps even for a short time. This is unfortunately happening now and ..	positive
		In our 5 shifts, we used self-seeding (SS) during Shifts 1-4. During Shift 1 (used for alignment), we noticed a nice enhancement of our signal due to SS. We did not see this enhan..	neutral
		In particular, we like to acknowledge the effort to provide beam (with lower intensity) even when part of the accelerator was not working, as this helped us to solve problems on the..	positive
		In the first four shifts we did not run in any special mode (49fs pulses at 8keV). The machine was running very stable. In the last shift, we had to change engy in 10eV steps in ord..	positive

First Conclusion

The downside with this analysis is that it gets approximately 80% of its classifications correct, which means we could miss out on some important reviews. That means that we still need someone to glance over the reviews before the meetings, but it will happen much more quickly than before thanks to this program

Data investigation

The other part to my summer project was to investigate some data that I was given. All of the variables were categorical, which really limited the amount of analysis that I could do on them; however, I was able to look into the data and make some visualizations of relationships that I found interesting. The dataset that I had access to had over seventy thousand observations, and for each observation we had 30 variables that went with it. The variables are: Country, Citizenship, Department, Eligible For Beamtime, Ethnicity, Facility, Facts Sensitivity, Funding Source, Institution Code, Institutions Code, Instrument, Run, Person Id, Person Type, Proposal Id, Research Area, Sex, Small Business, State, Submission, Cycle, Submission Cycle, Technique, Time Submitted, Year Of Birth, and Zip code. These were all collected from every experiment from every run at LCLS.

Frequent User Countries

I thought it would be interesting to investigate some of the variables that went along with the countries that have used the SLAC LCLS facilities more than one hundred times. The first important plot that I thought was necessary when looking into the frequent User countries was how many experiments have actually been run by each country while they were at SLAC.

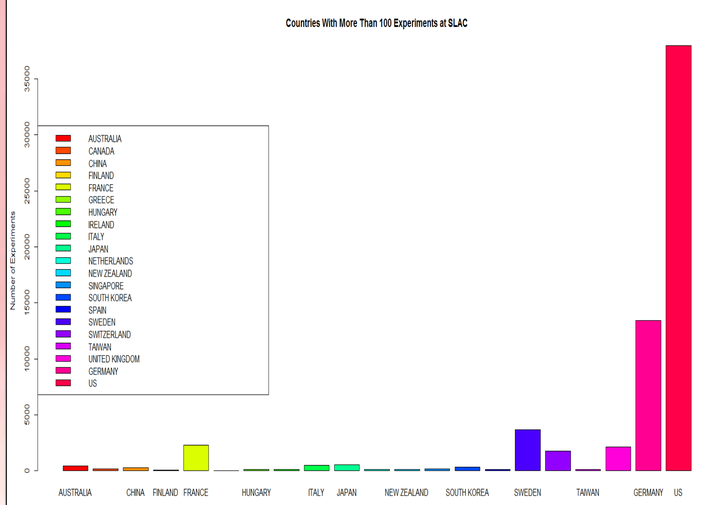


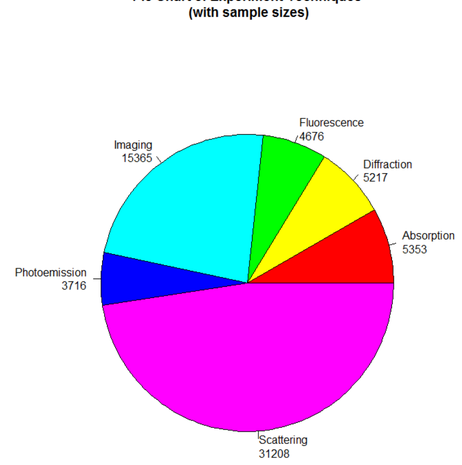
Figure 1: Bar graph of number of experiments associated with the country of the User group

First Conclusion

The downside with this analysis is that it gets approximately 80% of its classifications correct, which means we could miss out on some important reviews. That means that we still need someone to glance over the reviews before the meetings, but it will happen much more quickly than before thanks to this program

Techniques of Experiments

With all the experiments that are run at SLAC, I thought it would be interesting to see which experimental techniques were used most often.



Figures 2: Pie chart of techniques and number of times each technique was used

Research Areas

There are also many different areas of research that are done at SLAC. Over all the runs at SLAC in LCLS the areas of research and their amounts are in the plot below.

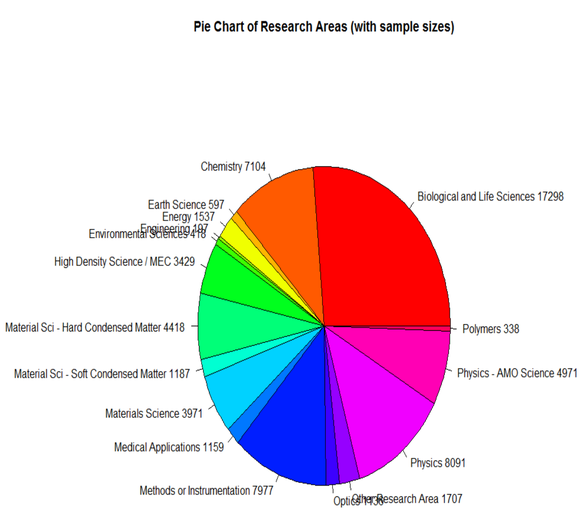


Figure 3: Pie chart of research areas and number of times each area was researched at LCLS

Second Conclusion

With 30 variables and seventy thousand observations there are countless possibilities for investigation, so I decided to post the three that I found most intriguing. With all the other variable associations that I investigated and to the many I didn't, I'm; sure there are many new things that we can find out about SLAC to help decide where they can focus their advertisements to other countries and other institutions in the United States, to help further the scientific discoveries at our great facilities.

Date: 09/07/2016