PREDICTING COMMUNITY ASSESSMENT

VOTING FOR QUESTIONS AND ANSWERS IN STACKOVERFLOW

FINAL PROJECT SPRING SEMESTER 2018

L665 APPLYING MACHINE LEARNING TECHNIQUES IN COMPUTATIONAL LINGUISTICS

SCOTT MCCAULAY

APRIL 17, 2018

OVERVIEW

- StackOverflow is a popular site for sharing questions and answers related to computer programming. It has specialized areas on Python, Java, etc.
- Over 10 million questions, over 4 million registered users, still heavily used despite criticism over perceived bullying/hazing
- Registered users can upvote/downvote both questions and answers
- Questions:
  - Can a machine learning analysis of the text content of a post predict community response? Secondarily, can it predict poster's experience level?
  - Is text content a better predictor of response than other known data, such as user's experience level, post history, years on the site?

CONTEXT FOR THIS WORK

- Our lab in CNetS (Center for Complex Networks and Systems Research) is interested in online social network formation
  - Game theoretical approach to online network formation
  - How online activity relates to individual's sense of identity
- These types of sites where communities assign value to other user's posts are potentially a rich source of research data
- Analysis of the text content of online discourse is an important part of this research (sentiment analysis, assessment of poster's intent…)

RESEARCH QUESTIONS:

(…for this project and future work)

- Can text content predict response better than known data about poster?
- Can a combination of user data plus text predict better than either alone?
- Do the scores for content versus user identity vary by situation?
  - Are questions and answers rated differently?
  - Is there evidence of newbie hazing?
  - Do cultural or regional differences in communication style impact ratings?
  - Is deference shown to frequent or well known posters?

DEEPER QUESTIONS

(…beyond current scope)

- Can we do a sentiment analysis of posts, and look for relationships of sentiment to feedback received?
- Can we observe changes in a user's understanding of the landscape over time?
  - As a user's experience with the site increases, are their posts better received?
  - How does their experience level show in the content of their posts? Enhanced technical expertise or conformity to the community's preferred communication style?

PREVIOUS WORK

- “Predict Closed Questions” competition on Kaggle
- Questions on StackOverflow can be “closed” by other users as off topic, redundant, not a question or other reasons
- Kaggle had a competition to predict which questions would be closed
- This was 5 years ago, so the tools in common usage have changed radically

PREVIOUS WORK

- 100+ papers using Cornell movie review paper
- Much previous and contemporary work on sentiment analysis in text has some commonality with this project
- In a sense the work proposed here goes further, predicting how a post will be received by a diverse audience of peers, rather than matching a single assessment of whether the post itself is positive or negative

PROJECT OVERVIEW: DATA

- StackOverflow has made all user-contributed data available online, in anonymized, zipped XML files
- The most recent data from the main StackOverflow site is > 40GB
- Smaller subsites are available separately, for example the Unix and Linux Stack Exchange is only 350MB (over 300K posts)
- Data is available for user (location, age, cumulative statistics), and for each post (text, date, votes)
- User data is self reported and unvalidated, so of limited value. ☹

PROJECT OVERVIEW: APPROACH

- Use a workable subset of data
  - Picked a smaller subset of StackOverflow (using Linux/Unix exchange)
  - Selected only very high and low rated posts, for a polar classification (> 8K posts)
- Will use a traditional predictive method such as regression analysis to predict community response based on all available poster data
- Will use a CNN to predict community response based solely on vectorized text content of question and answer posts
- Could additionally try to train a CNN to predict other information, such as a poster's length of time in the community

PROJECT STATUS

- StackOverflow data retrieved from Google Cloud BigQuery site
- Selected Linux/Unix Exchange data to use for the analysis (small enough to be processed reasonably, hopefully general enough to be representative)
- Using a combination of keras and sklearn tools for analysis
- Incremental approach, start with simple prediction of polar responses and work toward more complex analysis
- Started trying to predict two categories, up vote or down vote, with the vectorized text of the answer as input
  - Only getting 80% accuracy so far from keras CNN, many options to try to improve on those results
  - Working both on improving the quality of the data and the architecture of the neural network

EXPLORATORY LOOK AT POLARIZED DATA

- Some example phrases from heavily downvoted answers:
  - “Format the drive.”
  - “Wrong.”
  - “Just symlink that directory.”
  - “In other words it is a hack…”
  - “You are mistaken.”

EXPLORATORY LOOK AT DATA BY LOCATION

usage by location, from annual developer survey

EXPLORATORY LOOK AT DATA BY LOCATION

- Locations with best received posts
  - United States
  - European Union
  - Germany
  - Athens Greece
  - Indiana
  - France
  - Virginia
  - London England
  - Aztec NM
  - Hanoi Vietnam
  - London United Kingdom
  - Ontario Canada
  - Sweden
  - Italy
  - Izmir Turkey
  - Mountain View CA
  - Netherlands
  - Fort Lauderdale FL
  - Melbourne Victoria Australia
  - Europe

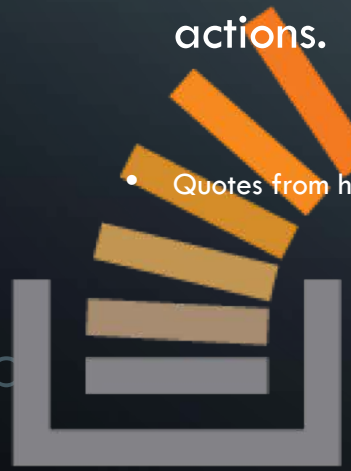
- Locations with worst received posts
  - Germany
  - Bangalore Karnataka India
  - India
  - United States
  - Tehran Iran
  - France
  - Berlin Germany
  - Switzerland
  - Valparaíso Chile
  - USA
  - United Kingdom
  - Mumbai India
  - Kiev Ukraine
  - California
  - Sweden
  - Netherlands
  - Paris France
  - Czech Republic
  - Mumbai
  - Moscow Russia

NEXT STEPS ON MACHINE LEARNING ANALYSIS

- Filter data
  - Select posts with a minimum number of characters
- Better pre-processing of tokens
- Work with a larger set of data
  - May do better if words occur more frequently
- Try different techniques for the NN architecture
  - Worth doing, but problems are more likely with the data
- Hybrid approach – add user features to post content
  - Unfortunately, some of the user features are missing or inconsistent

FEEDBACK AND DISCUSSION

- How to push past the trolls and get the help you need on Stack Overflow
  - When you're new to coding, Stack Overflow can be a scary place. It's an amazing resource for newbies. But it's also a place where bullies troll for new victims.
  - … there's also a vocal minority of people who will respond to your questions with snark or responses like "Read the Freaking Manual (RTFM)". They may flag your question as a duplicate without taking time to read it, or take any number of other passive-aggressive actions.
  - Quotes from https://medium.freecodecamp.org/how-to-push-past-the-trolls-and-get-the-help-you-need-on-stack-overflow-52fd42ebe7c4



stackoverflow