ARE YOU SURE THAT'S REAL?

An explorative journey through Natural Language Processing and Craigslist

Heather Laree Dykstra

Fall 2014

ABSTRACT

Using the skills learned in my Natural Language Processing and Artificial Intelligence Courses as well as the resources on the internet and the CU Department of Computer Science, I plan to analyze Craigslist posts to determine if they are real or fake. I will use techniques discussed in papers on Data Mining, sentiment analysis and spam filtering to spur my understanding of complex phrases and ad analysis. Many web services have moved to using CAPTCHA's and other tools to determine if something is human or not, and if implemented correctly, we can remove the need for this service.

Introduction

Statement of Problem

The internet has become a huge part of everyone's life, and determining if the content on each page is real or fake has become a serious problem. Craigslist is a site where people can go and post all kinds of Ad's, anything from wanting a relationship to "I have free puppies" or needing a specific position filled in a job. The whole thing is semi-monitored by computers and the site owners, but it relies a lot on general feedback from the community on if a post is real or not. Users can go and flag posts and get them removed, or after a successful transaction, report that everything worked out fine. I want to see if I can make an automated system more robust. By looking into specific user data, pulling down how many times a post has been flagged, and the language each user is using to post the ad.

Purpose of Research

In this project, I am setting out to use sentiment analysis and machine language learning to determine if a new post on Craigslist is real.

Significance of Research

This research will allow site owners to better implement bot protection and tell if their site is being spammed by someone or something. This is will also allow users of the site to browse and reply with the knowledge that the post is real, cutting down on time spent trying to reach out to a system that may never contact you back or a system that might try to steal information from you.

BACKGROUND

On some scale, this problem has some solutions. There are many different companies who have implemented their own algorithms to determine if something is fraud, but as we grow and can better detect these posts, the bots grow too and can better avoid being detected.

In her talk about Sift Science, Katherine Loh, discussed fighting just this cause with machine language learning. At the end of her talk she specifically pointed out that “Noise is Everywhere” and gave a few key examples as to what is Noise and what needs to be improved. They were:

- Wrong labels
- Duplicate labels
- Bad integrations
- Incomplete integrations
- Missing fields
- Bugs
- System downtime

It is with this knowledge that we can begin to fix problems. I feel the simplest approach to fixing some of these is to assess wrong and duplicate labels.

A few years ago, a team of people working with Microsoft Research published a paper on detecting spam web pages. They were able to compile a lot of information and analyze a lot of

data on a webpage to determine if it was real or not. The difference between this research and my own is that I will be pulling much less data, and therefore will have to make much larger assumptions about language and scale then this paper does.

They also say “Effectively detecting web spam is essentially an “arms race” between search engines and site operators.” This alone says that there will always need to be more research done in this field and that any

DESCRIPTION OF PROPOSED RESEARCH

I plan on using a few different tools for this research. To get my data I plan on using a python data scraper. I will then notate by hand if a post is real, and what features of each post lead me to believe this is so. Following this I will need to get more people to verify and notate the same files and others to get as much data for each post in our test set as possible.

Once I have all of my initial data, I will need to place it into either an ontology or a database. I will need to then determine which features mean more and use a heuristic algorithm to give weight to those features.

Following the initial set up, I will then have to determine what algorithm to use to ultimately go through all the information in a post and decide if it is real or not.

The end result of this will be that given a post or blurb of text from craigslist, we should have an output that will tell us if a bot or person generated the post as well as some reasoning or a percentage of results.

CHALLENGES

For this project, I see lots of challenges arising. The first of which is that I will need many people to go through each post I want to analyze, by hand to determine if they feel the post is real or fake. The biggest issue with this is we will not have any real way of knowing if the post was in fact posted by a bot. The only way we could determine this would be to email each post and this can be seen as spam.

Another challenge for this, is given the hardware on my laptop and my external drives, I will not be able to compute much for large sets of data. I will need to invest in some sort of cloud service or request access to the CSEL servers. Because this is a low scale project for an undergraduate course, I am not sure I will have the proper resources available to me.

SOURCES

1. http://brianabelson.com/open-news/2013/12/17/scrape-the-gibson.html, Scrape the Gibson: Python skills for data scrapers, Brian Asbelson, 2013
2. http://kavita-ganesan.com/opinion%20mining%20tutorial , Opinion Mining Tutorial, Kavita Ganesan and Hyun Duk Kim, 2008
3. http://en.wikipedia.org/wiki/Sentiment_analysis , Sentiment Analysis, Wikipedia, 10/12/14
4. http://text-processing.com/demo/sentiment/ , Demo Tool for positive or negative sentiment
5. http://genderanalyzer.com/ , Gender Analyzer
6. http://web.eecs.umich.edu/~kulesza/pubs/mteval_thesis.pdf , A Learning Approach to Improving Sentence-Level Machine Translation Evaluation, Alex Kulesza, 2004
7. http://research.microsoft.com/pubs/65140/www2006.pdf , Detecting Spam Web Pages through Content Analysis, Alexandros Ntoulas and Marc Najork and Mark Manasse and Dennis Fetterly, 2006
8. http://www.seaglass.com/postfix/spam-detection.html , Kyle Dent, 10/12/14
9. http://systers.org/wiki/communities/doku.php?id=wiki:ghc:ghc14:real_world_data_science_at_scale_panel , Real World Data Science at Scale Panel, Surabhi Gupta, Alexandra Brasch, Andrea Burbank, Ayse Naz Erkan, Cristina Scheau, 10/9/14 (Grace Hopper Celebration Panel)

10. http://systers.org/wiki/communities/doku.php?id=wiki:ghc:ghc14:data_science_in_practical_applications_presentations Data Science in Practical Applications, Fraud Detection with Machine Learning: A Case Study from Sift Science, Katherine Loh, 10/9/14 (Grace Hopper Celebration Speaker)
11. http://myresearchdiaries.blogspot.com/2014/10/the-power-of-context-in-real-world-data.html The Power of Context in Real-World Data Science Applications #GHC14, Shivani Rao, 10/12/14
12. Lecture Notes