Monday, May 22, 2017 Max Rohde

Yelp Readability Analysis Project

We have Yelp review data.

Main website: https://www.yelp.com/dataset_challenge
Example code: https://github.com/Yelp/dataset-examples

Let’s take the text of all of the reviews and run them through a function that returns a readability metric, which tends to be either numerical, or a academic grade level (i.e., 9th grade).

```
readability_metric_array = []

for text in list_of_reviews_text:
      metric_array.append(readability(text))
```

Now attached to all of our reviews, we have the readability score associated with it

```
(review, readabilityScore)
```

These python packages below could work to provide a readability score

https://pypi.python.org/pypi/readability
https://pypi.python.org/pypi/textstat

What does readability mean?
- longer words?
- higher average syllables?
- higher vocabulary level?
- longer sentences?
- longer length of text?

It varies by the type of metric used. So we need to look into a good readability metric that matches our goals of what readability should be based on.

Examples:
https://en.wikipedia.org/wiki/Automated_readability_index
http://en.wikipedia.org/wiki/SMOG
https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests
https://en.wikipedia.org/wiki/Coleman%E2%80%93Liau_index
https://en.wikipedia.org/wiki/Gunning_fog_index
https://en.wikipedia.org/wiki/Dale%E2%80%93Chall_readability_formula

I suggest that we select a promising few methods and test them on examples from the data.

For example, if we have n methods.

1. Rank reviews by method [n] score
2. View the highest and lowest readability reviews (about 10?)
3. Is there a clear difference in the readability? Will this work for our project?
4. Repeat for the other methods.

What can we do with this data? (some ideas)

- Plot where in a city more/less readable reviews are found (on a map)
- Are more readable reviews associated with higher or lower star ratings (histogram?)?
- What kinds of businesses are associated with more/less readable reviews?
- What does readability tell us about the users associated with the reviews?
    - check-in times?
    - number of reviews?
    - average rating?
    - number of friends?
- Comparisons of readability between cities?

Some considerations

We need to be provide representative samples of high / low readability review texts, so that the audience has reference.

Why?

- Builds trust that our scoring algorithms are working properly
- User find it interesting to read some actual review, not just aggregated metrics

Would be interesting to do this within the visualization automatically as seen in this visualization: https://www.tweetping.net/#/

Work in progress, will be expanded