GDELT: Predicting The Tone of Media Reports

Presentation Outline

- What is the GDELT data set?
- Predicting Average Tone
- Models and Results
- Applications
- Future Work and Improvement

GDELT: Global Data on Events, Location and Tone

What is GDELT?

- Collects information on events from media reports around the world.
- Utilizes the CAMEO code system.
    - Stores events as one actor performing a particular verb on another one.
- It’s massive! GDELT 1.0 is the “small” one and is over 30GB.

Predicting Average Tone

Our Goal

- GDELT implements an algorithm which calculates tone from media reports discussing the event being recorded and then averages them. This is AvgTone.
- AvgTone provides information about the “intensity” or “severity” or “scale” of an event.

Features

- Actor1CountryCode (Actor2)
- Actor1Type1Code (Actor2)
    - Type or role of Actor. “Police Forces” or “Media”, etc.
- EventRootCode
    - Cameo code root code. 1->20 ordered increasing degrees of aggressiveness.
    - 1 = Make Public Statement, 20 = Engage in Unconventional Mass Violence

Models and Results

Data Characteristics

- Sparse.
- A lot of it.
- High dimensional.
- Majority categorical data.

Preparing for Models

- Convert Actor2CountryCode into a boolean determining whether Actor1CountryCode is the same.
- Build a model for each separate Actor1CountryCode separately.
- Or alternatively we take the computational cost of building on everything.

And the winner is...

- Linear Regression with Ridge (L2) regularization.
  - R^2 on test set is .272 when applied to all data.
  - (Unweighted) Mean R^2 on all country by country models is .184.
  - (Weighted) Mean R^2 on all country by country models is .207.
- Gradient Boosting performs about as well, but takes significantly longer to train and isn’t as interpretable.

Applications

What can we do with this?

- Guiding humans who need to know what to pay attention to.
- Interpreting the model could give insight into change media perceptions on actions between countries.

Future Work and Improvements

Future Work and Limitations

- Expand the amount of data from GDELT that is used to build the model.
- Building a more robust model that benefits from increased computing power, probably Gradient Boosting.
- More sophisticated feature engineering and dimensionality reduction.

References

- GDELT Website - https://www.gdeltproject.org/
- GDELT Data Format Documentation - http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf
- CAMEO Code Manual - https://www.gdeltproject.org/data/documentation/CAMEO.Manual.1.1b3.pdf
- Full Report on Github - https://github.com/eistre91/thinkful-projects/blob/master/GDELT%20Supervised%20Learning%20Capstone.ipynb

Erik Istre

e.istre91@gmail.com

https://github.com/eistre91

https://www.linkedin.com/in/erikistre/

Image:
https://www.pexels.com/photo/green-water-fountain-225769/