Datathon 2019: Tips and Suggestions for Participants

(a) Missing Data

- Note that some of the topic IDs listed in training.csv are not actually contained in interest_topics.csv
- It is up to you to decide what to do with this missing information

(b) Imbalanced data

- Choose appropriate evaluation metrics! (F1 score, balanced accuracy, precision, recall, etc.)
- Undersampling, oversampling, etc.
- Imbalanced-Learn package in Python
- Use class weights inside your model, to weigh the smaller class more heavily

(c) Sparse data

- Consider dimensionality reduction methods
- Also, consider combining topics (for example, combine all topics within the 'Arts and Entertainment' category into just one topic)

(d) Dealing with 'big data'

- In R: consider using readr::read_csv() instead of read.csv()
- Python: DictVectorizer may come in handy
- Use vectorization (rather than for loops) when possible!

(e) Other tips

- Start working on your presentations early! Explaining your results is often just as important as producing them, and it may take longer than you expect
- Using visualizations to explain your work can be very helpful