Outline:

1. Problem formulation
2. Data acquisition and cleaning: we have the code to crawl arXiv.org (slightly less than 1M distinct papers) and to generate fake papers using snarXiv.
3. Literature review
4. Feature selection
5. ML algorithm to use: SVM, NN, k-NN
6. Preliminary results: just show some results using SVM
7. More work: different feature set and algorithm (bias analysis), explore variance across different time in arXiv paper and different grammar file in snarXiv.

Tasks:

1. Problem formulation: need some writing

Discuss about snarXiv, SCIGen, and SCIGen detection.

2. Data acquisition: done. Need to record the date (year-month) of arXiv file, and find more grammar file (not important for now)
3. Feature selection: gain some insight from literature review
4. Method: try Connor's method first
5. Result: discuss both false positive and false negative, but we want to minimize false negative first

References:

SCIGen detection: https://hal.archives-ouvertes.fr/hal-00641906v2/document
arXiv v.s. snarXiv: http://snarxiv.org/vs-arxiv/
Conclusion from arXiv v.s. snarXiv:
http://davidsd.org/2010/09/the-arxiv-according-to-arxiv-vs-snarxiv/

Report location