**CSC 180 Intelligent Systems)**

**List of Datasets for Final Project (updated Fall 2022)**

Below is a list of \*potential\* datasets for CS 180 final project. You are also encouraged to find and use a dataset not on this list. Your data set should be reasonably large (> =500KB).

An **overview** about where to find data sets:

* <https://www.dataquest.io/blog/free-datasets-for-projects/>

[Stanford Large Network Dataset Collection](https://dev.twitter.com/docs/streaming-apis/streams/public)

* <http://snap.stanford.edu/data/>

Outlier Detection (attack and security) Datasets

* <http://odds.cs.stonybrook.edu/>

UCI Machine Learning Archive

* <http://archive.ics.uci.edu/ml/index.html>

COVID-19:

* <https://www.kaggle.com/c/stanford-covid-vaccine>
* <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>
* <https://crisisnlp.qcri.org/covid19>
* <https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset>

Tweets:

* <https://data.gesis.org/tweetskb/>
* <https://data.gesis.org/tweetscov19/>

Fake News:

* <https://github.com/KaiDMML/FakeNewsNet>
* <http://www.multimediaeval.org/mediaeval2016/verifyingmultimediause/>

Yelp

* <http://www.yelp.com/dataset_challenge/>

Government Data:

* <https://www.data.gov/>

California CHHS Open Data:

* <https://data.chhs.ca.gov/>

[Sacramento County Open Data](https://data.saccounty.net/)

* https://data.saccounty.net/

NYC Open Data

* <https://opendata.cityofnewyork.us/data/>

Various Science Research Data

* <https://data.mendeley.com/>

Various datasets from ASU:

* <http://socialcomputing.asu.edu/pages/datasets?att=numAttributes&order=ASC>

Foursquare

* <https://sites.google.com/site/yangdingqi/home/foursquare-dataset>
* <https://archive.org/details/201309_foursquare_dataset_umn>

KD Nuggets

* <http://www.kdnuggets.com/datasets/>
* Dozens of datasets

Yahoo

* <http://webscope.sandbox.yahoo.com/>
* The Yahoo! Webscope™ Program is a reference library of interesting and scientifically useful datasets for non-commercial use by academics and other scientists.

Other

* Recommender systems: MovieLens, Million Song Database, ...
* MovieLens: <http://www.grouplens.org/node/12>
* MERL motion sensor data: <https://sites.google.com/a/drwren.com/wmd/home>

StackOverflow data

* [http://www.ics.uci.edu/~duboisc/stackoverflow/](http://www.ics.uci.edu/%7Eduboisc/stackoverflow/)
* online question-and-answer site for programmers
* Predicting Answerers
* Predicting Time until Answer
* Expert Finding Algorithms

Modis

* <http://modis.gsfc.nasa.gov/data/>
* Earth image data

NSF Award Search

* List of all National Science Foundation Awards
* <http://www.nsf.gov/awardsearch/download.jsp>

Medical Subject Headings

* <http://www.nlm.nih.gov/mesh/>
* Tree of 1000s of Medical Subject Headings
* Used by MEDLINE/PubMed

Wikipedia

* <http://en.wikipedia.org/wiki/Wikipedia:Database_download>
* Many many text mining projects possible on this data

Clinical Trials

* <http://www.clinicaltrials.gov/>
* Combine with NIH and/or PubMed

ACL Corpus

* <http://clair.eecs.umich.edu/aan/index.php>
* Corpus of Association of Computational Linguistics articles
* Lots of metadata extracted

Lists of Datasets from others’ classes:

* <http://www.cs.cmu.edu/~epxing/Class/10701/project.html>
* <http://www.cs.washington.edu/education/courses/cse599c1/13wi/project.html>
* [http://www.cs.cmu.edu/~awm/10701/project/data.html](http://www.cs.cmu.edu/%7Eawm/10701/project/data.html)
* <http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html>
* <http://www.stanford.edu/class/cs341/data.html>
* <http://www.stanford.edu/class/cs341/01-qa_session.pdf>