Tian

<https://www.kaggle.com/competitions?hostSegmentIdFilter=5>

Covid detection based on lung-image

<https://www.kaggle.com/thomasnibb/image-classification-covid19-xray>

<https://www.kaggle.com/c/landmark-recognition-2021/overview>

<https://www.kaggle.com/c/deepfake-detection-challenge/data>

<https://www.kaggle.com/c/nlp-getting-started/overview>

<https://www.kaggle.com/c/landmark-retrieval-2021/overview>

GAN

<https://www.kaggle.com/c/gan-getting-started/data>

Semeval-2021

<https://semeval.github.io/SemEval2021/tasks>

Semeval-2022

<https://semeval.github.io/SemEval2022/tasks.html>

From Canvas:

Datasets:

* [data.gov](https://www.data.gov/)
* [(Links to an external site.)](https://www.data.gov/)
* [UCI Datasets](https://archive.ics.uci.edu/ml/datasets.php)
* [(Links to an external site.)](https://archive.ics.uci.edu/ml/datasets.php)
* [Kaggle](https://www.kaggle.com/datasets)
* [(Links to an external site.)](https://www.kaggle.com/datasets)
* [Awesome Public Datasets (github)](https://github.com/awesomedata/awesome-public-datasets)
* [(Links to an external site.)](https://github.com/awesomedata/awesome-public-datasets)
* [Stanford SNAP Datasets](https://snap.stanford.edu/data/)
* [(Links to an external site.)](https://snap.stanford.edu/data/)
* [Google Dataset Search](https://toolbox.google.com/datasetsearch)
* [(Links to an external site.)](https://toolbox.google.com/datasetsearch)

Nice Collection of APIs:

* <https://github.com/public-apis/public-apis>
* [(Links to an external site.)](https://github.com/public-apis/public-apis)

Web Scraping Tools:

* [Selenium](https://www.seleniumhq.org/)
* [(Links to an external site.)](https://www.seleniumhq.org/)
* [Beautiful Soup](https://www.crummy.com/software/BeautifulSoup/)
* [(Links to an external site.)](https://www.crummy.com/software/BeautifulSoup/)
* [Scrapy](https://scrapy.org/)
* [(Links to an external site.)](https://scrapy.org/)

Data Cleaning:

* [Open Refine](http://openrefine.org/)

Also, text data in stack overflow can be collected using built-in APIs

Under Investing on kaggle

--------------------------------------------- topic candidates

## Lexical Complexity Prediction

​​<https://sites.google.com/view/lcpsharedtask2021/call-for-participation?authuser=0>

<https://competitions.codalab.org/competitions/27420#learn_the_details-overview>

Sarcasm evaluation

<https://sites.google.com/view/semeval2022-isarcasmeval#h.t53li2ejhrh8>

<https://github.com/silviu-oprea/iSarcasm/blob/master/isarcasm_train.csv>

* No full data yet

Patronizing and condescending Language Detection

<https://sites.google.com/view/pcl-detection-semeval2022/>

* Train data available

Bakery store data mining

* Frequent itemset, associate rule learning
* Membership data
* Mostly bread, (no rule?) bread→ drinks, evaluate?

To do for Tian

* Clean up dataset
* Delete NaN date row
* Translate column names
* Replace personal info with token
* Output to a new csv
* Check duplicate code
* Add category
  + Bread,

**About data:**

* Negative amount means returned items
  + Check how many returned items, then decide either to remove the transaction or do some special treatment

**Project to do**:

Rules searching

* Frequent itemset mining
  + Top popular items
* Association rules
  + Bread type, bread and drinks/icecream?
* Buying pattern with time/location
  + Morning, afternoon, night
  + Different stores
* Customer clustering
  + Based on their preferences, group by bread type, e.g. cake, drink, muffin
  + Group by money spent
  + Personal recommendation (harder)

Business Insights

* Membership vs. no membership portion
  + Change with time?
* Payment method analysis
  + Wepay vs. alipay vs. cash vs. credits/points
* Visualization tools
  + Popular items sales trend, e.g. mooncake

**Questions:**

* To do recommendations, grading criterias?
* Methods, baselines?
* Creativity? Recommendation system new methods?
  + Focus on practical analysis or apply new methods?
* What if no good results?

**LDA Hyperparameter Optimization via Hyperband**

* Find optimal hyperparameters for LDA under restricted resources (time, iterations)
  + Hyperband
    - <https://arxiv.org/abs/1603.06560>
  + Baseline: grid search
  + <https://www.jmlr.org/papers/volume18/15-595/15-595.pdf>
* Quantify LDA results
  + Coherence within a topic, semantic features
    - <https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0>
    - Semantic features
  + Distance between topics, perplexity
    - Jaccard distance
  + Baseline: human judge

------------------------

Potential problems:

HW2 new dataset for social good

HW2 data all about politics

To do:

* New comprehensive dataset with multiple topics modeling
* Literature review on LDA tuning/optimization
* LDA results quantification metrics
* Random search / grid search with LDA
* Hyperband code with LDA

References

<https://ourcodingclub.github.io/tutorials/topic-modelling-python/>