05/03/20

**Modeling Workflow**

1. Obtain the raw data into csv
2. Pass the raw data (i.e., meta data) title and abstract into paperClassifier to identify the class and subclasses.
   1. Generate a classified csv files
3. The classified csv files will be used for extracting relationships that can be represented numerically:
   1. Topic modeling
      1. See the trend of the topic change allow paper publication date
      2. What topic is more interesting to public research now than before and vice versa
      3. How does the topic distribution change by the origin of data (i.e., country)
      4. What are the topics? What are the keywords in the topics? Can we interpret them? What are the corpus keywords? What are the key papers for each topic?
      5. How the topics change in subclass and class?
      6. Correlate topics to incidence data and motility data (from another Kaggle competition)
   2. Relationship extraction
      1. Entity classification
      2. Entity relationship
         1. Quantify the **relationship** between two entity, e.g., “covid-19”, and “age”
            1. Use topic modeling to define the relationship

E.g., looking into topic and find X correlated to gender

* + - * 1. Human-defined

Define what exactly the relationship is we are looking at for each class? Classes: risk factor, diagnosis, treatment, outcome,

Risk factor: is X the risk factor of Covid-19 (statistically significantly)?

1. Visualize the information in temporal-space
   1. By publication date; ideally by the data source date
   2. Filters:
      1. E.g., countries, journal (high and low impact journal), class, subclasses
   3. Include the incidence and mortality rate projections
2. Potential expansion to link to other data sources:
   1. Stock
   2. Twitters, reddit
   3. Media newspapers

**Data**

From the original data sources

* Title
* Author name, author origin (.e.g, university, which country)
* Abstract
* Shad ID, paper id
* Publication date
* Main text
  + Data date (i.e., when the data is collected)

From the incidence data

* Another Kaggle competition