Exercise

- Replicate Kelly et. al. methodology on USPTO cancer patent data. Download here.
- Goal: Use patent titles and filing years to build a measure of patent quality.
- Exploratory text analysis: examine the terms that are used in the most patent titles.
- Study the distribution of pairwise similarity between patents.
- Analyze how cancer patent quality varies through time.

Exercise Step #1 Data Preparation and Exploration

- Data Cleaning: Consider patents from 1963 to 2000 to keep computational burden low (approx. 43K samples).
- Text preprocessing: lowercase, remove punctuation, numbers, stopwords.
- Create Term-Document Matrix with sklearn's CountVectorizer.
- Replace multiple counts by 1 and sum the number of patents mentioning a term (replication of Appendix Table A1).
- View the most frequently mentioned terms.

Exercise Step #2 Measuring patent similarity and quality

- Compute the standard TF-IDF vectors for each patent via sklearn.
- Create a pairwise cosine-similarity matrix of size approx. 43K x 43K.
- Apply the filters used in Kelly et. al.: set similarity coefficients lower than 0.05 to 0.
- Plot the CDF of the similarity coefficients.
- Compute measures of backward and forward similarity.
- Compute the measure of patent quality.
- Plot the evolution of this measure through time.
- Identify breakthrough patents (above 90th percentile of patent quality) and plot evolution through time.

Exercise: Going Further

- Add more preprocessing steps: lemmatizing, stemming.
- Use different time-windows for:
  - Determining the CDF of pairwise cosine similarity
  - Computing patent quality
- Use all of the cancer patent sample, i.e. after 2000.
- Modify scikit-learn's TfIdf.Transformer() method to implement the backward-IDF calculation.