Author: Chunru Zheng (cz8yb@virginia.edu (mailto:cz8yb@virginia.edu))

Class: DS 5001 Spring 2023

1. Introduction.

This project aims to analyze the textual similarities between patents and their citations filed by Chinese firms between 2000 and 2007. The motivation behind this research is to merge the patent data with firm information and potentially uncover insights into the knowledge diffusion process within the Chinese corporate ecosystem. This project provides a foundation for understanding the knowledge diffusion process among Chinese firms and the relationships between patents and their citations. With further development and analysis, it has the potential to contribute valuable insights to the field of Economics and Intellectual Property research.

It is important to note that the current project focuses on creating a data structure that allows for the merging of different data sources in a 'firmID-patentID-citationID-citedID' format (as shown below). The project does not yet aggregate the distances at the firm level, but this is an area of further research that will be pursued during the course of the Econ PhD program. Additionally, the current project only calculates similarities with the citing-citations, and the cited-citations data is still being scraped. The plan is to incorporate this information into the analysis during the summer.

IndustryID	FirmID	PatentID	CitationID	CitedID
01	SZ10000639	CN101029148	CN100390228C	TBD
01	SZ10000639	CN101029148	CN1145295A	TBD
01	SZ10000639	CN101029148	CN1250621C	TBD

The project consists of three main parts of coding, stored in the 'doflies' folder:

> **(1) Data Cleaning and Preparation**: The data cleaning process is documented in two Jupyter Notebook files:

- **0.Clean_Match_Chunru_20230422.ipynb:** This file clean the ASIF data and match the patent_ID with firm_ID, and citation_ID.
- **0.Clean_Scrape_20230422.ipynb:** This file clean the patent abstracts I got from different sources and got a file prepared for texting analysis: 'pooling_patent_data.csv', with 262458 rows × 3 columns.

> **(2) Calculating Similarity Metrics:**

- **1.Similarity_SKL_Chunru_20230429.ipynb:** This part involves using various techniques, such as TF-IDF, PCA, Topic Modeling, and Word Embedding, to calculate the similarity measurements between patents and their citations. The result is a DataFrame containing the distance between each patent and its citations: 'distance_df_combine.csv'.

> *(3) Visualization and Analysis: *

- **2.Analysis_Chunru_20230425.ipynb:** In this stage, the calculated distances are visualized using network graphs to showcase the relationships between patents and their citations.

2. Source Data

1) Provenance: Data Sources and URLs

The data used in this analysis originates from two primary sources:

- **PATSTAT Global:** https://www.epo.org/searching-for-patents/business/patstat.html (https://www.epo.org/searching-for-patents/business/patstat.html) A dataset containing bibliographical data for over 100 million patent documents from leading industrialized and developing countries. This dataset provides a list of patents filed by Chinese firms along with their English abstracts, allowing for comprehensive comparisons with global citations.
- **Google Patents:** https://patents.google.com/ (https://patents.google.com/) This platform was used to scrape the citations for each patent, obtaining the abstracts of both the cited patents and the patents that cite them.
- **ASIF(Annual Survey of Industrial Firms):** https://www.census.gov/programs-surveys/asm.html (https://www.census.gov/programs-surveys/asm.html) I got the access to this datasets with information of all firm, including ownership, financial performance and patent innavation.

2) Location: Link to Source Files

Due to data confidentiality and size constraints, only a sample of 200 observations from the source data is provided. However, the complete coding process and results for the entire sample are demonstrated. **All data files are stored together in a GitHub repository:** https://github.com/Chunru1995/Patent (https://github.com/Chunru1995/Patent)**.**

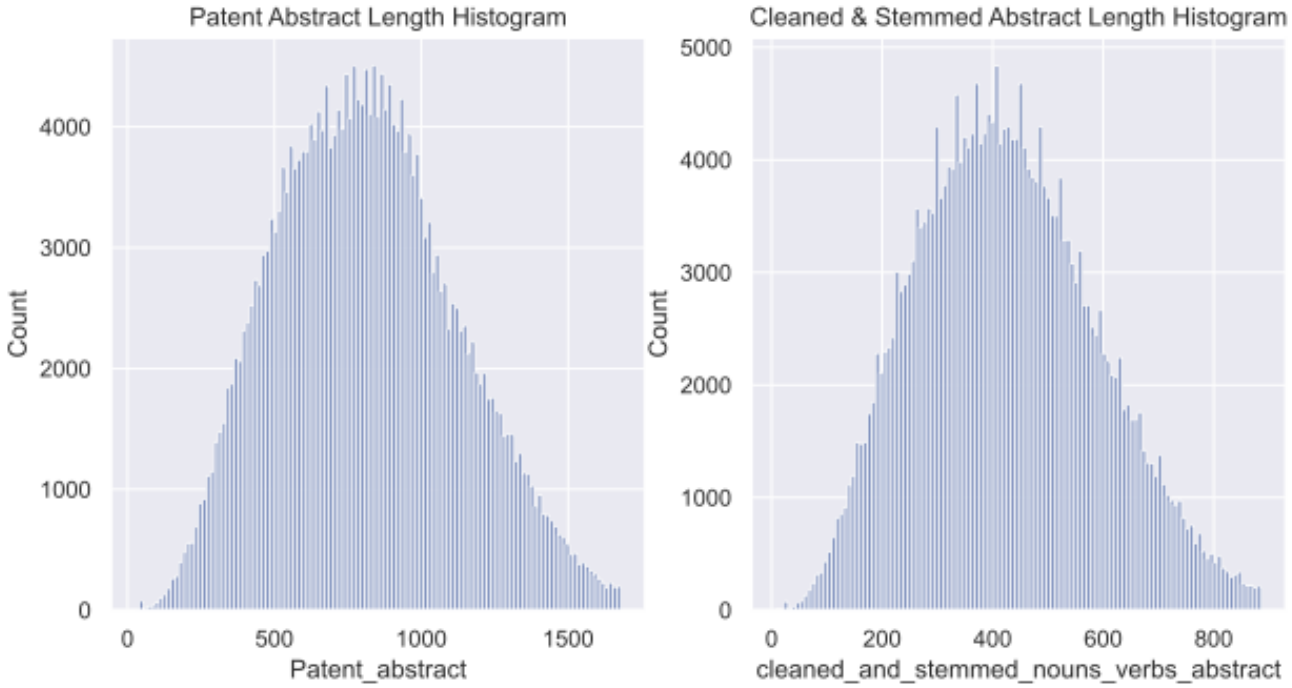
3) Format: File Formats and Internal Structure

The data was scraped and stored in CSV files, which were then imported as dataframes. As the abstracts are not overly lengthy, it was not necessary to use LIB or store them as text files.

4) Description: General Subject Matter, Observations, and Document Length

The dataset consists of 54,522 patents filed by Chinese firms, with English abstracts. The citations of these patents include 263,672 observations. All patents and their citations are combined in a file named 'pooling_patent_data.csv'.

The word length distribution of each patent is analyzed before and after data processing. Data processing involves removing punctuations, stopwords, digits, and non-English words, as well as applying the Porter stemmer to nouns and verbs. This reduces the dimensions of the TF-IDF table and accommodates the large dataset.



3. Data Model.

Describe the analytical tables you generated in the process of tokenization, annotation, and analysis of your corpus. You provide a list of tables with field names and their definition, along with URLs to each associated CSV file.

Here is the list of my tables: all with index of 'Patent_ID'

> **tfidf_table**: tfidf_table_filter.csv (with 262458 rows × 2605 columns)

> **pca_table**: pca_table.csv (with 262458 rows × 30 columns)

lda_table: lda_table.csv (with 262458 rows × 30 columns)

wbed_table: wbed_table.csv (with 262458 rows × 100 columns)

senti_emot_table: sentiment_and_emotions_df.csv(with 200 rows × 10 columns) Note that sentiment_and_emotions analysis is not that suitable for professional text like patents, so for the aim of this course project, I only do that with a sample of 200.

With all these tables, I calculate the distance of each patent to their citations. To do this, I need a dataframe showing the citing relationships of patents, which is called **'PAIRS.csv'**. Then I calculate the distance of each pair, using TFIDF, PCA, Topic model and word embedding, respectively; and change the measurement of ditstances, including euclidean, cosine, jensenshannon and jaccard (for TFIDF only).

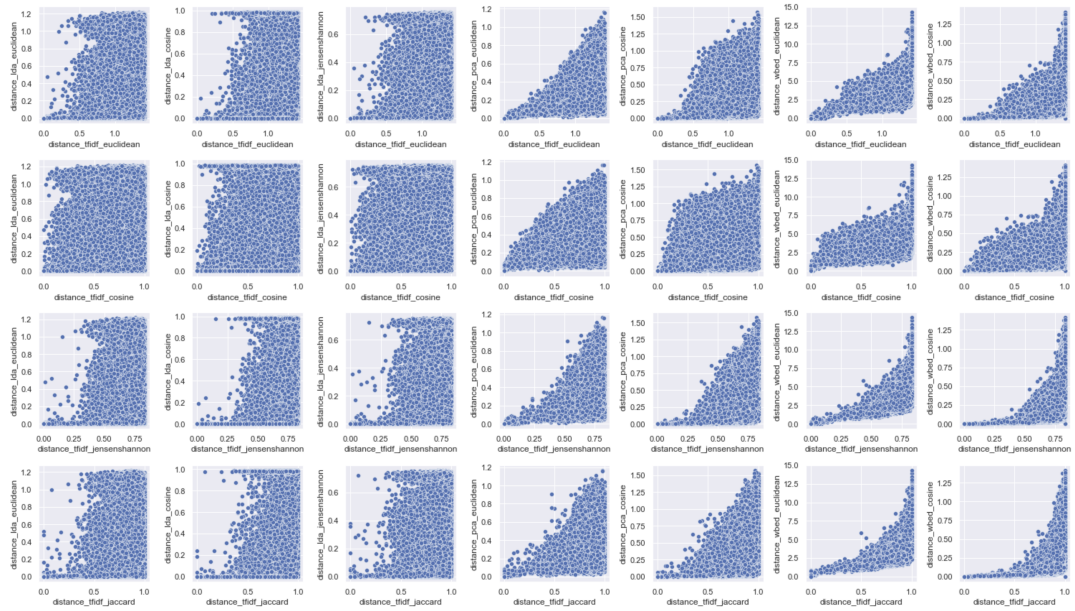
distance_table: distance_df_combine.csv (with 255032 rows × 13 columns).

4. Exploration.

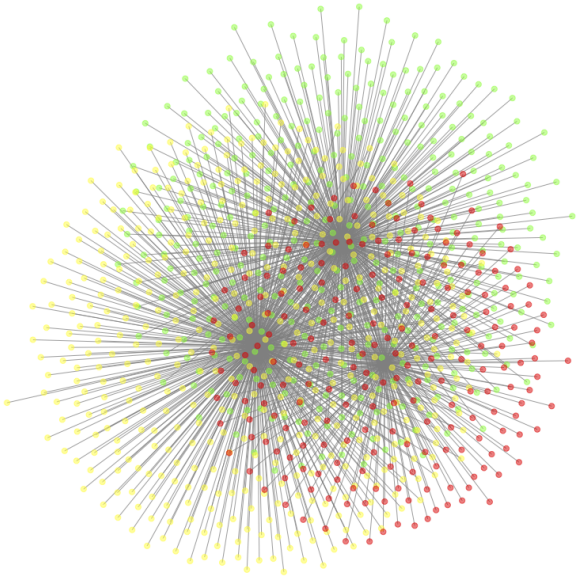
Describe each of your explorations, such as PCA and topic models. For each, include the relevant parameters and hyperparemeters used to generate each model and visualization.

Here I use the distance_table, indexed with a pair of **['APPLN_ID_SIPO' (ID of the patent they file), 'PubNum_google (ID of the patent the cite)']**, I groupby each patent ID and caculate the most nearest citation, and the average distance of each patent to their citations. **All the process are included in the '2.Analysis_Chunru_20230425.ipynb'**

1) corr matrix of different measurement of distance:



2) Show an example of 3 patents with their nearest citations:



For this part, I also create a dropdown bar in the notebook to show an interaction: when inputting different patent_ID, it show the network of its citation, accordingly.

3) sentiment distribution of samples:

```
sentiment_and_emotions_df['Sentiment_Polarity'].sample(20).sort_values().plot.barh();
```