ARTS1422 Final Project

THEME ONE

Introduction

Patent data is one of the key indicators for measuring innovation capabilities and technological advancement. Behind each patent lies the emergence of new technologies and products, which not only drive industry development but also have a profound impact on economic growth and social progress. In the context of globalization and rapid technological advancement, studying patent data can help us gain insights into industry trends, identify technological hotspots, and forecast future directions.

This topic will focus on patent data across the United States. Through this dataset, you will have the opportunity to explore the level of innovation activity in various fields, understand regional innovation differences, and observe trends in patents across industries. By analyzing dimensions such as time, region, and technical classification within the patent data, students can use visualization tools to intuitively reveal the meaning behind the data, thereby enhancing their data analysis and visual presentation skills. It is hoped that, in completing this project, students will not only improve their skills but also develop a strong interest in data science and innovative technology.

Dataset

This dataset records patent information registered in the United States from 1989 to 2020, totaling 893,656 entries. Note that this does not represent the actual number of patents, as some patents have multiple inventors, which leads to duplication in the count.

Data Format

Header	Description
application_id	The unique identifier for a patent application.
doc_id	The unique identifier for a patent document.
filing_date	The date the patent application was filed, in the form of YYYY/MM/DD.
assignee_sequence	The serial number of the patent assignee, indicating the order of the assignee in the patent.
assignee_id	The unique identifier of the assignee.
original_organization	The name of the original assignee, indicating the company or institution to which the patent originally belonged.
assignee_location_id	The unique identifier of the assignee's location, used to identify the geographic location of the assignee.
inventor_sequence	The serial number of the inventor, indicating the order of the inventor in the patent.

Header	Description
inventor_id	The unique identifier of the inventor.
male_flag	A flag indicating the gender of the inventor, 0 for female and 1 for male.
inventor_location_id	The unique identifier of the inventor's location, used to identify the inventor's geographic location.
assignee_city	The name of the assignee's city.
assignee_state	The name or abbreviation of the assignee's state.
latitude	The latitude of the assignee's location, used for geographic positioning.
longitude	The longitude of the assignee's location, used for geographic positioning.
assignee_MSA_GEOID	Geographical identifier of the metropolitan statistical area in which the transferee is located. Check Metropolitan Statistical Areas for more information.
assignee_MSA_name	Name of the metropolitan statistical area in which the transferee is located.
patent_id	The unique identifier of the patent, usually the same as the document ID.
cpc_class	The Cooperative Patent Classification (CPC) main class code for patents. Check CPC Scheme for more information.
cpc_sub_class	The Cooperative Patent Class (CPC) subclass code of the patent.
cpc_class_title	Specifies the title of the CPC subclass.
patent_title	The title of the patent that briefly describes the subject matter of the invention.
patent_abstract	The abstract of the patent that Outlines the main contents and uses of the invention.
beCited	The circumstances in which the patent is cited. Note that the id inside this column is **not included in this dataset**, if you want to use this column, you can simply count the number of citations.

Detailed Declaration for some headers

1. latitude & longitude: The latitude and longitude here are accurate to the city, that is, if two patents are filed in the same city, then their latitude and longitude are the same.
2. cpc_class: Patents in the CPC system are classified according to a three-level structure.

   For example, if a patent's cpc_sub_class is **F01C**, it first belongs to the **F** main class (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING), then to the **F01** class (MACHINES OR ENGINES IN GENERAL; ENGINE PLANTS IN GENERAL; STEAM ENGINES), and finally to the **F01C** subclass (ROTARY-PISTON OR OSCILLATING-PISTON MACHINES OR ENGINES).

   You can check CPC Scheme to learn more about it.

Topics

Topic 1

Technological Trends and Industrial Development require you to have an in-depth understanding of the current state of technological advancements and the strengths of various industries across states. By utilizing existing data and geographic information, you will detail the research status and trends in different fields. In Topic 1, your visual analysis tasks include:

1. Describing the current state of technological development in various fields based on patent categories, and providing corresponding supporting evidence.
2. Identifying the industrial structure of each state based on geographic information, and summarizing key findings.
3. Predicting potential technological fields and determining their development trends.

If you choose Topic 1, You might want to look at the data in filing_date, assignee_city, assignee_state, latitude, longitude, patent_id, cpc_class, cpc_sub_class, cpc_class_title, beCited.

Topic 2

Career Development focuses on institutions and individuals, examining the relationship between research potential and talent mobility within research institutions. In Topic 2, your visualization analysis tasks include:

1. Evaluating the level of investment in research and the research patterns of different institutions, and categorizing them accordingly;
2. Exploring talent mobility dynamics, assessing the quality of talent, and analyzing its impact on research outcomes;
3. Providing career planning advice for job seekers and making corresponding recommendations.

If you choose Topic 2, You might want to look at the data in filing_date, original_organization, inventor_id, male_flag, patent_id, cpc_class, cpc_sub_class, cpc_class_title, beCited.

THEME TWO

Introduction

“Open the Door to Free Trade!”

In an increasingly globalized world, international trade serves as the cornerstone of economic prosperity for nations worldwide. From electronic gadgets to agricultural goods and industrial raw materials, thousands of products traverse the globe daily through intricate logistics networks. However, this movement involves not just the mechanics of transportation; it also encompasses deeper trade relationships and economic interactions among countries.

The international trade regulator you work for has collected detailed data on individual companies' trade activities over recent years—capturing where they trade, with whom, in what products, and at what economic value. This data paints a complex portrait of global trade relations, far beyond mere numbers. As a visual analytics expert at the agency, you have been tasked with analyzing this data. What insights can you extract to help the agency uncover hidden trends, thereby enhancing its ability to regulate.

Dataset

The dataset consists of 1 main dataset of trade networks and 12 supplementary datasets of prediction-complementary networks in the same format, consisting of nodes and directed links (from the shipper to the receiver).

Data Format

Nodes

Attribute	Description
id	Name of the company.
shpcountry	Country the company most often associated with when shipping.
rcvcountry	Country the company most often associated with when receiving.
dataset	Always 'MC2' and not essential in this project.

Links

Attribute	Description
arrivaldate	Date the shipment arrived in YYYY-MM-DD format.
hscode	Harmonized System code for the shipment, representing different types of the shipment.
valueofgoodsusd	Customs-declared value of the shipment, in dollars.
volumeteu	The volume of the shipment in 'Twenty-foot equivalent units', roughly how many 20-foot standard containers would be required.
weightkg	The weight of the shipment in kilograms.
generated_by	Name of the program that generated the link. (Only appeared on the 12 sets of predicted trade networks.)
dataset	Always 'MC2' and not essential in this project.
source	Name of the supplier of this trade.
target	Name of the consignee of this trade.

Detailed Declaration

1. Some data was anonymized leading to some supplier and consignee names/countries being omitted. These are represented by numerical names in the dataset.
2. Due to the agency's limited data collection capacity, some detailed data in "nodes" and "links" is partially missing.

Topics

Topic 1

`Analyze global freight network`

What insights can you derive from analyzing the data to assist the agency in identifying hidden trends? These trends may reveal the potential for designing targeted “little” service fees for specific countries, especially considering that some nations move shipment more frequently than others, which could indicate significant economic linkages or even hint at policy changes.

However, due to data privacy policies, some information is anonymized to protect privacy—such as the countries of origin for companies and the exact name of the shipment. You will need to explore alternative methods to fill these gaps. In Topic 1, the challenge you need to tackle is:

1. Analyze the value and transportation characteristics of various cargo types within the global freight network.
2. Examine the trade profiles of various companies using available data to assess their growth potential.
3. Investigate the positions of countries within trade networks, analyze their trade relations, and provide supporting evidence for your findings.

If you select Topic 1, you may not need to view the data in "Recommendations".

Topic 2

`Identify abnormal companies`

There are some companies engaged in illegal trade, and past experience has shown that when these companies are forcibly shut down, they often reemerge under different names. This severely undermines the security of the trading network, making it essential to analyze the activities of these companies over time to identify anomalies.

Due to their secretive nature, there is a significant lack of data collection by the agency on these illegal trades. Consequently, the agency employs a variety of tools, including artificial intelligence, to analyze the trade network and provides 12 recommendations for recovering trade links. Your task is to evaluate these recommendations to determine which tools are most reliable for patching the network. (The degree of overlap between the predict and the actual network can reflect their relevance). In Topic 2, your task is:

1. Identify temporal patterns in trade records between companies and categorize the types of business relationship patterns you discover.
2. Evaluate the 12 sets of predicted trade networks and determine which sets are the most reliable for enhancing the trade network.
3. After incorporating the reliable predictive links, use visual analytics to identify new patterns and anomalies in the knowledge graph. Highlight any anomalous companies and provide specific evidence to support your findings.

Tips: Pay special attention to new temporal patterns or anomalies that emerge when incorporating predicted complementary networks.