C M U - S V S E

PORTFOLIO

By **Wan Yin Chen**

Highlighting Projects in
Scalable Systems, Legal Tech,
and AI Solutions

Carnegie Mellon University
Silicon Valley

2025

About Me

I am a law graduate from National Taiwan University with a year of experience exploring the intersection of law and technology. Over the past year, I have honed my skills through a variety of projects, research, and self-learning, gaining experience in **machine learning**, **natural language processing**, and **data visualization**.

At Carnegie Mellon University, I aim to broaden my technical expertise and gain a deep understanding of scalable software systems, machine learning, and data-driven decision-making. The SESV program's interdisciplinary focus and emphasis on real-world applications align perfectly with my goal to contribute to innovative solutions across various industries and drive meaningful change through technology.

- **github link:** https://github.com/wyinchen/CS-DS-Portfolio
- **Resume**

- **Languages:** Python, C++, R, SQL, JavaScript
- **Frameworks/Tools:** Tableau, HTML/CSS, TF-IDF
- **Techniques:** Machine Learning, Data Visualization, Statistical Analysis, Web Scraping, Text Mining

Academic & Technical Foundations

Self-Learning: LeetCode Practice

- Solved 92 problems in one month, focusing on:
- Advanced: Dynamic Programming, Backtracking, Divide and Conquer.
- Intermediate: Trees, Binary Trees, Hash Tables.
- Fundamental: Arrays, Strings, Two Pointers, Matrices.
- Profile: LeetCode

Related Coursework

National Taiwan University (NTU)

29 credits with an average GPA of **4.11 / 4.3**

- Programming for Data Science (A)
- Programming and Web Scraping (A)
- Computer Programming in Python (A-)
- Digital Decision Making: Data Visualization and Machine Learning (A+)
- Text Analysis with Python (A)
- Statistical Learning (A+)
- Calculus (General Mathematics I & II) (A+, A+)
- Introduction to Computer Science (A+)
- Seminar on Legal Analytics (A)
- Digital Intelligent Court & Empirical Legal Studies (A)

National Taiwan Normal University (NTNU) TRANSCRIPTS

12 credits, Non-Degree Academic Coursework

- Data Structures (A+)
- Object-Oriented Programming (A+)
- Linear Algebra (B+)
- Statistics (A+)

Experience

Research

- Empirical Research on Guardianship Declaration in Taiwan

Awards

- Second Place, Taiwan’s Legal Tech Hackathon
- Honorable Mention, National Humanities Big Data Competition
- Bachelor’s Thesis Award, National Taiwan University

Technical Projects

Course Projects

- Price Comparison Platform
- Defamation Ruling Analysis for Civil Damage Compensation Prediction
- Drunk Driving Fatality Sentencing Analysis
- Examination Essay Sample Analysis

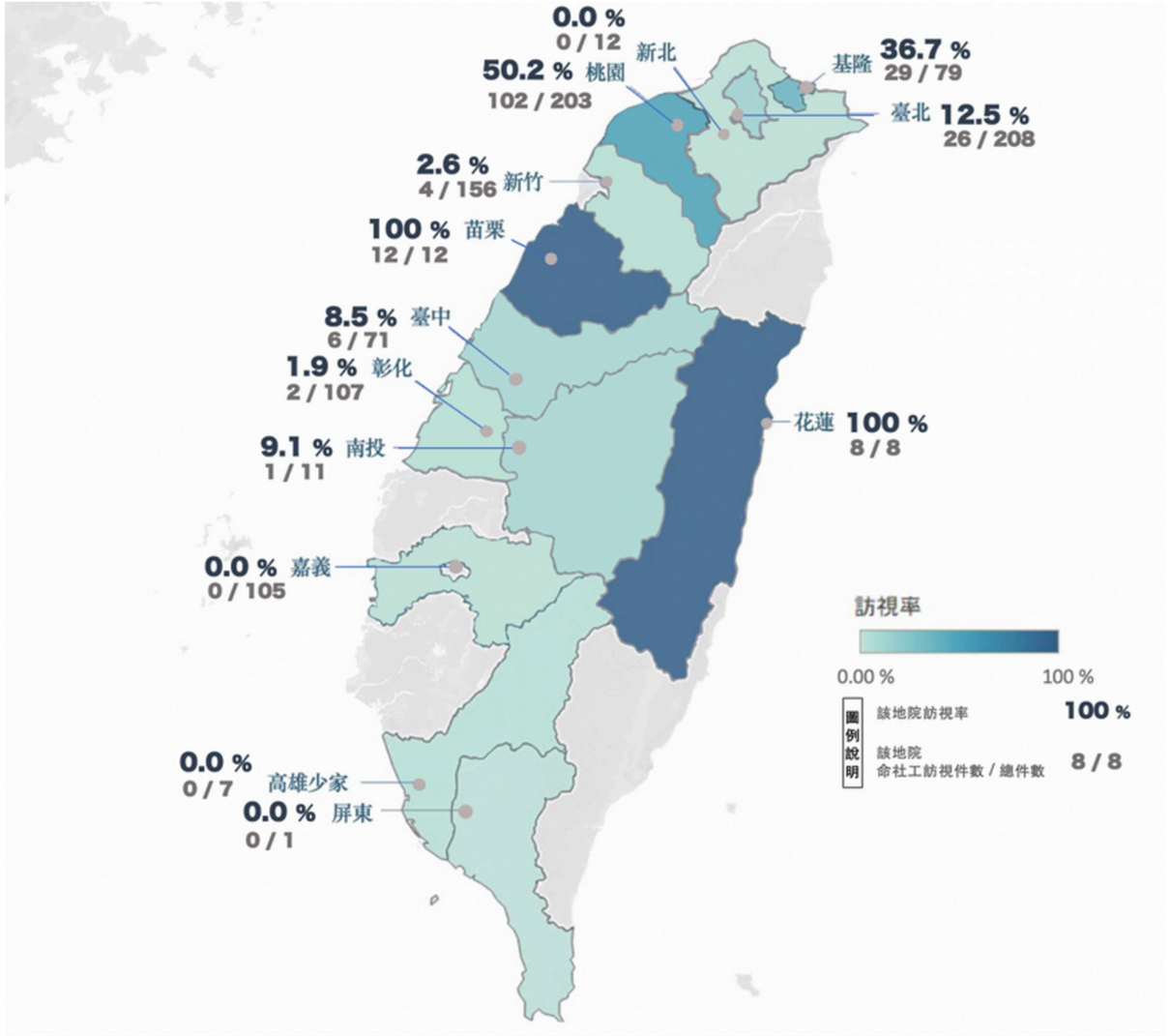
Competitions

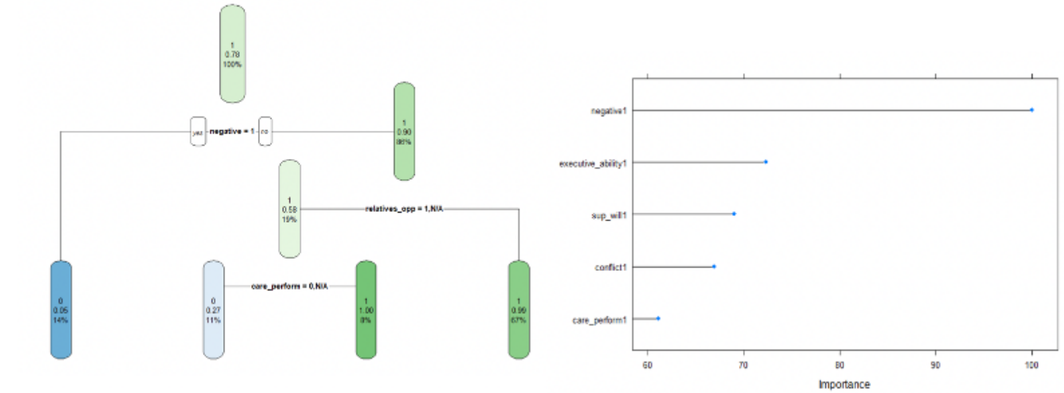
- AI-Assisted Judgment System for Jury Decision Support
  （Legal Tech Hackathon）
- Compensation Prediction System for Marital Rights Violation Cases
  （National Humanities Big Data Competition）

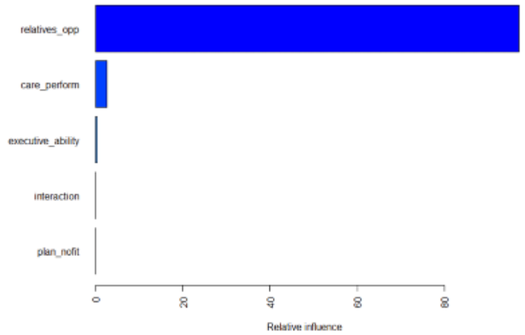
Empirical Research on Guardianship Declaration in Taiwan

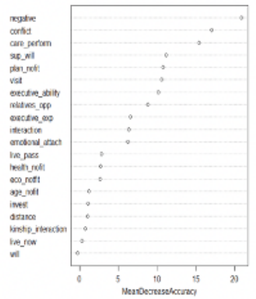
Undergraduate Research Fellowship, National Science and Technology Council (NSTC), Taiwan

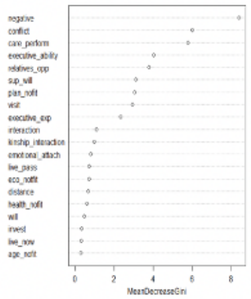
- Advisor: Prof. Sieh-Chuen Huang
- Honor: NTU Undergraduate Thesis Excellence Award
- Publications:
  - Huang, S.-C., & Chen, W.-Y. (2023). An empirical study on the applications for guardianship declarations. **Court Case Times, 129**, 99–110.
  - Huang, S.-C., & Chen, W.-Y. (2023). An empirical study on the factors influencing court-appointed guardianship in Taiwan. **Court Case Times, 136**, 104–117.

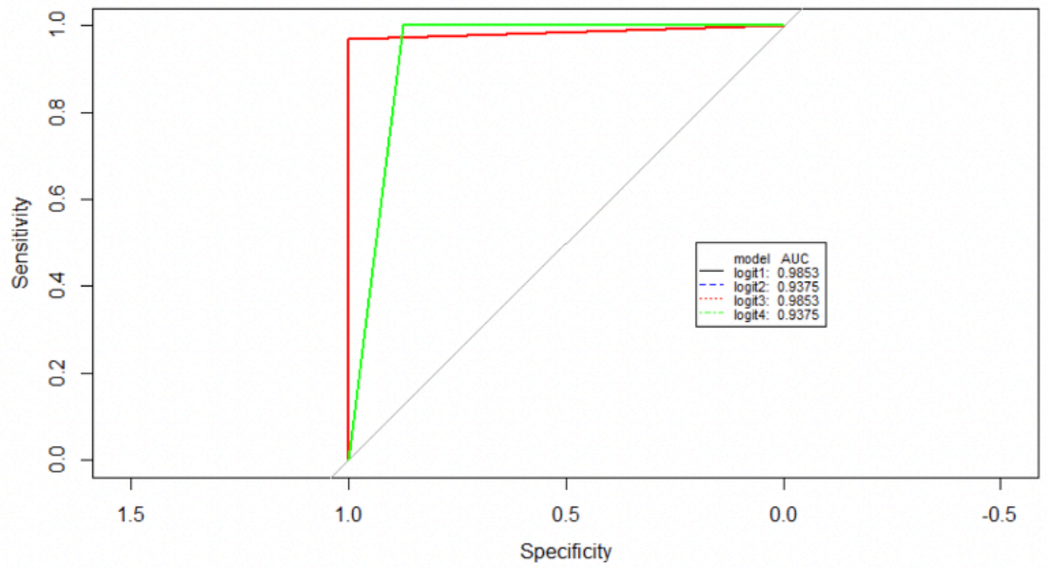












- **Research Focus**

  Analyzed 1,155 guardianship cases in Taiwan (2009–2021), focusing on judicial selection logic in 171 cases with multiple candidates.

- **Techniques**

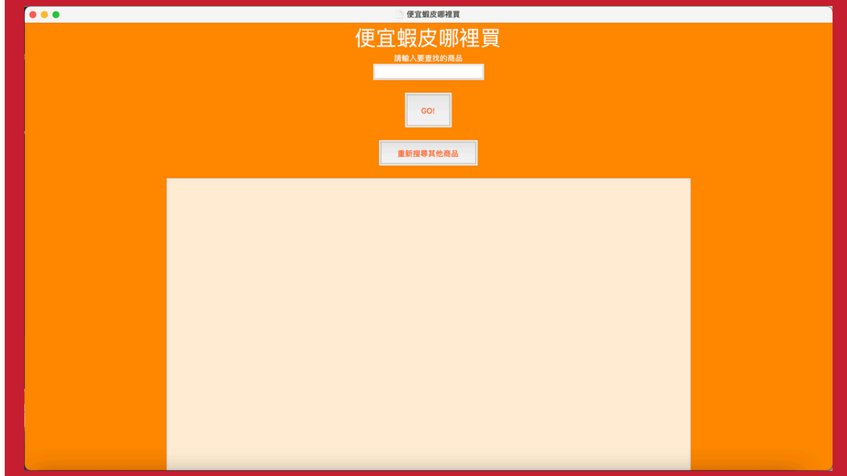
  Applied machine learning models (e.g., decision trees, random forests, GBM), logistic regression, and data visualization techniques to identify key decision factors and streamline analysis

- **Outcomes**

  Achieved 90.9% prediction accuracy with an AUC of 0.9423, providing actionable insights to enhance transparency and efficiency in legal decision-making processes

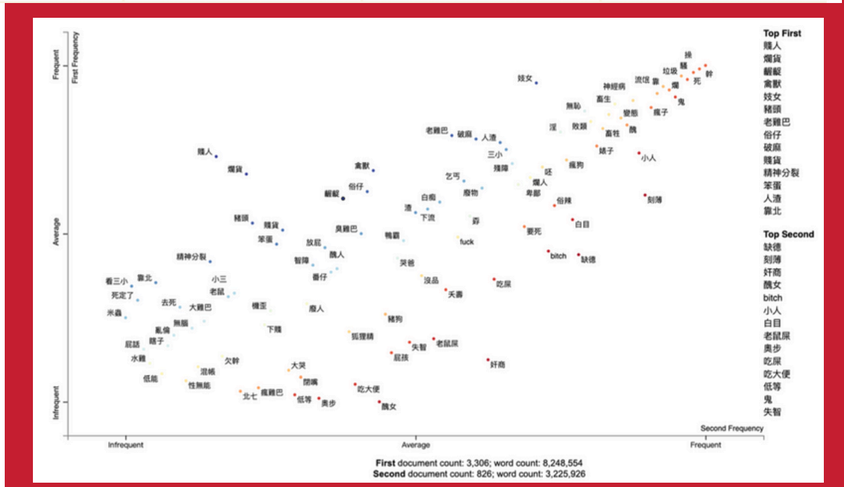
Price Comparison Platform (Shopping Website Scraping)

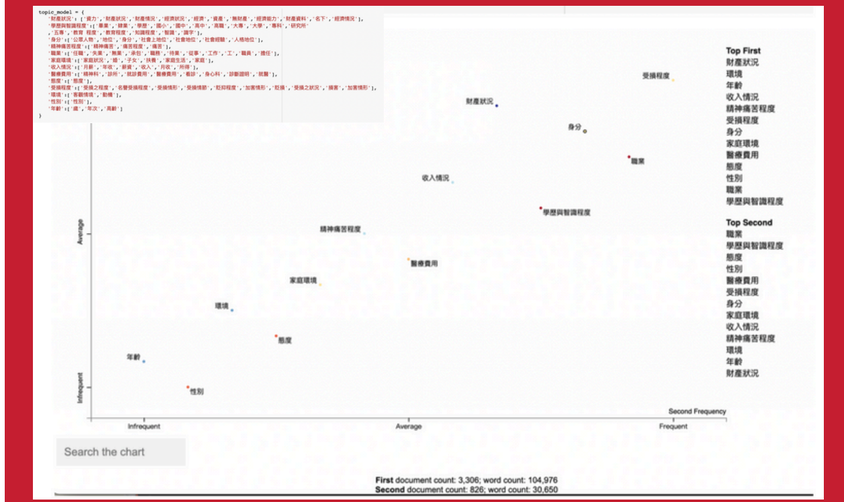
- **Course**: Programming and Web Scraping
- **Github**
  - power point
  - Presentation Video
- **Objective**
  - Simplify the price comparison process for online shopping by creating a tool that automates product price ranking and optimizes coupon application.
- **Key Features**
  - Extracted product data and coupon information from Shopee using Python and web scraping techniques.
  - Designed a GUI interface using Tkinter for user-friendly interaction.
  - Implemented backend algorithms to rank product prices, accounting for shipping costs and discount applications.
- **Impact**
  - Enabled users to make real-time, efficient price comparisons.
  - Reduced manual effort by automating data collection and sorting processes.
- **Technologies Used**
  - Programming: Python, Asyncio, Requests
  - Tools: Tkinter, JSON for data formatting



Defamation Ruling Analysis for Civil Damage Compensation Prediction

- **Course**: Text Analysis with Python
- **Github**
  - power point
  - Term project
  - Dynamic Charts
- **Objective** To analyze defamation rulings in Taiwan's district courts (2008–2018), evaluate compensation patterns
- **Key Features**
  - Collected and analyzed 4,132 court rulings, focusing on compensation amounts and frequently litigated terms.
  - Applied Natural Language Processing (NLP) to extract and rank derogatory terms by frequency and legal relevance.
- **Impact**
  - Identified regional inconsistencies in damage awards, supporting judicial standardization.
  - Revealed biases in the "Insult Pricing Table," advocating data-driven improvements.
  - Improved transparency to inform policymakers and legal professionals.
- **Technologies Used**
  - Python: For web scraping, data preprocessing, and analysis.
  - NLP Techniques: Used TF-IDF to rank defamatory terms.
  - Tableau: To visualize regional trends and disparities in court rulings.
  - Statistical Analysis: Applied to validate compensation patterns.

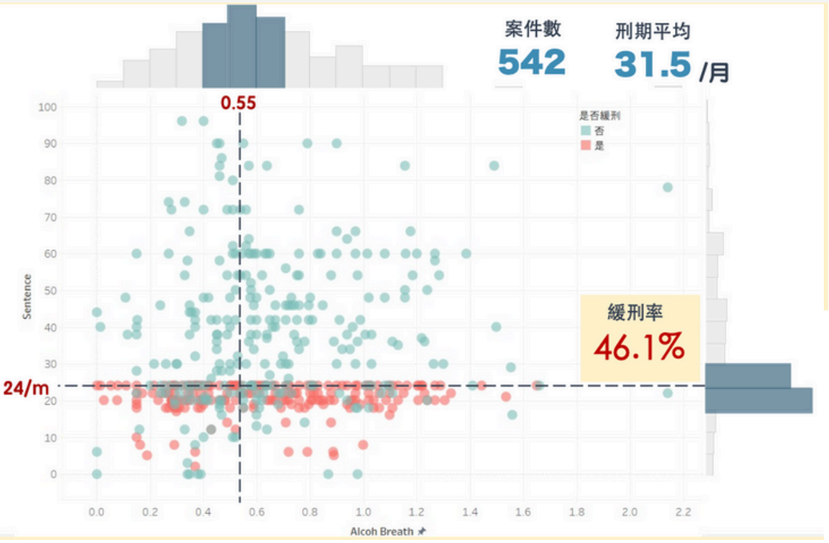




Drunk Driving Fatality Sentencing Analysis

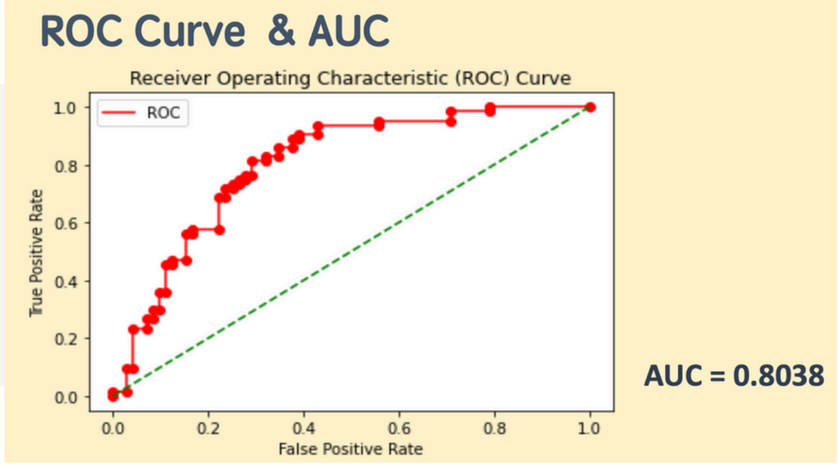
- **Courses**: Seminar on Legal Analytics, Data Visualization and ML, Statistical Learning
- **Github**
  - Digital Humanities and Legal Analytics Educational Webpage: term project
  - Term projects
- **Objective** Analyze the sentencing factors influencing drunk driving fatality cases in Taiwan to identify judicial trends and factors affecting probation rulings, providing insights for legal practitioners and policymakers
- **Key Features**
  - Developed structured datasets from 542 court rulings (2015–2021) using web scraping and data preprocessing.
  - Utilized linear regression to identify factors impacting sentencing length, such as alcohol concentration and post-offense behavior.
  - Constructed logistic regression models to predict probation likelihood, focusing on variables like compensation, criminal history, and judicial region.
- **Impact**
  - Identified inconsistencies in sentencing practices and probation rulings across regions.
  - Offered data-driven recommendations for improving sentencing transparency and equity.
  - Contributed tools and insights to enhance judicial policy evaluation and public communication.
- **Technologies Used**
  - Data Processing & Analysis: Python, R
  - Visualization: Tableau
  - ML & Statistical Modeling: MLP, Decision Trees, and Linear and Logistic Regression Techniques

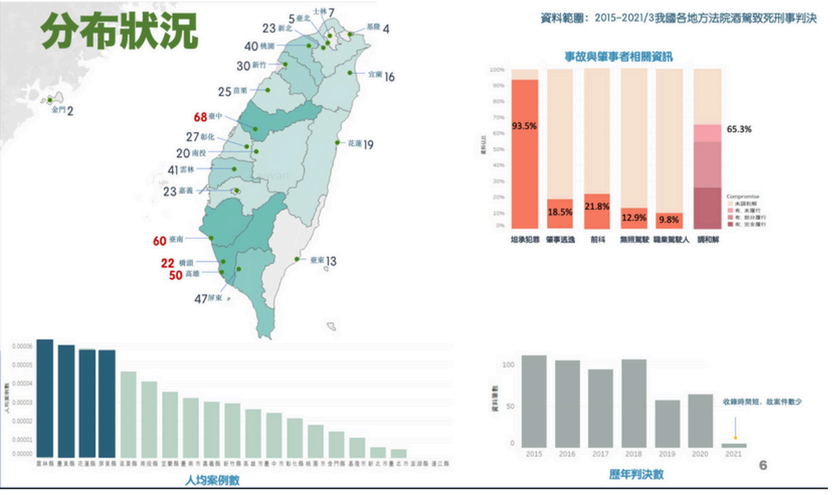
Drunk Driving Fatality Sentencing Analysis

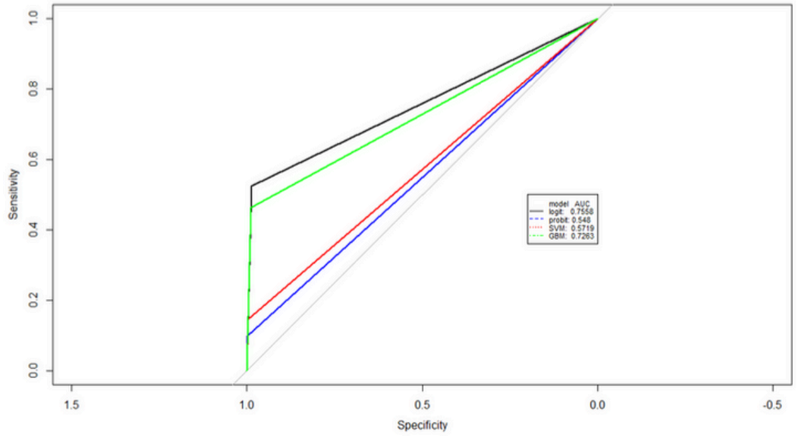


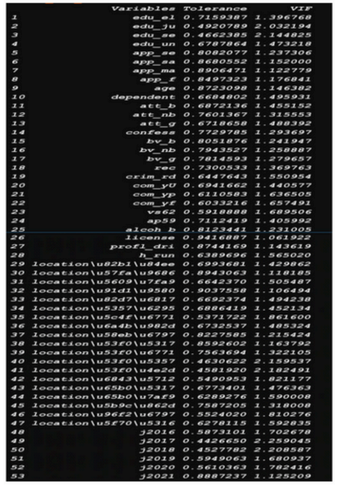
近十年酒駕肇事案件與受傷人數

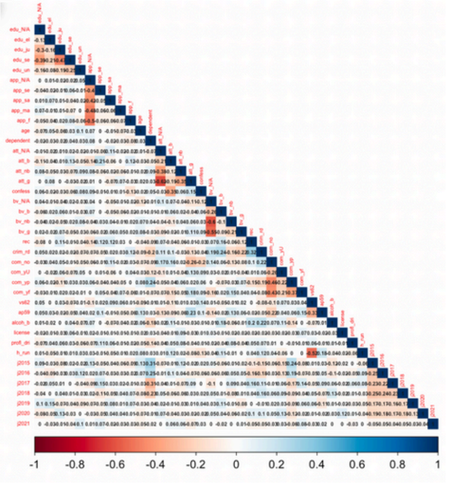
近十年酒駕肇事死亡人數











Examination Essay Sample Analysis

- **Course**: programming for data science
- **Github**
  - term project website
  - Term project
- **Objective** To analyze patterns and qualities in high-achieving essays from Taiwan's university entrance exams, aiming to understand the linguistic and structural elements that contribute to their success
- **Key Features**
  - Transcribed and standardized 208 handwritten essays (2006–2021) into machine-readable text.
  - Analyzed word count, paragraph structure, and sentence patterns to identify key trends.
  - Evaluated quote usage, historical references, and vocabulary complexity through statistical methods.
  - Used CKIPtagger for semantic analysis, comparing essay similarities and measuring uncommon word usage.
- **Impact**
  - Identified key factors of high-quality essays to improve teaching strategies.
  - Highlighted linguistic and structural disparities, promoting fairer evaluation criteria.
  - Laid groundwork for future research on academic writing standards in Taiwan.
- **Technologies Used**
  - Python (CKIPtagger, Jieba) for natural language processing and semantic analysis.
  - R (ggplot2) for data visualization and statistical analysis.
  - Regular Expression for pattern detection in textual data.
  - SimHash for semantic similarity assessment.

Examination Essay Sample Analysis

使用ckiptagger斷詞與分析詞性

佛祖見迦葉而拈花微笑，
兩人之間的會心無法為外人共享。

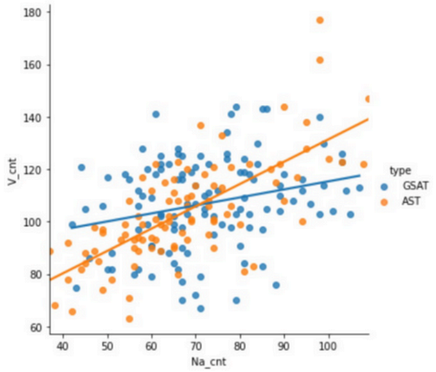
佛祖(Na) 見(VE) 迦葉(Nb) 而(Cbb) 拈花(VA) 微笑(VA) ，(COMMACATEGORY)

兩(Neu) 人(Na) 之間(Ng) 的(DE) 會心(Na) 無法(D) 為(P) 外人(Na) 共享(VJ) 。(PERIODCATEGORY)

中研院平衡語料庫詞類標記集

104學測作文<獨享> 佳作

ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
ADV	D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbc	/*專有名稱*/
N	Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
N	Ncd	Ncda, Ncdb	/*位置詞*/
N	Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
DET	Neu	*Neu*	/*數詞定詞*/



結果

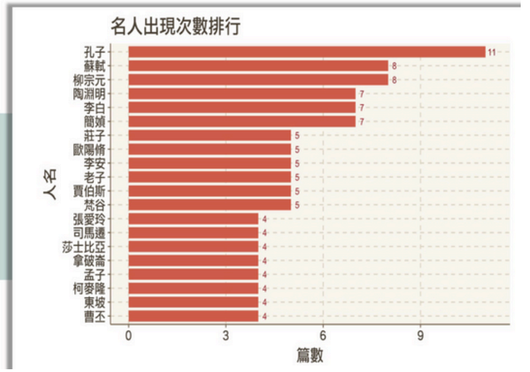
名詞：Na(普通名詞)

動詞：除 V_2(有) 外所有動詞

名人偉人 出現次數排行

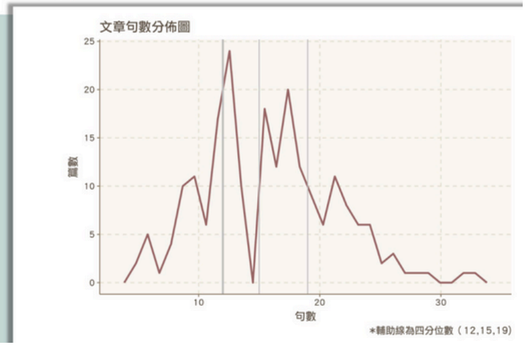
發現

引用之名人類型以中國古代文人為主



佳作文章句數

- 平均：15.75
- 標準差：5.17
- 中位數：15
- 眾數：13



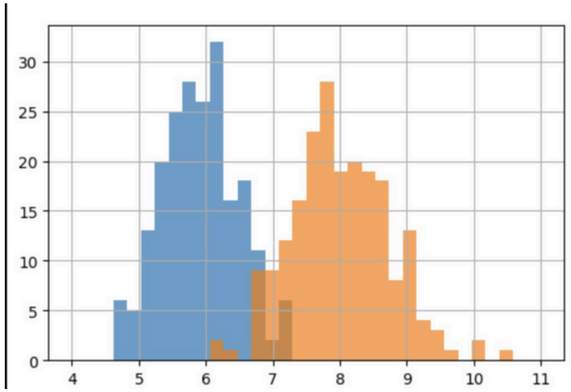
相似度分析

量化比較文本之間的相似度

$$d(\vec{p},\vec{q}) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_n-q_n)^2}$$

$$cos(\theta) = \frac{\vec{p}\cdot\vec{q}}{\|p\|\|q\|}$$

AST_100_1.txt	AST_100_5.txt	AST_100_11.txt	AST_100_14.txt	AST_100_2.txt
1.0000000	0.9649792	0.9228791	0.8584520	0.8566117
AST_100_9.txt	AST_100_8.txt	AST_100_3.txt	AST_100_6.txt	AST_100_13.txt
0.8483397	0.8353900	0.8117013	0.8043561	0.7923995
AST_100_12.txt	AST_100_7.txt	GSAT_99_5.txt	AST_100_4.txt	AST_100_10.txt
0.7863128	0.7645895	0.5575813	0.5367095	0.5320222
GSAT_104_10.txt	GSAT_101_4.txt	GSAT_103_3.txt	GSAT_104_6.txt	AST_101_6.txt
0.5244957	0.4695750	0.4048583	0.3981934	0.3897797

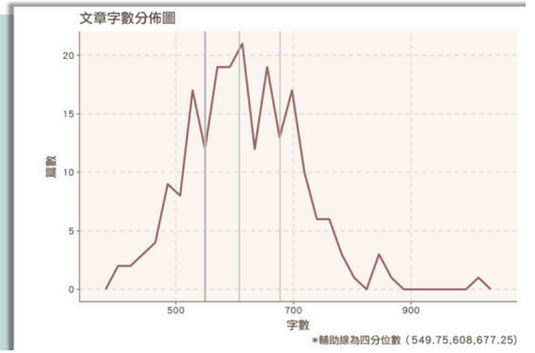


平均詞意量 = 5.911

詞意量非零詞彙平均詞意量 = 8.013

佳作文章長度

- 平均：615.21
- 標準差：94.16
- 中位數：607
- 眾數：650



Competition & Honor

2020 Legal Tech Hackathon

Silver Award

AI-Assisted Judgment System for Jury Decision Support

- **Web** Legal Tech Hackathon 2020 Facebook Page
- **Final Round Replay** Watch Here
- Developed an AI-assisted judgment system to support jury decision-making in mental competency cases, addressing expertise gaps between citizen and professional judges.
- Leveraged machine learning to analyze judicial patterns and designed an intuitive interface tailored for both legal professionals and laypeople.
- Collaborated with interdisciplinary peers from computer science, psychology, and law, utilizing pseudo-labeling techniques to tackle overfitting challenges and improve model accuracy

2020 National Humanities Big Data Competition

Honorable Mention

Compensation Prediction System for Marital Rights Violation Cases

- **Web** Event and Award-Winning Project Exhibition Website
- **Objective** To create an ML-powered system predicting compensation in marital rights cases, improving accessibility for non-expert users.
- **Key Features**
  - Developed ML models, including decision trees and random forests, to predict compensation with high accuracy.
  - Designed an intuitive interface to simplify legal data, providing case-specific compensation estimates.
  - Conducted data preprocessing and regression analysis to identify key judicial decision factors.
- **Impact**
  - Provided accessible legal insights, enabling non-experts to make informed decisions.
  - Identified patterns and disparities in judicial rulings, promoting discussions on standardizing compensation policies.
  - Equipped policymakers and legal professionals with data-driven tools for evaluating case outcomes.
- **Technologies Used**
  - ML: Decision trees, random forests, and regression models
  - Data visualization and UI/UX design for an intuitive prediction interface
  - Data preprocessing and feature engineering for judicial case datasets

Carnegie Mellon University
Silicon Valley

THANK YOU

FOR REVIEWING MY PORTFOLIO

I LOOK FORWARD TO THE OPPORTUNITY TO CONTRIBUTE TO CMU-SV AND FURTHER MY JOURNEY IN SOFTWARE ENGINEERING

Contact Information

- **Wan Yin Chen**
- **Email** wynnewyc@gmail.com
- **GitHub** https://github.com/wyinchen/CS-DS-Portfolio