AI for Justice

Final Presentation
UCLA CAM REU 2024

PI: Professor Deanna Needell, Mentor: Dr. Minxin Zhang
Shreya Balaji, Dakota Lin, Anshuman Singh, Kyle Torres

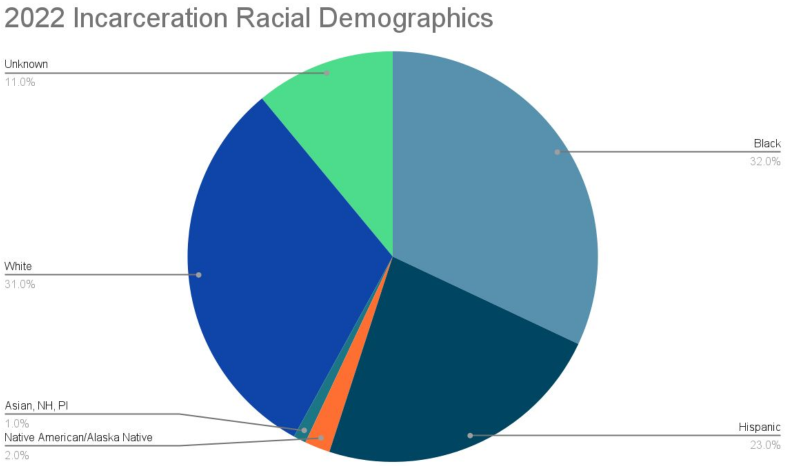
01

Background

Injustice in Our Criminal Justice System

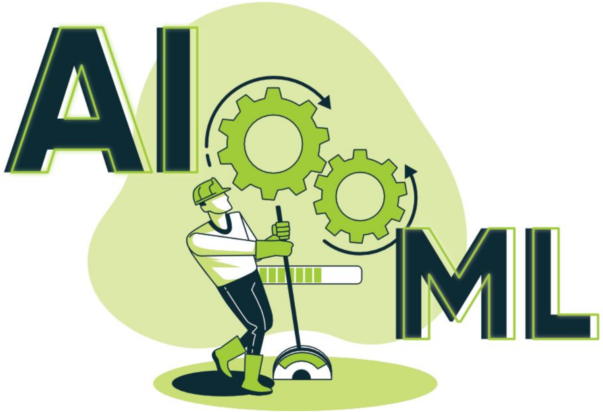
Disproportionate Incarceration Rates

- Dating back to 1999, 49% of prison inmates were African American, despite African Americans comprising only 13% of the overall population
- Estimates suggest that 5-10% of the incarcerated population are innocent
- Study shows that 4.1% of incarcerated individuals under a death sentence could be exonerated



The Purpose in Our Work

- Enhance the use of AI and ML technologies within the criminal justice system
- AI technologies should be fair, reliable, and transparent
- Mitigate bias that is inherent in the system due to historical data
- Secure justice for all and protecting humanity
- Test models against historical decisions to ensure reliability in our work



THE INNOCENCE CENTER

- Nonprofit Organization dedicated to exonerating wrongfully convicted individuals
- Advocates for policy and practice changes to prevent wrongful convictions
- Assists clients with post-release life adjustment
- Raise awareness through partnerships with educational institutions

The National Registry of EXONERATIONS

- Database of wrongfully convicted individuals who have been exonerated
- Raises awareness of systemic issues and advocates for criminal justice reforms
- Contains annual reports with trends and patterns that highlight issues
- Partners with innocence organizations, legal clinics, and academic institutions

02

Our Data

Data Sources & Filtering

Preliminary Goal: 100-200 documents of murder case opinions (50-100 documents of exonerated/non-exonerated cases)

Data Sources:

- Exonerated cases: The National Registry of Exonerations
- Non-exonerated cases: Casetext or Westlaw

Data Filtering:

- Murder cases with exonerations within the last ten years
- Excluded federal Supreme Court cases

casetext
Smarter legal research

THOMSON REUTERS
WESTLAW™

Data Selection Process

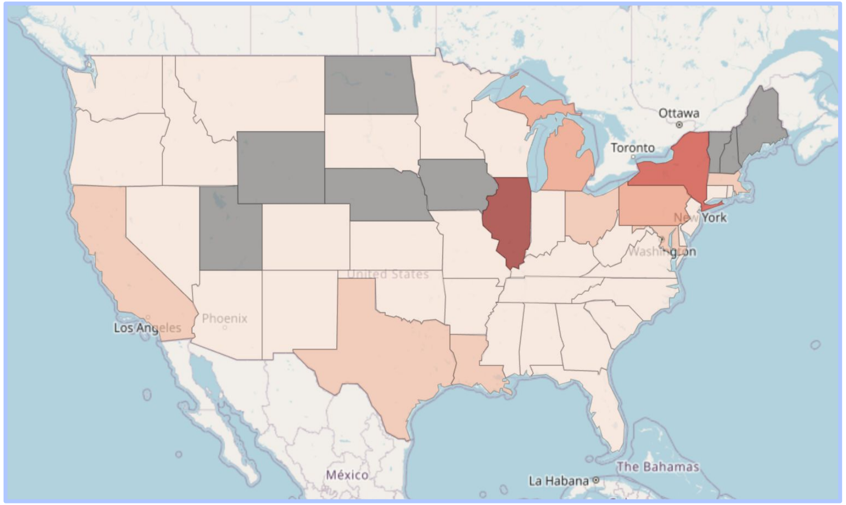
- Randomly selected one case from each state
- Randomly selected additional cases to reach ~100 data points
    - Located corresponding documents on Casetext and Westlaw
- Eliminated cases with unavailable documents
- Repeated the process until reaching a sufficient number of data points in the desired range

Final dataset contains 140 cases total (70 exonerated & 70 non-exonerated)

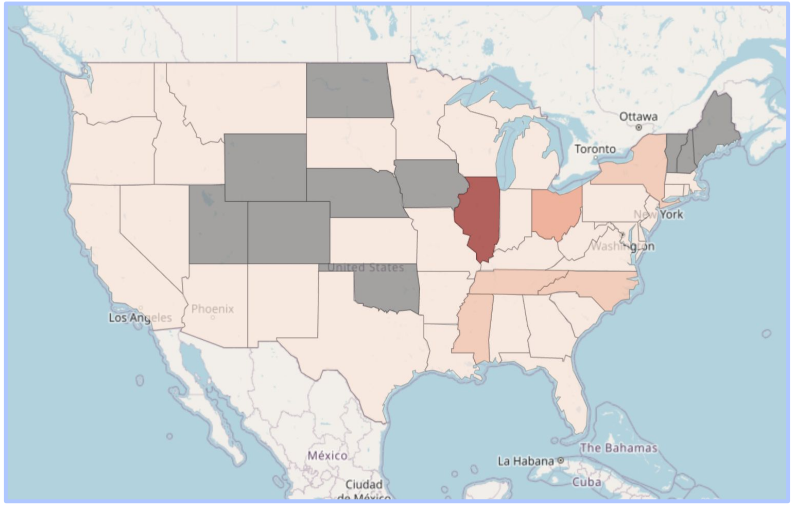
Last Name	First Name	Age	Race	ST	County of Crime	Tags	OM Tags	Crime	Sentence	Convicted	Exonerated	DNA	MWID	FC	P/FA	F/MFE	OM	ILD
Count= 3550																		
Abbitt	Joseph	31	Black	NC	Forsyth	CV, IO, SA		Child Sex Abuse	Life	1995	2009	DNA	MWID					
Abbott	Cinque	19	Black	IL	Cook	CIU, IO, NC, P	OF, WH, NW	Drug Possession or Sale	Probation	2008	2022				P/FA		OM	
Abdal	Warith Habib	43	Black	NY	Erie	IO, SA	OF, WH, NW, WT	Sexual Assault	20 to Life	1983	1999	DNA	MWID			F/MFE	OM	
Abernathy	Christopher	17	White	IL	Cook	CIU, CV, H, IO, JV, SA	OF, WH, NW, INT	Murder	Life without parole	1987	2015	DNA		FC	P/FA		OM	
Abney	Quentin	32	Black	NY	New York	CV		Robbery	20 to Life	2006	2012		MWID					
Abrego	Eruby	20	Hispanic	IL	Cook	CDC, H, IO	OF, WH, NW, WT, INT, PJ	Murder	90 years	2004	2022		MWID	FC	P/FA		OM	
Acero	Longino	35	Hispanic	CA	Santa Clara	NC, P		Sex Offender Registration	2 years and 4 months	1994	2006							ILD
Adams	Anthony	26	Hispanic	CA	Los Angeles	H, P	OF, WH, NW, WT	Manslaughter	12 years	1996	2001				P/FA		OM	
Adams	Cheryl	26	White	MA	Essex	F, NC, P		Theft	Probation	1989	1993				P/FA			
Adams	Darryl	25	Black	TX	Dallas	CIU, IO, NC, P, SA		Sexual Assault	25 years	1992	2017	DNA			P/FA			
Adams	Demetris	22	Black	IL	Cook	CIU, IO, NC, P	OF, WH, NW	Drug Possession or Sale	1 year	2004	2020				P/FA		OM	

Where Is Our Data From?

Geographic Distribution of Exonerations

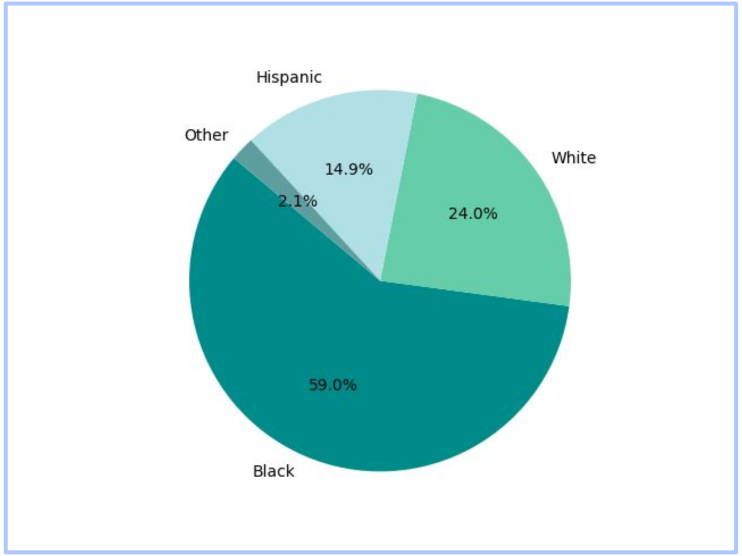


Heat Map of Original Data

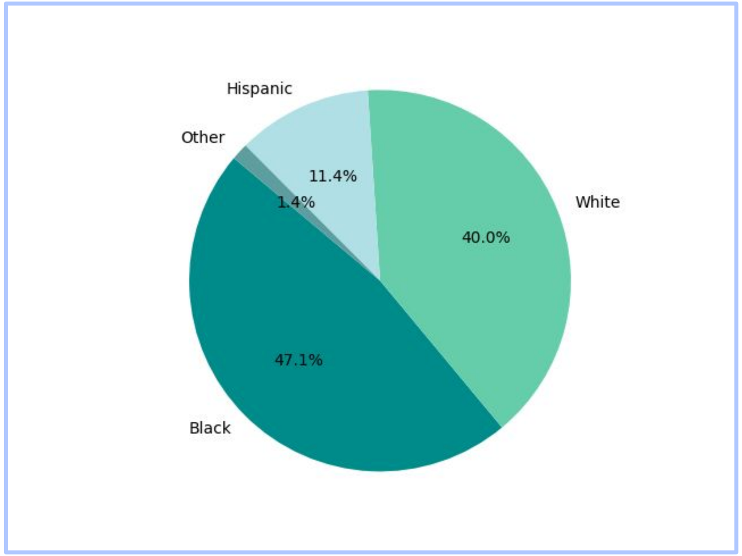


Heat Map of Sample Data

Racial Distribution of Exonerees



Original Data



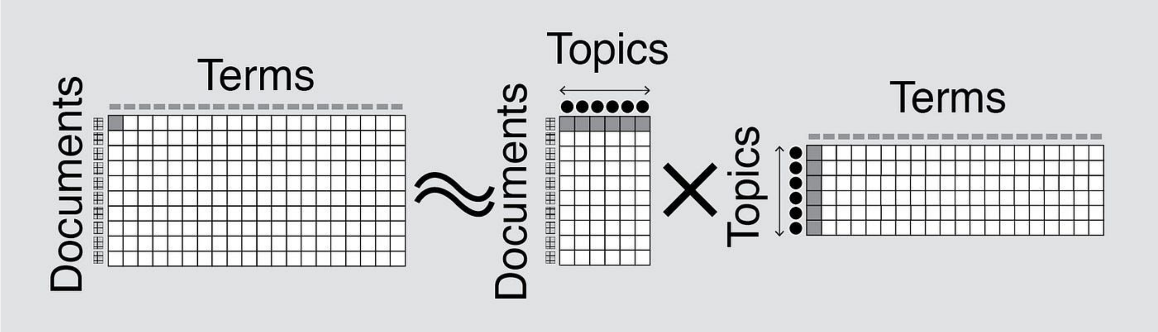
Sample Data

03

Methodology

Nonnegative Matrix Factorization (NMF)

Vanilla NMF framework:



(Data Matrix) ≈ (Feature Matrix) x (Basis Matrix)

Semi NMF

- Semi NMF is a variation of NMF, where the basis matrix ***F*** can have positive and negative values, while the coefficient matrix ***G*** is non-negative
- Used for document embeddings, which are represented as column vectors of the input matrix ***X***
- The flexibility in ***F*** allows for a better representation of our complex mixed-sign data
- The sparse, non-negative ***G*** helps us identify the most significant features in our data
- Our algorithm[1] minimizes the objective function to achieve matrix factorization:

$$J_{K\text{-means}} = \sum_{i=1}^{n} \sum_{k=1}^{K} g_{ik} \|\mathbf{x}_i - \mathbf{f}_k\|^2 = \|X - FG^T\|^2$$

- This factorization transforms ***X*** into a product of ***F*** and $\boldsymbol{G^T}$ for better data interpretation

1. C. H. Q. Ding, T. Li and M. I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 1, pp. 45-55, Jan. 2010, doi: 10.1109/TPAMI.2008.277.

Convex NMF

- Convex NMF is a variation of NMF where the basis vectors ***F*** (represented by ***W***) are combinations of the input data columns, similar to how cluster centroids work
    - This ensures that the basis vectors lie within the column space of the input matrix ***X***
- Used for non-negative and mixed-sign data, and it produces sparse factors which highlight key features in our data
- Our algorithm[1] transforms ***F*** into a product of ***X*** and ***W*** for better data interpretation:
    - $$\mathbf{f}_\ell = w_{1\ell}\mathbf{x}_1 + \cdots + w_{n\ell}\mathbf{x}_n = X\mathbf{w}_\ell, \quad \text{or} \quad F = XW$$

1. C. H. Q. Ding, T. Li and M. I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 1, pp. 45-55, Jan. 2010, doi: 10.1109/TPAMI.2008.277.

Semi-Supervised NMF (SSNMF)

- SSNMF incorporates both labeled and unlabeled data during factorization process, and it helps the model generalize better to new, unseen data.
  - The labeled data helps the model understand the specific features or categories of interest.
  - The unlabeled data ensures the model captures the overall data distribution.
- We want to minimize $\|\boldsymbol{W} \odot (\boldsymbol{X} - \boldsymbol{AS})\|^2 + \lambda \|\boldsymbol{L} \odot (\boldsymbol{Y} - \boldsymbol{BS})\|^2$, where lambda is a weight parameter, Y is the label matrix (document x class), B is the basis matrix for Y

H. Lee, J. Yoo and S. Choi, "Semi-Supervised Nonnegative Matrix Factorization," in *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 4-7, Jan. 2010, doi: 10.1109/LSP.2009.2027163.

Kernel SSNMF: Our Extension

- We project the data to a higher dimensional space (kernelize the data vectors).

$$\mathbf{x}_i \rightarrow \phi(\mathbf{x}_i), \quad \text{for } i = 1, 2, \ldots, n$$

- Our objective function becomes

$$\mathbf{Z} = \mathbf{Z}\mathbf{W}\mathbf{G}^T,$$

where

$$\mathbf{Z} = \begin{bmatrix} \phi(\mathbf{X}) \\ \lambda \mathbf{Y} \end{bmatrix}.$$

- Our method is semi-supervised because we have stacked it with a label matrix and we follow the update rules of Convex NMF, thereby restricting the F matrix to be a convex combination of the data matrix, Z.

Kernel SSNMF: Computational Strategy

- We overcome the need for computing phi(X) by directly computing the kernel matrix below which would be expensive for large number of features.
- Our objective function for minimizing the error becomes

$$\min \left\|\mathbf{Z}-\mathbf{ZWG^T}\right\|^2 = \mathrm{Tr}(\mathbf{D}-2\mathbf{DWG}^\top+\mathbf{GW}^\top\mathbf{DWG}^\top)$$

, where $D = \phi^T(\mathbf{X})\phi(\mathbf{X}) + \lambda^2 Y^T Y$. $\phi^T(\mathbf{X})\phi(\mathbf{X})$ is our kernel matrix, so the objective function did not depend on $\phi(\mathbf{X})$, but it depended on the kernel matrix.

- Also, similar to SSNMF, A (our basis matrix for phi (X)), B (our basis matrix for Y), and S (feature matrix) becomes

$\mathbf{A} = \phi(\mathbf{X})\mathbf{W}$ and $\mathbf{B} = \lambda\mathbf{YW}$, and $\mathbf{S}$ is G^T

Kernel SSNMF Classification Theory

Theorem 9. *Since* $\mathbf{A} = \phi(\mathbf{X}_{train})\mathbf{W}$, *then the* S_{test} *matrix was given by*

$$\mathbf{S}_{test} = \mathbf{A}^{\dagger}\phi(\mathbf{X}_{test}),$$

where $\mathbf{A}^{+}$ *denotes the Moore-Penrose pseudoinverse of* $\mathbf{A}$, *and*

$$\mathbf{A}^{\dagger} = \begin{cases} \mathbf{W}^{+}\left(\phi(\mathbf{X}_{train})^{T}\phi(\mathbf{X}_{train})\right)^{-1}\phi(\mathbf{X}_{train})^{T}, & \text{if } \mathbf{X}_{train} \text{ is a tall matrix,} \\ \mathbf{W}^{+}\phi(\mathbf{X}_{train})^{T}\left(\phi(\mathbf{X}_{train})\phi(\mathbf{X}_{train})^{T}\right)^{-1}, & \text{if } \mathbf{X}_{train} \text{ is a wide matrix,} \end{cases}$$

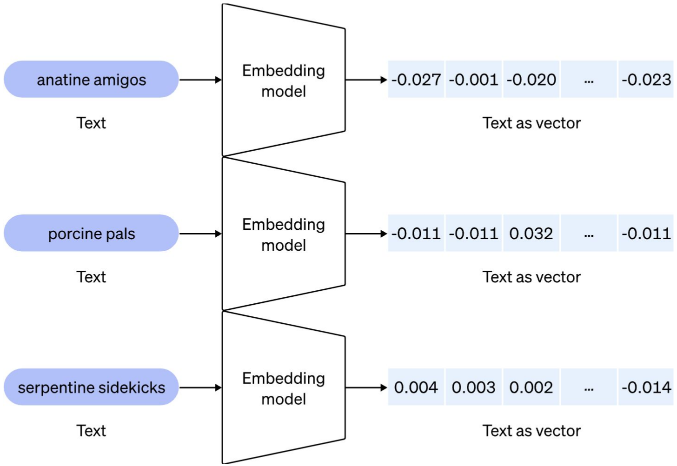
- We are primarily concerned with testing our algorithm on a tall matrix because here we would only compute the inner product verses for a wide matrix where we would compute phi for all features.

04

Experiments & Results

LLMs and Embeddings

- Large Language Models are designed using deep learning architecture known as the transformer which uses vector encodings to transfer human text.
- We transformed our text to vectors and performed simple classification tasks using SSNMF and SVM to classify cases as exonerated or non-exonerated.



Layered Summaries using GPT 3.5

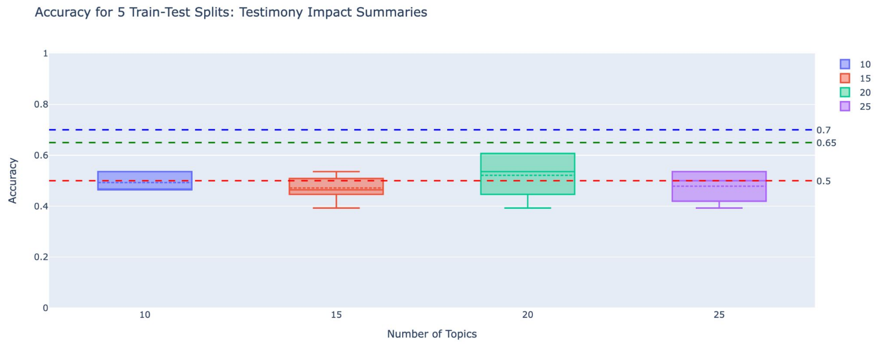
Testimony Prompt:

"Evaluate how the accuracy and reliability of eyewitness testimony influenced the outcome of this case, considering factors such as the witnesses' credibility, consistency, and potential biases."

Outcome Prompt:

"Give me a good summary for this case to help the judge decide whether exonerated or non-exonerated."

Choosing Summary Prompt



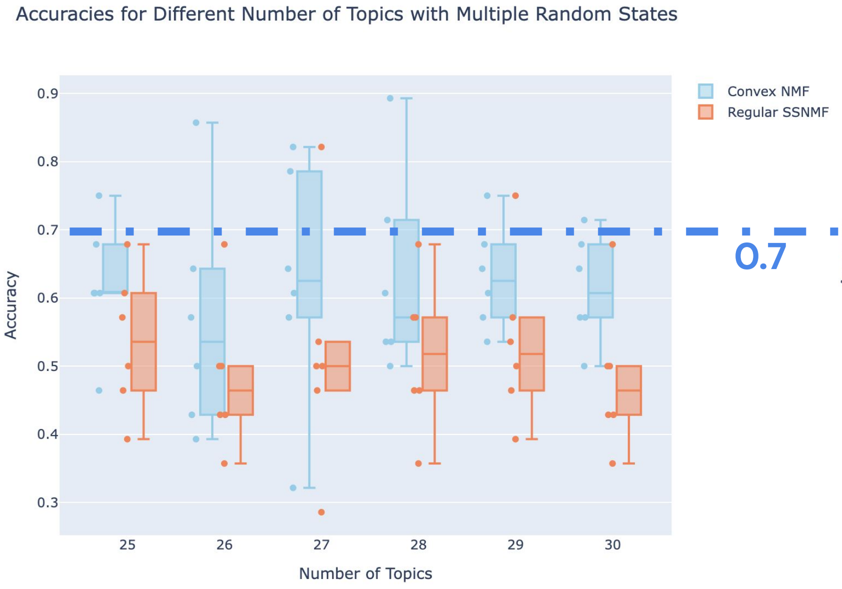
Test Kernel SSNMF on Embeddings

We test Kernel SSNMF for predicting wrongful convictions with LLM embeddings

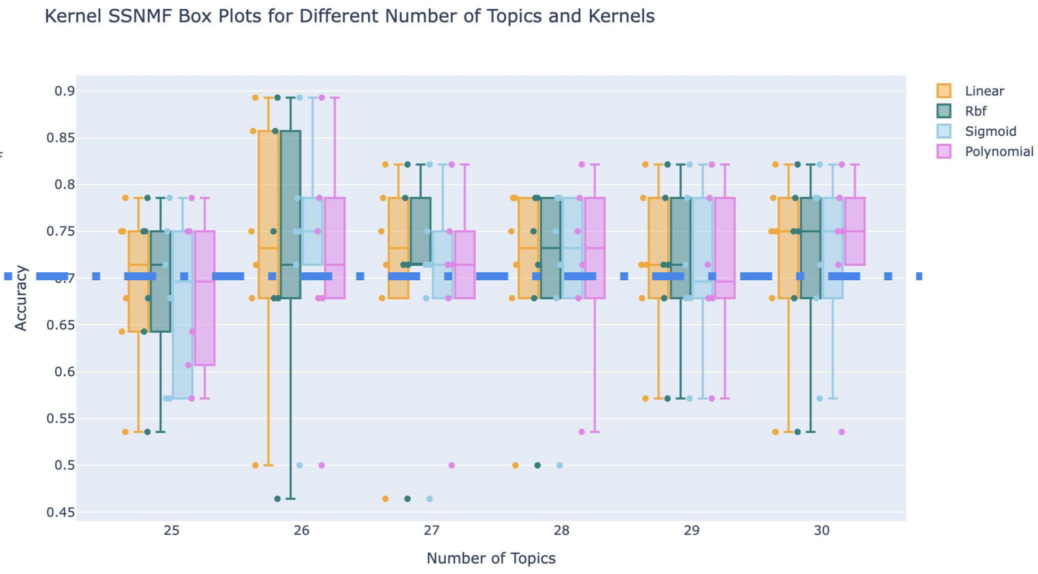
Steps for Testing:

1. Set regularization parameter λ=1, max_iter=1000 for consistency (kernel SSNMF has much faster convergence)
2. Select different numbers of topics and 6 random states for train test split
3. Run kernel SSNMF with linear, rbf, sigmoid, and polynomial kernels
4. Train SVM classifier & grid search with the reduced feature matrix to compute the test accuracy for each experiment
5. For comparison, perform the same procedure using Convex NMF and regular SSNMF

Compare Algorithm Performance

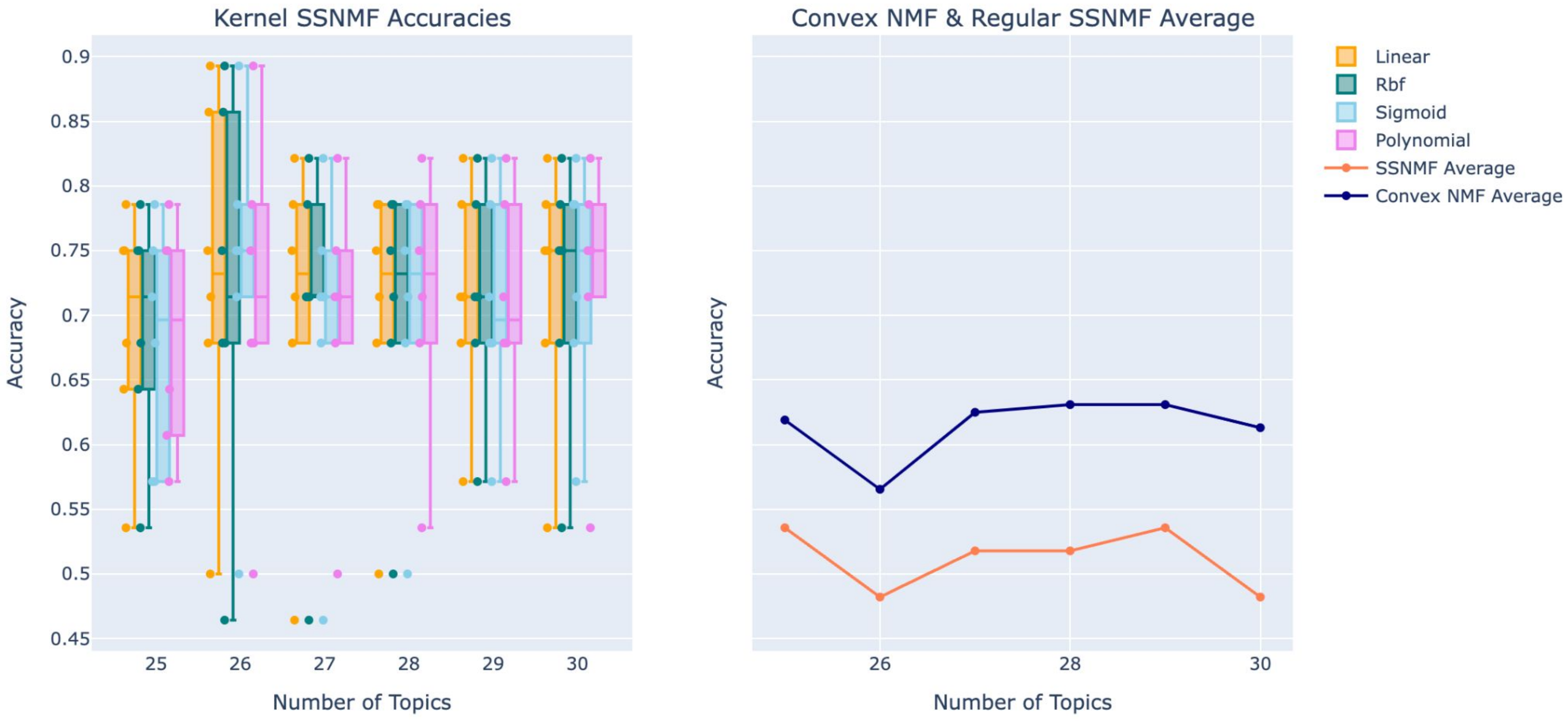


0.7



Combined Results

Kernel SSNMF Analysis



05
Evaluation & Future Directions

Evaluate Our Experiments

Strengths:

- Multiple metrics are applied to reduce randomness in testing
- LLM word embeddings of the summaries reduce dimension and cut down computation time

Future Improvements:

- Try more random states and experiments
- Get access to specialized legal LLM for more reliable summaries
- Current embeddings are at document-level. Will try interpret the textual meanings of the topics detected

Evaluate Kernel SSNMF

Strengths:

- Demonstrate robust performance in learning LLM word embeddings of long legal documents compared with benchmark algorithms
- Does not impose non-negative constraint on the data matrix
- Fast convergence
- Incorporate labeling information in training stage
- Flexibility in choice of kernels and regularization parameters

Potential Improvements:

- Implement on a wider variety of datasets to learn about its general performance

References

[1] "Prisoners in 2022 – Statistical Tables | Bureau of Justice Statistics."

[2] M. Mauer, "The crisis of the young african american male and the criminal justice system 1," in *Impacts of incarceration on the African American family*, pp. 199–218, Routledge, 2018.

[3] S. R. Gross, B. O'brien, C. Hu, and E. H. Kennedy, "Rate of false conviction of criminal defendants who are sentenced to death," *Proceedings of the National Academy of Sciences*, vol. 111, no. 20, pp. 7230–7235, 2014.

[4] C. E. Loeffler, "Measuring self-reported wrongful convictions among prisoners," *Journal of Quantitative Criminology*, vol. 35, no. 1, pp. 259–286, 2019.

[5] E. Ben-Michael, D. J. Greiner, M. Huang, K. Imai, Z. Jiang, and S. Shin, "Does ai help humans make better decisions? a methodological framework for experimental evaluation," *arXiv preprint arXiv:2403.12108*, 2024.

[6] Z. Sun, "A short survey of viewing large language models in legal aspect," *arXiv preprint arXiv:2303.09136*, 2023.

[7] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336–1353, 2013.

[8] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 4–7, 2009.

[9] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2000.

[10] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems* (T. Leen, T. Dietterich, and V. Tresp, eds.), vol. 13, MIT Press, 2000.

[11] M. Febrissy, A. Salah, M. Ailem, and M. Nadif, "Improving nmf clustering by leveraging contextual relationships among words," *Neurocomputing*, vol. 495, pp. 105–117, 2022.

[12] C. H. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45–55, 2010.

[13] P. Li, C. Tseng, Y. Zheng, J. Chew, L. Huang, B. Jarman, and D. Needell, "Guided semi-supervised non-negative matrix factorization," *Algorithms*, vol. 15, p. 136, 04 2022.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding.," *Journal of Machine Learning Research*, vol. 11, no. 1, 2010.

[15] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space," *IEEE Transactions on Neural Networks*, vol. 19, no. 6, pp. 1090–1100, 2008.

[16] M. Gao, J. Haddock, D. Molitor, D. Needell, E. Sadovnik, T. Will, and R. Zhang, "Neural nonnegative matrix factorization for hierarchical multilayer topic modeling," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 6–10, IEEE, 2019.

[17] R. Budahazy, L. Cheng, Y. Huang, A. Johnson, P. Li, J. Vendrow, Z. Wu, D. Molitor, E. Rebrova, and D. Needell, "Analysis of legal documents via non-negative matrix factorization methods," *arXiv preprint arXiv:2104.14028*, 2021.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, 10 2001.

[19] A. Ziegler and I. R. König, "Mining data with random forests: current options for real-world applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 1, pp. 55–63, 2014.

[20] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[21] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*, pp. 101–121, Elsevier, 2020.

[22] S. Bera, D. Chakrabarty, N. Flores, and M. Negahbani, "Fair algorithms for clustering," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] C. Zhang, S. H. Cen, and D. Shah, "Matrix estimation for individual fairness," in *International Conference on Machine Learning*, pp. 40871–40887, PMLR, 2023.

[24] H. Adams, L. Kassab, and D. Needell, "An adaptation for iterative structured matrix completion," *arXiv preprint arXiv:2002.02041*, 2020.

[25] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," *arXiv preprint arXiv:1903.03862*, 2019.

[26] "The Innocence Center - Securing Freedom For The Innocent."

[27] "Exoneration Detail List."

Thank you!