Exploring patterns in Computer Science Publications using the DBLP dataset.

Group Data Project (12/5/2025)

Group Members: Saloni Sharma (2170524), Aisulu Tulyeujan (2335026), Charity Smith (2095250)

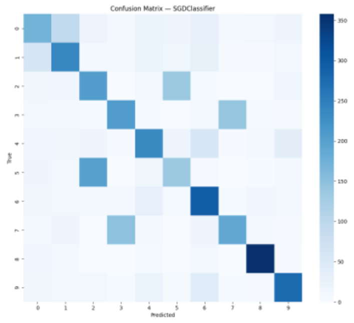
Tasks to Perform:

Regular Classification – Charity

Introduction: The goal of Regular classification is to Select high-quality venues, use venues as class labels, and construct classification models to classify papers and authors into different venues using their TF-IDF features. Compare different classifiers' performance using different evaluation metrics (e.g., accuracy, macro-F1, micro-F1).

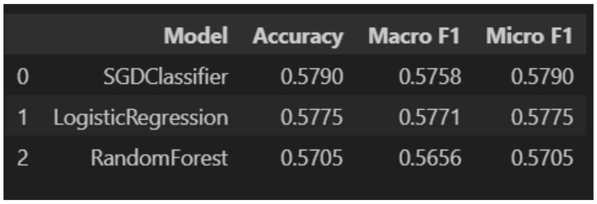
Methods:

1. Preprocessing:
    a. Removed incomplete entries
    b. Used TF-IDF vectorization on titles/abstracts using (max_features = 2000, stop_words='english'
2. Regular Classification Steps:
    1. Feature Extraction
        a. During extraction I utilized TF_IDF vectors of title and abstract
    2. Label encoding: Venue as categorical class.
    3. Model Training:
        a. Algorithms used: Logistic Regression, Random, Forest, SVM, LightGBM
        b. Evaluation metrics: Accuracy, Macro-F1, Training Time
        c. Data split: 80% training, 20% testing.
    4. Visualizations



Interpretation:

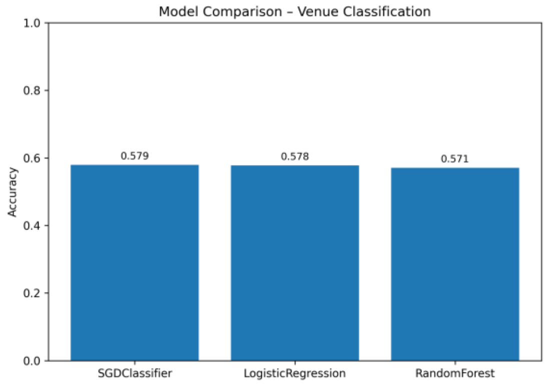
- The cells along the diagonal have the highest values, meaning that the model correctly classified most of the samples for most venues
- There is a good amount of cells off the diagonal, proving that many venues share overlapping text patterns, which lead misclassification
- Classes 6,8, and 9 have stronger diagonal values, showing that they have more distinctive vocabulary or topics, making them easier to classify.
- Other classes, much as 0, 2, and 4 show more spread-out predictions across multiple classes, suggesting topic similarity with other venues.



	Model	Accuracy	Macro F1	Micro F1
0	SGDClassifier	0.5790	0.5758	0.5790
1	LogisticRegression	0.5775	0.5771	0.5775
2	RandomForest	0.5705	0.5656	0.5705

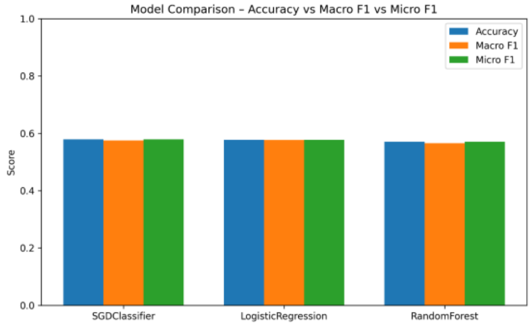
Interpretation:

- All models perform similarly, between 57%- 58% accuracy), which is somewhat reasonable given high venue overlap and short text length
- The SGDClassifier performed the best.
- Random performed the worst, probably attributed to the TF_IDF vectors that are sparse and high-dimension making linear models more effective.
- Moderate F1 scores indicate that many venues share vocabulary, causing misclassification



Interpretation:

- All three models: SGDClassifer, Logistic Regression, and Random Forest, perform very similarly across accuracy, macro-F1, and micro-F1 scores around 0.57- 0.58.
- The closeness for macro-F1 and micro-F1 suggests a balanced class performance. No one single venue class overwhelmingly dominates the dataset.
- The slightly higher performance from SGDClassifier suggests that linear classifiers handle TF-IDF features more effectively than tree-based methods in this dataset.



Interpretation:

- SGDclassifier yields the highest accuracy (0.579), followed by Logistic Regression (0.578)
- Random Forest has the lowest accuracy (0.571), confirming that non-linear tree models do not generalize as well on TF-TDF features.
- The small difference in accuracy across all models shows that classification task is inherently challenging due to overlapping venue language, not because of the model itself.
- The 58% accuracy suggests there is significant overlap between venues that limit prediction accuracy.

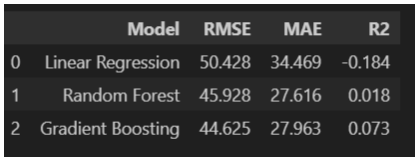
Paper Impact Prediction – Charity

Introduction: The goal of the Paper Impact Prediction is to predict a paper citation count using count using the title, year, author count, and venue dummy variables.

Expectation: Accurate and reasonable predictions as well as important predictive features e.g., papers with more authors often correlate with broader collaboration, which leads to higher visibility.

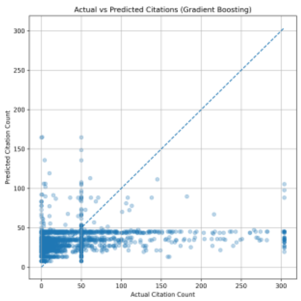
Methods:

1. Feature Extraction:
   a. Utilizing TF-IDF again for embedding for titles/abstracts
   b. Numeric metadata: year, author count, venue(one-hot encoded)
2. Models:
   a. Linear regression, Random Forest Regression, Gradient Boosting regressor
3. Evaluation metrics:
   a. RMSE (Root Mean Squared Error)
   b. MAE (Mean Absolute Error)
   c. R^2 (Explained variance)
4. Visualization interpretation
   a. Examining what factors impact most

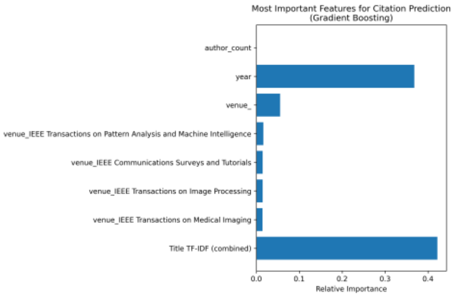


	Model	RMSE	MAE	R2
0	Linear Regression	50.428	34.469	-0.184
1	Random Forest	45.928	27.616	0.018
2	Gradient Boosting	44.625	27.963	0.073

Interpretation: While gradient Boosting performs better than Linear Regression and Random Forest, the R^2 values are close to zero, meaning the models explain only a small portion of the variance in citation counts. These results are to be expected because citation behavior is extremely noisy.



Interpretation: The scatterplot shows most of the predicted values cluster between 0 to 50 citations, even though actual citation counts go much higher. This reflects the highly skewed distribution of citations, where a small number of papers receive very high citation counts about most papers receive relatively few. Most points fall below the line. This suggests that the model underestimates high-impact papers consistently. This model captures general trends but fails to capture precision for extreme values.



Interpretation: The gradient Boosting model identifies title-based TF-IDF features as the strongest predictor of citation count, suggesting that the specific language and topics in a paper's title carry the most meaning for estimating impact.

Limitations: Because the DBLP dataset is very large and TF-IDF features are high-dimensional, training full linear SVM / logistic models on all papers exceeding the computational resources of my environment. Therefore, I (1) restricted our analysis to high-volume venues, (2) down sampled each venue to at most N papers, and (3) limited the TF-IDF vocabulary. This preserved class balance while making training feasible.

Conclusion: the classification models achieved moderate performance indicating that text features alone only partially differentiate venues. Citation prediction was more difficult, with low R^2 values, showing that impact is difficult to estimate from metadata and titles alone.

EDA – Saloni

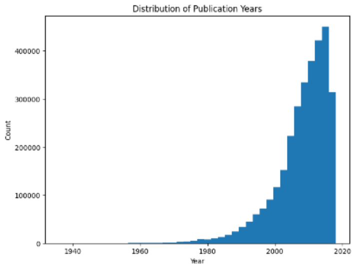
Introduction

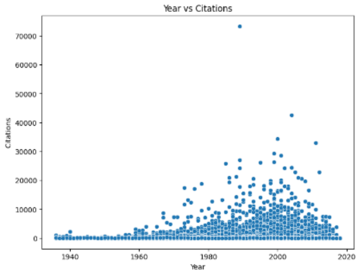
The goal of this Exploratory Data Analysis (EDA) was to uncover underlying patterns, relationships, and key characteristics within a dataset of academic publications by using engineering techniques like normalization and Principal Component Analysis (PCA). The initial findings indicate significant data skewness and the presence of outliers, which informs the subsequent data preprocessing and model building phases.

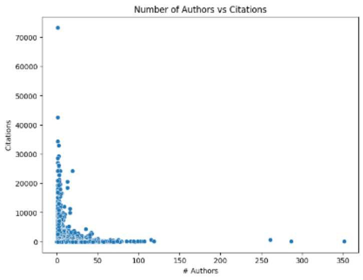
Methods:

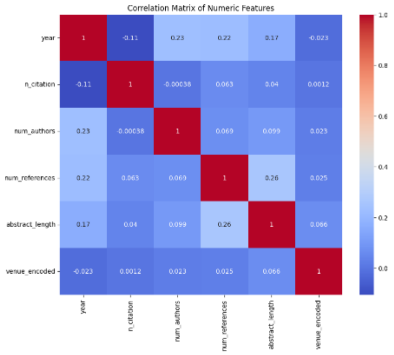
1. Univariate Analysis (Distributions): Histograms were used to visualize the distributions of continuous variables, specifically Citation Counts and Publication Years. This helps identify the central tendency, spread, and shape (e.g., skewness).
2. Bivariate Analysis (Relationships): Scatter Plots were used to investigate the relationships between key variables, such as Year vs. Citations and Number of Authors vs. Citations. This helps identify correlations and trends over time. A Correlation Matrix (Heatmap) was used to quantify the linear relationships between all numerical features.
3. Dimensionality Reduction Analysis: Principal Component Analysis (PCA) was applied to the numerical features at different stages of preprocessing (Original, Z-Score Normalized, and Min-Max Normalized) and visualized using 2D scatter plots (PC1 vs. PC2). This method helped assess the impact of normalization on the data's structure and its suitability for clustering or modeling.
4. Preprocessing: Cleaning involved dropping rows with missing title/year, ensuring year was an integer, and filling missing abstract values. Feature engineering then created four key metrics: num_authors, num_references, abstract_length (word count), and venue_encoded (categorical code). All resulting numerical columns were then grouped into the numeric_df DataFrame for subsequent analysis.

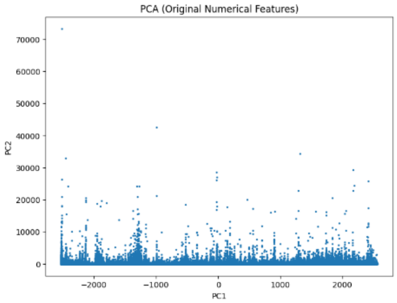
5. Visualizations:

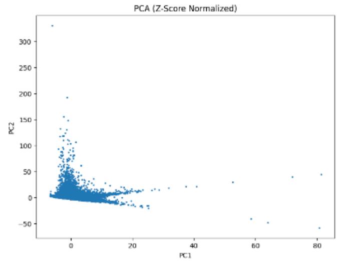












*not all graphs are in the report as we had to keep in under 10 pages. Rest can be found in code.

Conclusion

What worked: The visualizations confirmed an expected right-skewness in Publication Years, showing a clear increase in paper volume toward recent years. The scatter plots successfully identified influential outliers (papers with ≈ 75,000 citations), which stand out dramatically from the majority of the data. Crucially, the Correlation Matrix confirmed that n_citation has a very low linear correlation with all other numerical features, indicating simple factors alone are poor predictors of impact.

What didn't work: The histograms for Citation Counts and Number of Authors were uninformative due to extreme right-skewness, collapsing most data into the first bin and masking distribution details. PCA on the Original Data failed to show structure due to the dominant scale of features like raw citations. PCA on Min-Max Normalized Data was also ineffective, resulting in a compressed plot that removed too much useful variance. The Z-Score Normalized PCA was the most successful, yielding a spread-out structure that retained variance, making it the preferred scaling method for subsequent modeling.

Clustering – Saloni

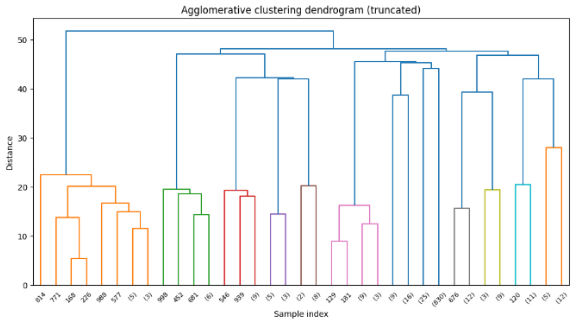
Introduction

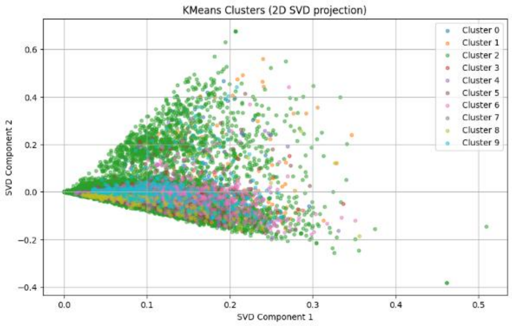
The goal of this analysis is to uncover meaningful groupings within a large collection of academic papers, venues, and authors. Clustering helps identify common topics, research areas, and communities of researchers. Clustering also helps further analysis, such as anomaly detection.

Methods:

1. Preprocessing: Each paper's text was constructed by combining its title and abstract. Missing or empty text entries were removed. Text features were converted to numerical vectors using TF-IDF with a maximum of 5,000 features. Dimensionality reduction via TruncatedSVD reduced these vectors to 100 components for clustering.

2. Clustering Experiments: KMeans (MiniBatch) clustering were applied to papers, venues, and authors with k = 2, 5, 10, 20 clusters. Silhouette scores, SSE (sum of squared errors), and cluster purity were computed to evaluate cluster quality.
    a. Papers (k=10): SSE ≈ 92.92, Silhouette ≈ 0.079
    b. Authors (k=10): Silhouette ≈ 0.033
    c. Venues (k=10): Silhouette ≈ 0.023–0.049 depending on normalization
3. DBSCAN: Density-based clustering with varying epsilon values showed that small epsilon values yield few clusters with many noise points. Larger epsilon values reduce noise but produce more clusters.
    a. At eps = 1.1: 34 clusters, Silhouette ≈ 0.367
    b. At eps = 1.3: 60 clusters, Silhouette ≈ 0.349
4. Hierarchical Clustering (Agglomerative): Hierarchical clustering was applied to papers with different distance thresholds:
    a. Threshold = 10 → 5,724 clusters
    b. Threshold = 25 → 597 clusters
    c. Threshold = 50 → 121 clusters
5. Visualizations





Dendrogram: Hierarchical clustering dendrogram reveals cluster hierarchies and relationships between papers.

KMeans Projection: The 10 clusters are visible in the 2D SVD projection, showing overlapping but distinguishable topic groups.

Conclusion

Papers: KMeans clustering with k=10 provides interpretable topics with moderate silhouette scores, indicating that broad thematic separation works well. DBSCAN highlights dense topics but generates many noise points for small epsilon values, revealing emerging or niche topics that do not fit neatly into major clusters. Hierarchical clustering captures a nested structure suitable for multi-scale analysis, showing both broad topics at higher levels and finer subtopics at lower levels. Overall, papers form recognizable clusters, but interdisciplinary or unique papers create natural outliers.

Authors: Clustering of authors indicates research communities with shared topical interests. However, silhouette scores are low, suggesting overlapping areas of expertise and significant collaboration across topics. This demonstrates that while communities exist, boundaries are fuzzy due to interdisciplinary research and multiple collaborations among authors.

Venues: Clustering reflects thematic similarity between venues, though the small number of venues limits distinct separation. Some venues cover multiple topics, making clusters less granular compared to paper-

or author-level clustering. Nevertheless, thematic patterns are still visible, providing insight into the broader research landscape.

Overall: The combination of KMeans, DBSCAN, and hierarchical clustering allows multi-perspective exploration of topics and research communities. KMeans provides interpretable main topics, DBSCAN identifies dense areas and outliers, and hierarchical clustering reveals nested, multi-scale relationships. Together, these methods highlight both clear clusters and overlapping interdisciplinary areas in papers, authors, and venues, illustrating the complexity of the research ecosystem.

Anomaly Detection – Aisulu

Introduction

The main idea of anomaly detection in this context is to find papers whose topics do not match the typical topics of the venue where they were published. These unusual papers might represent special sessions, interdisciplinary work, or possible data inconsistencies.

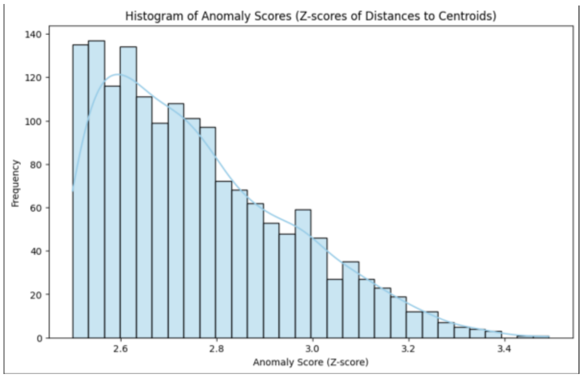
Our goal was to identify these outliers using clustering, TF-IDF features, and distance-based scoring.

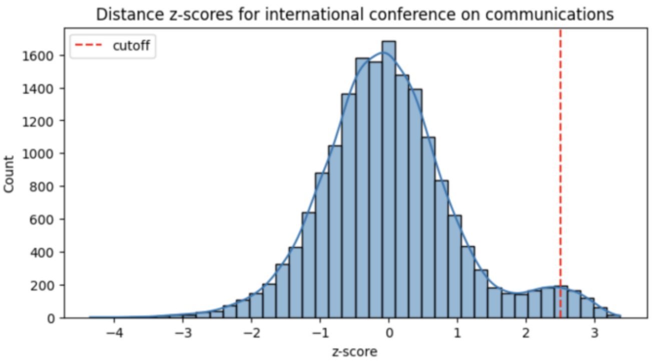
Data description

The dataset loading code iterated through JSON files and extracted papers from the top 5 largest venues, with 114,716 records.

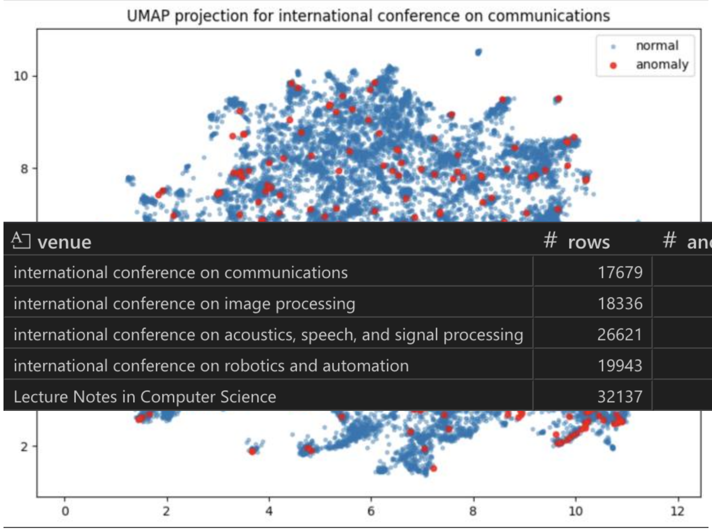
Methods

1. Preprocessing
   Each paper's text was built by combining title and abstract. The text was cleaned using a helper function to remove empty values.
2. TF-IDF Vectorization
   All paper texts are converted into vector features using TF-IDF.
3. Clustering per Venue
   For each venue:
   I. Extract all papers belonging to that venue.
   II. Fit K-Means clustering.
   III. Compute the distance of each paper vector to its cluster centroid.
   IV. Convert distances to z-scores.
   V. Mark papers as anomalies.
4. Visualization





The above histograms show how distances spread across papers. The first one shows among all the papers, while the second one shows only the “international conference on communications”. The histogram shows a long tail, meaning some papers are far from the mean distance, strong signal of real anomalies.


UMAP projection for international conference on communications

Above UMAP shows clusters and anomalies in 2-D space. Anomalies are colored red, normals blue. This helps visually confirm that the red points lie on the outer region.

Results

venue	# rows	# anomalies
international conference on communications	17679	447
international conference on image processing	18336	397
international conference on acoustics, speech, and signal processing	26621	363
international conference on robotics and automation	19943	214
Lecture Notes in Computer Science	32137	201

This table shows the number of anomalies detected in each top venue.

- Total papers: 114,716
- Total anomalies: 1,622
- Percentage: 1.41%

Conclusion

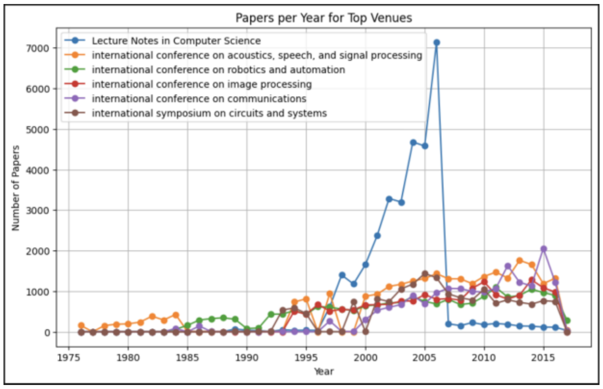
In this part of the project, I successfully applied clustering-based anomaly detection to the DBLP dataset. Using TF-IDF and K-Means, I identified about 1,622 anomalous papers, representing 1.41% of the total papers from the top venues. Visualizations such as histograms and UMAP projections supported these findings and helped confirm the structural differences between normal and anomalous papers.

Temporal Classification - Aisulu

Introduction

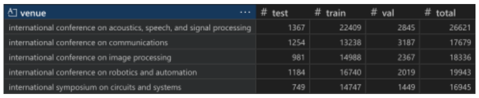
Goal: Predict the venue of future papers using only older historical papers.
Method: TF-IDF features + several machine learning classifiers evaluated over multiple time thresholds.

Data Description


Papers per Year for Top Venues

Same data as the anomaly detection task. But in this task, I used the 5 top venues excluding the “Lecture Notes in Computer Science”. Because that venue had extremely high number of papers published between the years 2000 and 2007. That caused an imbalance between the training and test data.

Train: 82122; Val: 11867; Test: 5535



venue	# test	# train	# val	# total
international conference on acoustics, speech, and signal processing	1367	22409	2845	26621
international conference on communications	1254	13238	3187	17679
international conference on image processing	981	14988	2367	18336
international conference on robotics and automation	1184	16740	2019	19943
international symposium on circuits and systems	749	14747	1449	16945

These venues have a balanced test, val and train data as shown in the above table.

Methods and Algorithms

The preprocessing and TF-IDF steps are the same as in the Anomaly Detection.

1. Choose year thresholds
   Thresholds: 2013, 2015
   Papers before the threshold are training data, until the next threshold is validation data and the rest are test data.
2. Class balance view: Compute counts per venue per split to see imbalance before modeling.
3. Majority-class classifier: always predicts the most common venue (gives baseline accuracy/F1 for val/test).
4. Logistic Regression + word TF-IDF (1–2 grams, max_features=10k, stopwords English, min_df=2), class_weight balanced. This is the main baseline.

Model tuning

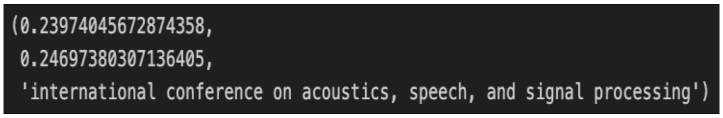
1. Sweep two Logistic Regression setups on the validation split:
   Word TF-IDF (same as above), C ∈ {0.5, 1.0}
   Char TF-IDF (3–5 char n-grams, max_features=50k, min_df=2), C ∈ {0.5, 1.0}
2. Pick the model with the best validation macro-F1; evaluate it on both val and test.

Interpretability: For both the initial word model and the tuned best model, list top weighted TF-IDF terms per venue from the Logistic Regression coefficients.

Fast mode: Optional capped sampling for speed; repeats word TF-IDF Logistic Regression with the same C grid to get quick feedback.

Results & analysis (structure)

1. The notebook prints accuracy and macro-F1 for val/test for majority, base word model, tuned best (word or char), and fast mode if used.
2. Classification reports (per-class precision/recall/F1) are shown for the main model and the tuned best model.
3. Top terms per venue help explain what the classifier learns.

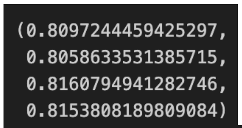


```
(0.23974045672874358,
 0.24697380307136405,
 'international conference on acoustics, speech, and signal processing')
```

If the most common venue is always guessed, it would be correct only about 24% of the time.

The models must beat this.

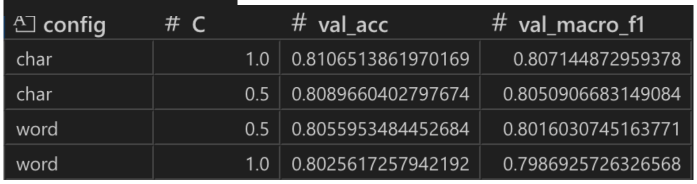
- Logistic regression results:



```
(0.8097244459425297,
 0.8058633531385715,
 0.8160794941282746,
 0.8153808189809084)
```

This model performs very well, far above the baseline (~0.24). It predicts venues correctly about 81% of the time.

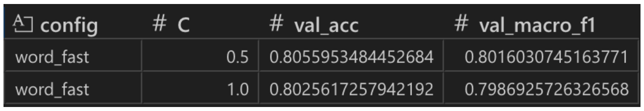
- Word vs character sweep



config	# C	# val_acc	# val_macro_f1
char	1.0	0.8106513861970169	0.807144872959378
char	0.5	0.8089660402797674	0.8050906683149084
word	0.5	0.8055953484452684	0.8016030745163771
word	1.0	0.8025617257942192	0.7986925726326568

The model is learning meaningful signal words that are representative of each field. This is a powerful interpretability result. Character n-grams performed best overall, beating the word-level model.

- Fast Sweep: This final section runs a simplified version of the model on smaller samples for debugging.



config	# C	# val_acc	# val_macro_f1
word_fast	0.5	0.8055953484452684	0.8016030745163771
word_fast	1.0	0.8025617257942192	0.7986925726326568

The fast sample still achieves strong performance, confirming consistency.

What worked:

1. Streaming JSON keeps memory low.
2. Temporal split prevents leakage from future papers.
3. TF-IDF + Logistic Regression trains fast and gives interpretable weights.

Conclusion

1. A simple word/char TF-IDF + Logistic Regression provides a solid, interpretable temporal baseline for venue prediction.
2. Validation macro-F1 is used to choose between word and char models; whichever wins on val is used on test.