Motor Insurance Part Mapping and Recommendation

Ctrl Alt Del

DecodeX

February 15, 2025

Himani Grover Khushi Gupta Shieladitya Basu

Contents

1 Business Objective

1.1 Background

The motor insurance industry is a very competitive industry and **AutoSure Insurance** has built a strong reputation for having an efficient and reliable claim settlement procedure. However, despite its operational excellence, the company faces a persistent issue within its **claims department**—there have been issues pertaining **inconsistencies in damage assessment and claim verification**.

Whenever an accident occurs, AutoSure Insurance follows a structured claim process:

- A **surveyor** assesses the vehicle's damage and manually enters the details into the company's system.
- The **garage** performing the repairs submits its own **list of damaged parts**.
- The **insurance team** compares these two lists before processing payments to verify claim accuracy and prevent fraud.

Now despite following this well-defined process, the following challenges arise:

- **Inconsistent Part Names**: Surveyors and garages often describe the same part differently (e.g., "Left Door" vs. "Driver-Side Door"), making it difficult to match records.
- **Data Standardization Issues**: The absence of a standardized part identification system leads to delayed claim processing and potential fraud.

1.2 Business Problem

To address these challenges, AutoSure Insurance **aims to implement an automated mapping system** to enhance the accuracy of **damage assessment and claim verification**.

The objective of this report is to enhance the accuracy of damage assessment and streamline claim verification for **AutoSure Insurance** by implementing a systematic mapping of damaged parts between the garage and surveyor datasets. Given the vast number of claims and the inconsistencies in part descriptions, ensuring precise alignment presents a significant challenge.

There are **huge business impacts** if a solution is not rapidly implemented. The company will face:

- **Increased claim processing costs** due to prolonged manual verification.
- **Higher risk of fraudulent claims**, negatively impacting financial sustainability.
- **Longer claim settlement times**, leading to reduced customer satisfaction.
- **Loss of competitive advantage**, as industry competitors adopt more automated, data-driven solutions.

1.3 Key Risks

The implementation of an **automated part mapping and claim verification system** presents several potential risks that must be carefully managed to ensure successful deployment.

1. Data Quality and Standardization Risks

A significant challenges in this project is ensuring **data consistency and accuracy**. Since different stakeholders (surveyors, garages, and insurance teams) enter part details manually, **inconsistencies, missing data, and duplicate entries** may affect the reliability of the system.

Inconsistent or incomplete data: If part descriptions are missing or incorrect, the mapping process may fail, leading to inaccurate claim

verification. Implementing **data validation checks** and leveraging historical data can help mitigate this risk.

2. Model Performance Risks

The accuracy of the system heavily depends on **Fuzzy Matching algorithm and Term Frequency-Inverse Document Frequency (TFIDF)**. If these models do not perform as expected, claim verification errors may occur.

3. Operational and Integration Risks

The success of the system depends on **seamless integration with existing workflows** and minimal disruption to ongoing claim processing. The new streamlit-based UI proposed system must integrate with AutoSure Insurance's existing claim management software and databases. If compatibility issues arise, the implementation may face delays.

2 Data Preparation and Availability

2.1 Data Description

The description of the data at hand is given bellow:

Dataset	Variable	Description
Garage Data	REFERENCE_NUM	Unique identifier for each record.
	PRODUCT_INDEX	Code representing the product category (e.g., car, truck).
	VEHICLE_MODEL_CODE	Code representing the specific vehicle model.
	CLAIMNO / CLAIM_NO / NUM_CLAIM_NO	Claim identification number, linking records to insurance claims.
	PARTNO	Unique part number for the claimed or repaired vehicle part.

Dataset	Variable	Description
	PARTDESCRIPTION	Text description of the part (e.g., "Brake Pad", "Fuel Pump").
	TOTAL_AMOUNT	Total cost associated with the claim, covering parts.
Surveyor Data	REFERENCE_NUM	Unique identifier for each surveyor's assessment record.
	PRODUCT_INDEX	Code representing the product category.
	VEHICLE_MODEL_CODE	Code representing the specific vehicle model.
	CLAIMNO / NUM_CLAIM_NO	Claim identification number linking to garage data.
	TXT_PARTS_GROUP_NAME	General category of the part (e.g., "Engine", "Brakes").
	TXT_PARTS_NAME	Specific part name (e.g., "Brake Pad", "Radiator").
	TOTAL_AMOUNT	Total cost associated with the claim, covering parts.
	NUM_PART_CODE	Unique numeric code representing the specific primary part.
Primary Part Code Data	PRODUCT	Category of the vehicle (e.g., "Private Car").
	SURVEYOR PART CODE	Unique code assigned by surveyors for each primary vehicle part.
	SURVEYOR PART NAME	Standardized name of the primary vehicle part.

2.2 Data Cleaning

To ensure consistency in part descriptions across the two datasets, we implemented a text-cleaning function to standardize naming conventions and remove unnecessary elements. The key steps of this process are as follows:

- Synonym Replacement:

  A predefined dictionary (synonyms) maps common automotive terms to their standardized equivalents. For example, "Bonnet" is replaced with "Hood," and "Boot" is replaced with "Trunk." This helps unify terminology differences across datasets.

- Removal of Stop Words:

  A list of irrelevant terms (stopwords), such as "Assembly," "Unit," and "Panel," is defined and removed from part descriptions. These words do not contribute to the unique identification of a part and are omitted to improve clarity.

- Special Character Removal:

  Non-alphanumeric characters are removed using a regular expression to maintain a clean and structured format.

- Case Standardization

  All text is converted to lowercase to eliminate case sensitivity inconsistencies.

- Handling Missing Data:

  Any rows containing missing values are removed to ensure data integrity before further analysis.

The cleaning function is applied to part descriptions in both the surveyor dataset (TXT_PARTS_NAME) and the garage dataset (PARTDESCRIPTION). The cleaned values are stored in new columns in the two dataframes.

3 Proposed Approach

The methodology employed for mapping motor insurance parts between surveyor and garage data consists of a series of text-processing and machine-

learning-based similarity techniques. The approach involves data preprocessing, vectorization, similarity computation, and hybrid matching using both TF-IDF and fuzzy matching.

The steps are outlined below:

1) TF-IDF Vectorization for Text Similarity

To quantify the textual similarity between surveyed parts and garage parts, the **Term Frequency-Inverse Document Frequency (TF-IDF)** technique was employed. This method transforms textual descriptions into numerical vectors while emphasizing distinguishing terms. The TF-IDF matrix was constructed using the `TfidfVectorizer` module.

Here, the `CLEAN_TXT_PARTS_NAME` from the surveyor dataset and `CLEAN_PARTDESCRIPTION` from the garage dataset were concatenated into a single corpus. The `TfidfVectorizer` then converted this corpus into a high-dimensional vector space, where each document (i.e., part description) is represented as a numerical feature vector.

2) Splitting TF-IDF Matrices

Once the TF-IDF matrix was generated, it was partitioned into two sub-matrices corresponding to the surveyor and garage datasets. The first segment, `survey_tfidf`, contains vectorized representations of surveyed parts, while `garage_tfidf` represents garage parts. This division enables pairwise similarity computation in the subsequent steps.

3) Cosine Similarity Computation

Cosine similarity was utilized to measure the degree of similarity between surveyed and garage parts based on their TF-IDF representations. The **cosine similarity metric** evaluates the angle between two vectors in the high-dimensional TF-IDF space, producing a similarity score between 0 (completely dissimilar) and 1 (identical). The computed **similarity matrix** contains pairwise similarity scores between all surveyed parts and all garage parts.

4) Identifying Best Matches

The best-matching garage part for each surveyed part was determined based on the highest cosine similarity score.

5) Hybrid Matching with TF-IDF and Fuzzy String Matching

To further refine the matching process, **Fuzzy String Matching** was incorporated alongside TF-IDF similarity. The **FuzzyWuzzy** library's `token_sort_ratio` function was used to account for spelling variations and word ordering discrepancies.

Here, the final similarity score was obtained by taking the maximum of the **TF-IDF cosine similarity score** and the **fuzzy matching score**. If the highest score met or exceeded the predefined **threshold of 0.5**, the match was considered valid.

6) Exporting Results

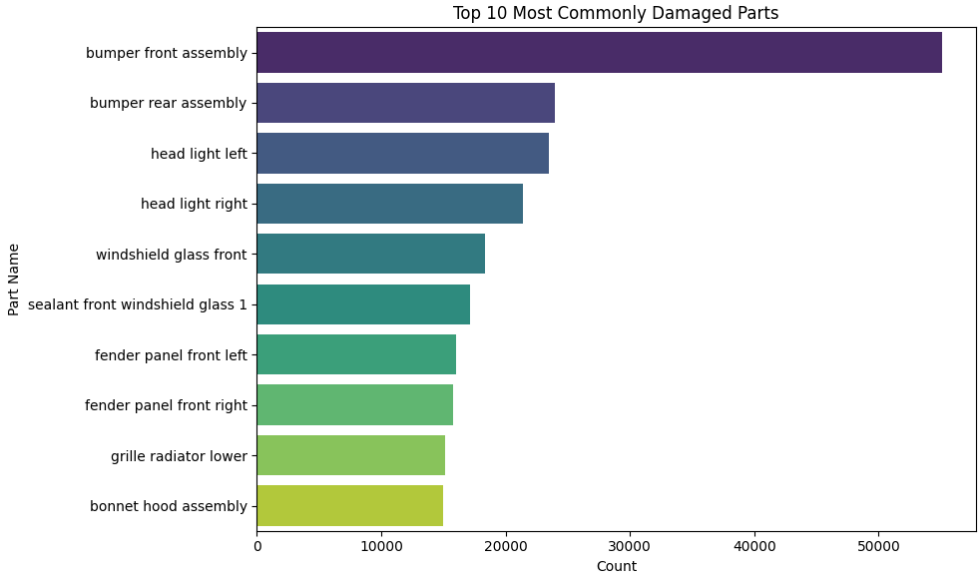
The finalized matching results, including part names, similarity scores, and the method used, were stored in a structured format and exported as a CSV file.

The proposed methodology ensures an accurate and efficient mapping of parts by leveraging a hybrid matching framework that integrates TF-IDF-based cosine similarity with fuzzy string matching. The combined approach enhances match quality, especially in cases where textual descriptions exhibit variations in word order, abbreviations, or minor spelling inconsistencies. The application of a similarity threshold further improves the robustness of the system by filtering out weak matches.

This approach provides a scalable and automated solution for motor insurance part mapping, enabling streamlined claims processing and inventory management in the insurance sector.

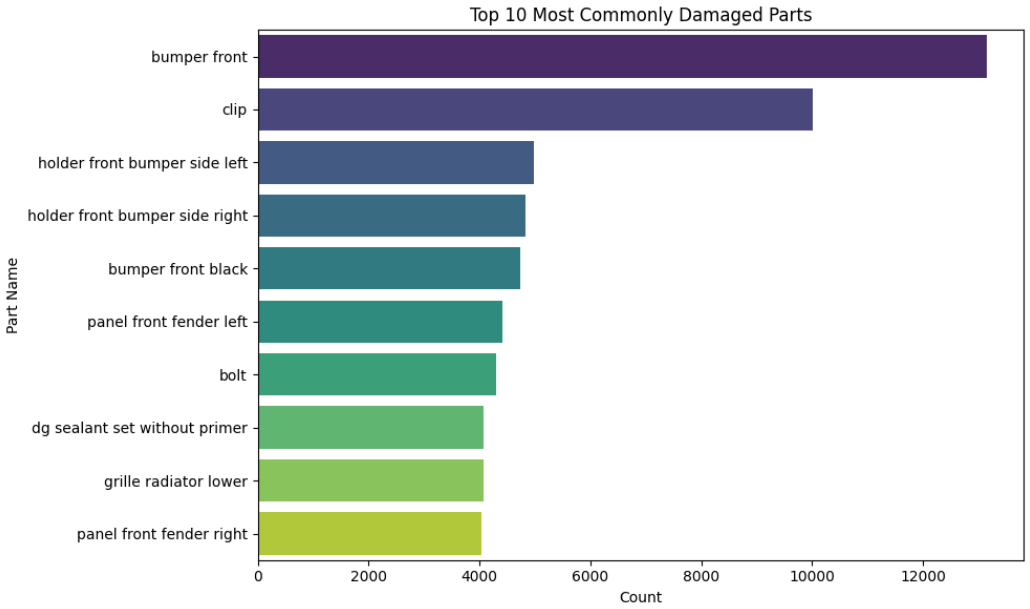
4 Data Analysis

- Here we have the most commonly damaged parts according to the **surveyors**, along with their count and percentages.



	Part Name	Count	Percentage
0	bumper front assembly	55085	5.827768
1	bumper rear assembly	23978	2.536775
2	head light left	23450	2.480914
3	head light right	21422	2.266360
4	windshield glass front	18369	1.943365
5	sealant front windshield glass 1	17169	1.816410
6	fender panel front left	16060	1.699083
7	fender panel front right	15754	1.666709
8	grille radiator lower	15171	1.605030
9	bonnet hood assembly	14960	1.582707

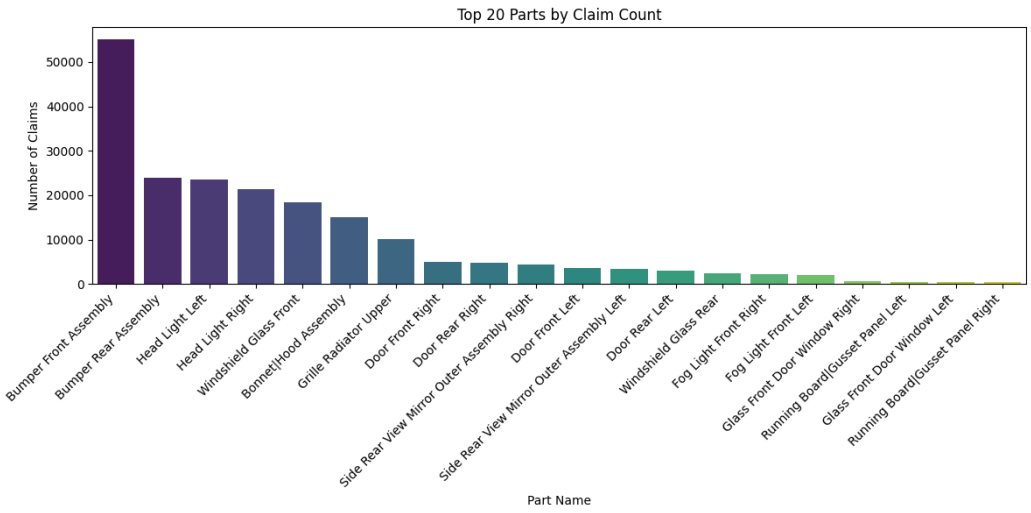
- Here we have the most commonly damaged parts according to the **garage**, along with their count and percentages.



	Part Name	Count	Percentage
0	bumper front	13155	3.603261
1	clip	10006	2.740724
2	holder front bumper side left	4974	1.362419
3	holder front bumper side right	4838	1.325167
4	bumper front black	4744	1.299420
5	panel front fender left	4418	1.210126
6	bolt	4297	1.176983
7	dg sealant set without primer	4072	1.115354
8	grille radiator lower	4067	1.113984
9	panel front fender right	4043	1.107410

Using the primary parts dataset and the surveyor data,

- The graph below shows the distribution of claims by **primary parts.**



Surveyor Part Name	Claim Count
Bumper Front Assembly	55074
Bumper Rear Assembly	23974
Head Light Left	23447
Head Light Right	21415
Windshield Glass Front	18355
Bonnet|Hood Assembly	14952
Grille Radiator Upper	10071
Door Front Right	4996
Door Rear Right	4847
Side Rear View Mirror Outer Assembly Right	4470
Door Front Left	3651
Side Rear View Mirror Outer Assembly Left	3304
Door Rear Left	3036
Windshield Glass Rear	2448
Fog Light Front Right	2285
Fog Light Front Left	2069
Glass Front Door Window Right	622
Running Board|Gusset Panel Left	510
Glass Front Door Window Left	481
Running Board|Gusset Panel Right	390

The table below shows **the top 10 secondary parts associated with the primary part** Bonnet|Hood Assembly. The respective claim counts are also mentioned.

Primary Part	Secondary Part	Claim Count
Bonnet|Hood Assembly	ac condenser assembly	11
	hinge bonnet hood left	9
	condenser assembly	5
	hinge comp front hood left	3
	back panel assembly	2
	hinge comp front hood left hand	2
	back s assembly front right hand	1
	bonnet hood lock panel	1
	connrod assembly	1
	controller assembly	1

Similar, here we have the associated secondary parts for Bumper Front Assembly and so on for the other primary parts.

Primary Part	Secondary Part	Claim Count
Bumper Front Assembly	bumper front upper	8
	bracket front bumper right	6
	cover fog light right	5
	bumper absorber front 1	3
	bumper front lower	3
	absorber comp front bumper lower	2
	bracket front bumper left	2
	cover fog light left	2
	crossmember front lower	2
	absorber front bumper lower	1

Recommending associated damaged parts using historical claim data. Our recommeder system works as follows:

- **Extract** historical claim data to identify frequently co-damaged parts.
- **Compute** part-pair co-occurrence counts and calculates Lift Scores.
- Identify **the most strongly associated parts based on Lift Scores.**

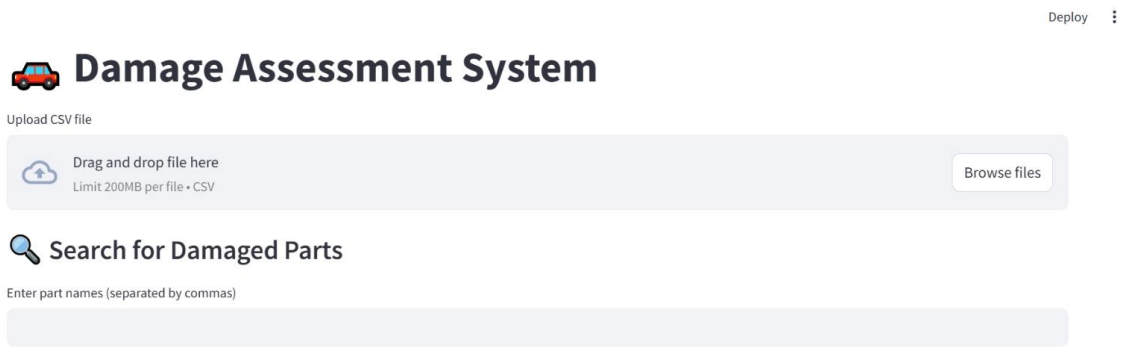
- Provide recommendations for associated parts when a user selects specific damaged parts.
- Displays results in a structured format.

This method enables dynamic part recommendation based on historical claim patterns, helping insurance analysts predict commonly co-damaged parts.

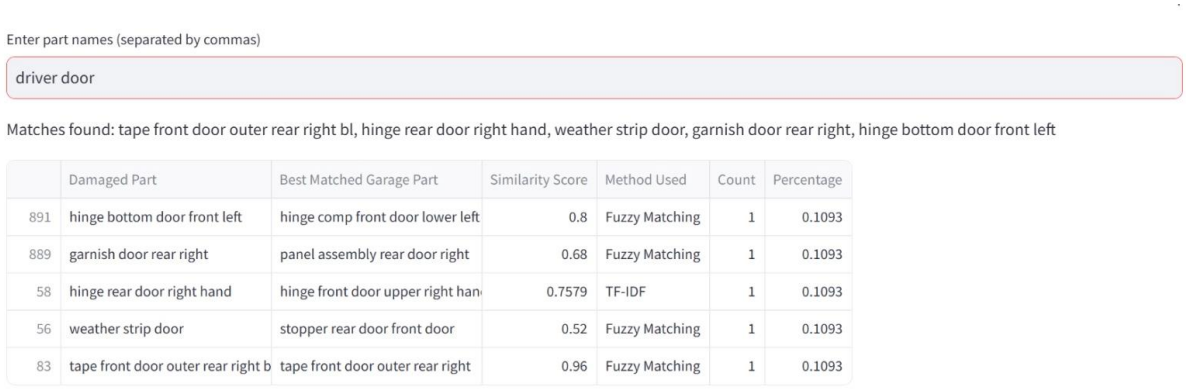
Selected Parts	Recommended Parts
('windshield glass front')	[('moulding front windshield', 0.207), ('sealant front windshield glass 1', 0.103), ('head light right', 0.069), ('bonnet hood assembly', 0.034), ('cable comp hood latch release', 0.034)]
('windshield glass front', 'bumper front assembly')	[('sealant front windshield glass 1', 0.103), ('grille radiator lower', 0.093), ('fender panel front left', 0.037), ('brace, lamp support r', 0.037), ('bonnet hood assembly', 0.034)]
('head light right', 'grille radiator upper')	[('member lamp support rh', 0.167), ('member ho lock c', 0.167), ('carrier radiator support upper member r', 0.167), ('bumper absorber front r', 0.167), ('fog light front left', 0.167)]
('fender panel front right', 'bonnet hood assembly')	[('bonnet hood assembly', 0.095), ('hinge bonnet hood left', 0.095), ('fender panel front left', 0.071), ('member comp hood lock r', 0.048), ('hose discharge', 0.048)]
('brace', 'lamp support')	[('airbag dashboard passenger side', 0.5), ('bulb', 0.5), ('lamp support lh', 0.5), ('brkt radtr support r', 0.5), ('carrier radiator support lower member', 0.333)]

5 UI Integration

A Streamlit-based UI will be integrated with AutoSure Insurance's claim management system for real-time mapping and recommendations. Here are a few snippets of the proposed beta-version of the site:

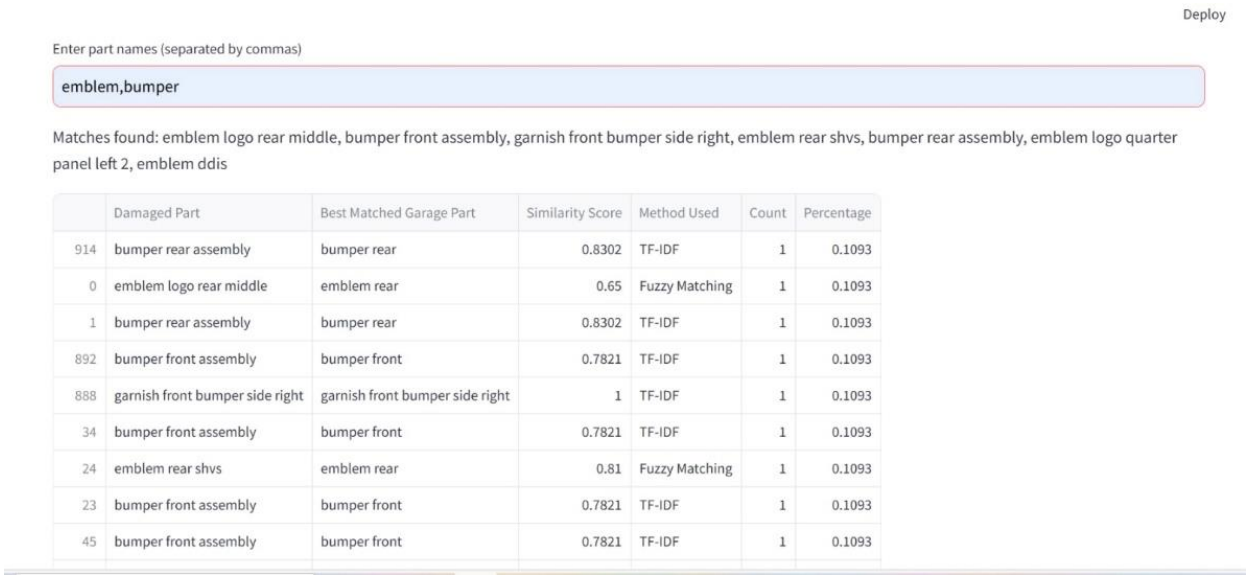


The **Damage Assessment System (DA)** interface allows users to upload a CSV file (up to 200MB) and search for damaged car parts by entering part names. It features a clean, user-friendly design with the purpose of mapping highly similar damaged parts as to maintain consistency in naming of damaged parts.



	Damaged Part	Best Matched Garage Part	Similarity Score	Method Used	Count	Percentage
891	hinge bottom door front left	hinge comp front door lower left	0.8	Fuzzy Matching	1	0.1093
889	garnish door rear right	panel assembly rear door right	0.68	Fuzzy Matching	1	0.1093
58	hinge rear door right hand	hinge front door upper right han	0.7579	TF-IDF	1	0.1093
56	weather strip door	stopper rear door front door	0.52	Fuzzy Matching	1	0.1093
83	tape front door outer rear right b	tape front door outer rear right	0.96	Fuzzy Matching	1	0.1093

The **DA System** further allows users to search for damaged car parts by entering keywords. It displays matched parts along with the best-matched garage part. It also displays the similarity score which is either calculated by the Fuzzy Matching method or the TF-IDF method. The system helps identify relevant replacements efficiently using advanced text-matching techniques.



	Damaged Part	Best Matched Garage Part	Similarity Score	Method Used	Count	Percentage
914	bumper rear assembly	bumper rear	0.8302	TF-IDF	1	0.1093
0	emblem logo rear middle	emblem rear	0.65	Fuzzy Matching	1	0.1093
1	bumper rear assembly	bumper rear	0.8302	TF-IDF	1	0.1093
892	bumper front assembly	bumper front	0.7821	TF-IDF	1	0.1093
888	garnish front bumper side right	garnish front bumper side right	1	TF-IDF	1	0.1093
34	bumper front assembly	bumper front	0.7821	TF-IDF	1	0.1093
24	emblem rear shvs	emblem rear	0.81	Fuzzy Matching	1	0.1093
23	bumper front assembly	bumper front	0.7821	TF-IDF	1	0.1093
45	bumper front assembly	bumper front	0.7821	TF-IDF	1	0.1093

The **DA System** allows users to input multiple part names at once, unlike the previous snippet where only a single input was used. This feature enhances efficiency by retrieving matches for multiple damaged parts simultaneously. As before the system uses **TF-IDF** and **Fuzzy Matching** and identifies similar parts, helping streamline damage assessment and replacement.

6 Conclusion

The implementation of an automated part mapping and recommendation system for AutoSure Insurance significantly enhances the accuracy and efficiency of damage assessment and claim verification. By leveraging TF-IDF vectorization, cosine similarity, and fuzzy string matching, the proposed approach effectively standardizes part descriptions, mitigates inconsistencies, and streamlines the claim processing workflow.

Key benefits of this system include:

- Reduced Claim Processing Time: Automation minimizes manual effort in matching part descriptions, leading to faster claim settlements.
- Improved Accuracy in Damage Assessment: Hybrid matching techniques ensure better alignment between surveyor and garage datasets, reducing mismatches.
- Fraud Prevention: Standardized part identification helps in detecting discrepancies and potential fraudulent claims.
- Enhanced Customer Satisfaction: Faster and more accurate claims processing strengthens AutoSure Insurance's reputation and customer trust.
- Scalability and Integration: The model can be seamlessly integrated into AutoSure Insurance's existing systems, ensuring minimal operational disruptions.

Overall, this data-driven solution provides a competitive edge in the motor insurance sector by optimizing claims management, improving financial sustainability, and enhancing customer experience. Future improvements could focus on refining matching thresholds, expanding synonym databases, and incorporating machine learning techniques to further enhance system robustness.

Please scan the QR below to get access to the Python code.