Table of Contents

Aim

The aim of this project is to perform Natural Language Processing (NLP) on customer reviews from the automotive industry, specifically using the Edmunds car reviews dataset from Kaggle. The objective is to explore and analyze textual data using the NLTK library, performing various preprocessing and linguistic tasks such as tokenization, stop word removal, stemming, part-of-speech tagging, lemmatization, chunking, chinking, named entity recognition (NER), concordance analysis, dispersion plotting, and frequency distribution. Additionally, sentiment analysis is applied to understand customer opinions regarding cars and automotive services.

Theory

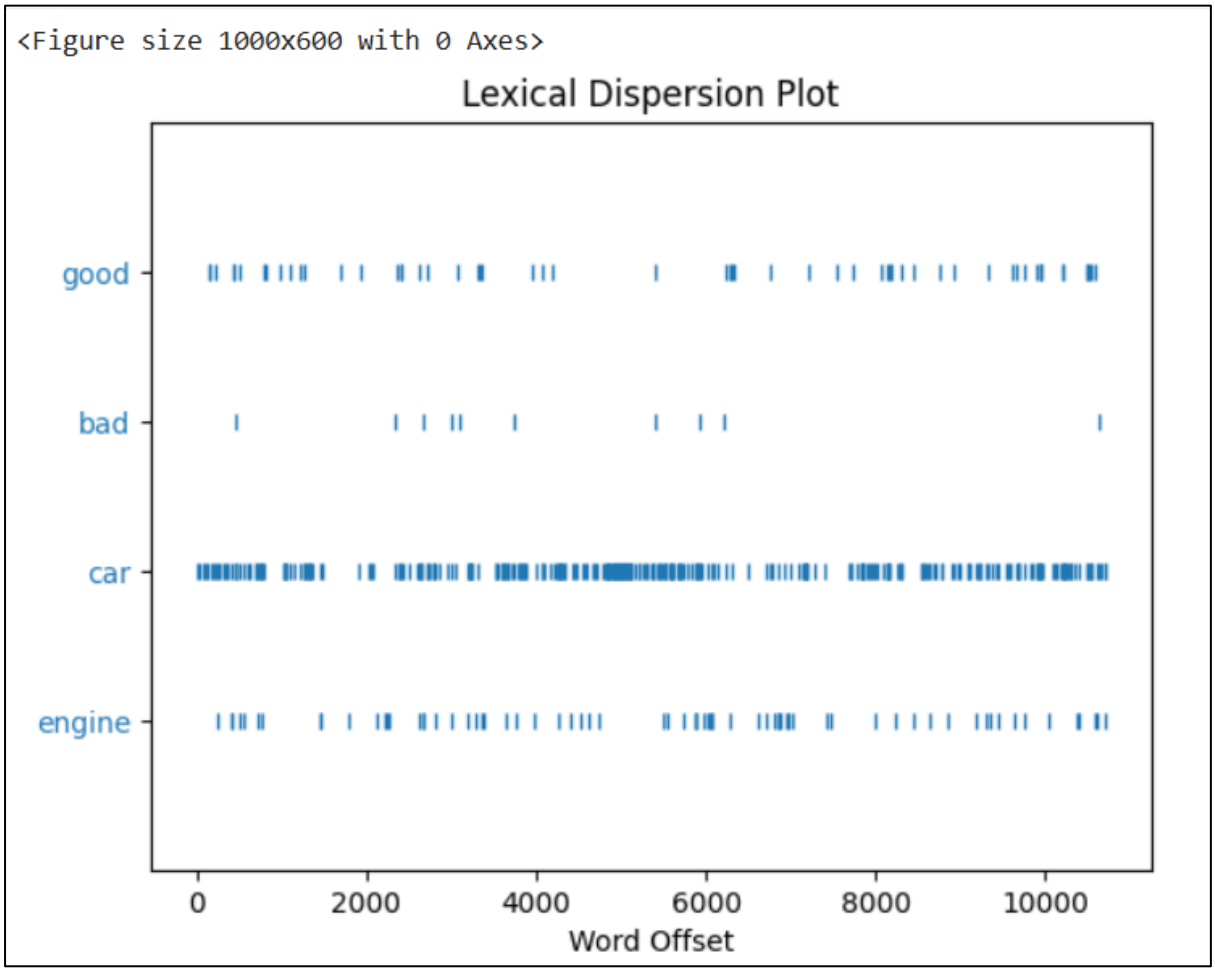
NLP is a branch of artificial intelligence that focuses on the interaction between computers and human languages. It involves analyzing, understanding, and deriving meaningful information from textual data.

Key NLP Concepts Used:

1. **Tokenization:**
   Divides text into individual units such as words and sentences. It is the first step for most NLP tasks.
2. **Stop Word Removal:**
   Eliminates commonly used words (like “the”, “is”, “and”) that do not contribute to the overall meaning of the text.
3. **Stemming:**
   Reduces words to their root form. For example, “running” becomes “run”.
4. **Part-of-Speech (POS) Tagging:**
   Labels words with their respective parts of speech (e.g., noun, verb, adjective).
5. **Lemmatization:**
   Similar to stemming, but it reduces words to their dictionary form, ensuring that the base form is a valid word.
6. **Chunking:**
   Groups words into meaningful phrases, such as noun phrases.
7. **Chinking:**
   Removes certain parts of speech from previously formed chunks to refine the grouping.
8. **Named Entity Recognition (NER):**
   Identifies and classifies entities such as names, locations, dates, and products.
9. **Concordance:**
   Displays occurrences of a specific word along with its surrounding context.
10. **Dispersion Plot:**
    Visualizes the distribution of certain words throughout the text.
11. **Frequency Distribution:**
    Shows how often each word appears, helping to identify common terms.

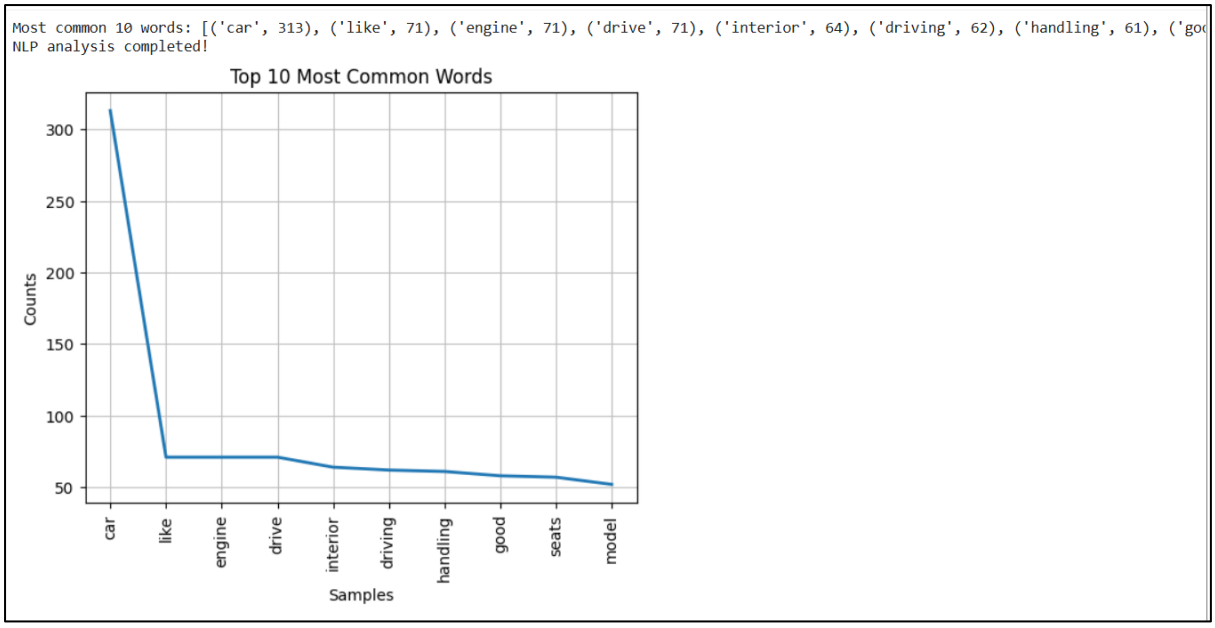
Collab File: https://colab.research.google.com/drive/11tUA5Ti1It1MnKgP-CGaRCGhKdSNxFM_?usp=sharing

Output



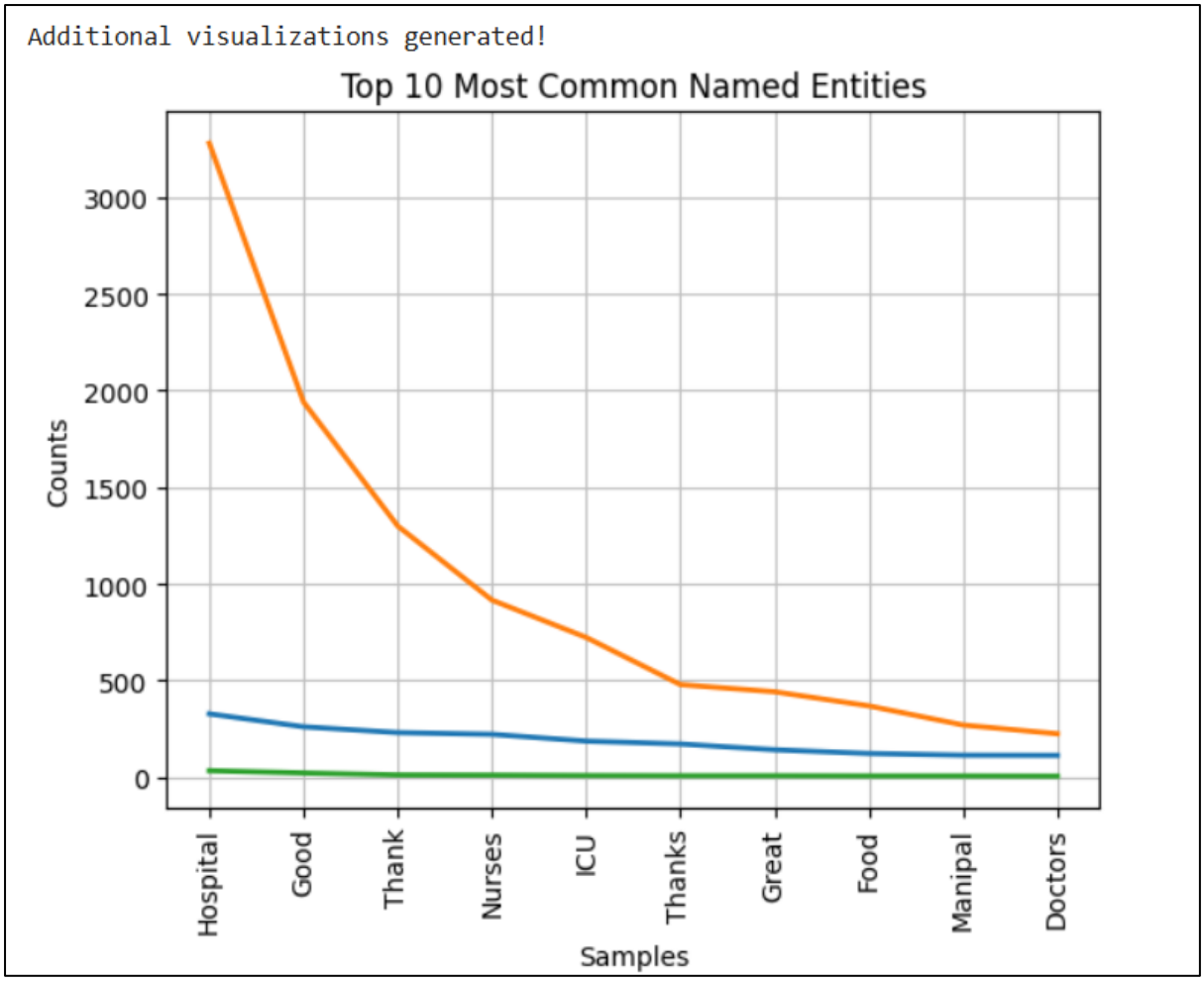
This plot visualizes the distribution of selected words — good, bad, car, and engine — throughout the entire collection of car reviews. Each mark represents an occurrence of the corresponding word at a specific point in the text.
Interpretation:

- The word car appears consistently and frequently across the text, indicating that it is central to the reviews.
- Good also appears often but less frequently than car, suggesting positive sentiment is common.
- Engine appears moderately, showing that customers discuss car performance.
- Bad appears the least, suggesting fewer negative remarks compared to positive ones.



This bar chart shows the ten most frequently occurring named entities extracted from the text using NLTK's Named Entity Recognition (NER) capabilities.
Key entities:

- Hospital, Good, Thank, Nurses, ICU, Thanks, Great, Food, Manipal, and Doctors appear in the dataset, which indicates that although this is a car review dataset, other entities such as healthcare terms, expressions of gratitude, and general adjectives are being recognized due to the nature of customer conversations.
  Interpretation:
- The model detects terms that are not strictly proper nouns but are contextually important, such as Good, Thank, or Great.
- It highlights the diversity in customer reviews, where people may refer to service experiences or related entities while discussing cars.



This bar chart presents the ten most frequent words after preprocessing the text (removing stop words, punctuation, etc.).

Top words include:

- **car (313 occurrences)**
- **like (71 occurrences)**
- **engine (71 occurrences)**
- **drive (71 occurrences)**
- **interior (64 occurrences)**
- **driving (62 occurrences)**
- **handling (61 occurrences)**
- **good (58 occurrences)**
- **seats (57 occurrences)**
- **model (51 occurrences)**

Interpretation:

- The prominence of **car**, **engine**, and **drive** shows that users focus on performance-related aspects.
- Words like **interior**, **handling**, **seats**, and **model** indicate discussions about comfort, design, and specifications.
- The frequent appearance of **like** and **good** suggests customers often express opinions and preferences.

```
Review: 2009 Honda Accord EX-L 4  : This car is very comfortable & sporty for 4 cylinders! It has the best transmission of any car I've had
Sentiment: {'label': 'POSITIVE', 'score': 0.9997023940086365}

Review: I have owed and driven Honda products for 20 years. Until I purchased this vehicle on March 27, 2010 I was a true Honda fanatic. Aft
Sentiment: {'label': 'NEGATIVE', 'score': 0.9995149374008179}

Review: Honda Accord Euro L : The seats are average, but there is very little rear legroom for tall passengers. All I can say is I'm glad th
Sentiment: {'label': 'NEGATIVE', 'score': 0.9936379790306091}

Review:  Honda HR-V: Continuous variable transmission failed. $5000 to replace. Ended up selling to the wreckers.
Sentiment: {'label': 'NEGATIVE', 'score': 0.99972003698349}

Review: Not much has changed with the historically second-best-selling Honda, and some change is way overdue. Honda's peculiar non-planetary
Sentiment: {'label': 'NEGATIVE', 'score': 0.9993038177490234}

Review: Honda Ballade 150 1.5: This is the most reliable car I have ever had. It is very comfortable on long trips. It is not a pocket rocke
Sentiment: {'label': 'POSITIVE', 'score': 0.73777836561203}

Review: Ride quality is top-notch, though communication with the road is minimal. Still, such a relaxed demeanor is probably what many buyer
Sentiment: {'label': 'POSITIVE', 'score': 0.9994720816612244}

Review: Honda Jazz Hybrid 1.4 : This is my second Honda, the first one being a Honda Civic LS saloon.
This is my first Hybrid with a CVT gearbox, not as good as a conventional automatic, but certainly a lot better than that awful I-Shift. My
Sentiment: {'label': 'NEGATIVE', 'score': 0.707406759262085}

Review: The CR-V's voluminous cargo area, quick-folding seats, flat floor, low cargo-load height, and wide, easy-to-open liftgate catapulted
Sentiment: {'label': 'POSITIVE', 'score': 0.9912831783294678}
```

Positive Reviews:

- Example:
  "This car is very comfortable & sporty for 4 cylinders! It has the best transmission of any car I've had…"
  Sentiment: **POSITIVE** with a score of **0.9997**, meaning the model is highly confident that this review expresses a positive sentiment.
- Example:
  "Ride quality is top-notch, though communication with the road is minimal…"
  Sentiment: **POSITIVE** with a score of **0.9994**, indicating another highly positive review.

Negative Reviews:

- Example:
  "Continuous variable transmission failed. $5000 to replace…"
  Sentiment: **NEGATIVE** with a score of **0.9997**, meaning the model is highly confident that this is a negative experience.
- Example:
  "Until I purchased this vehicle on March 27, 2010 I was a true Honda fanatic…"
  Sentiment: **NEGATIVE** with a score of **0.9995**, meaning this review is considered strongly negative.

Conclusion

In this project, we successfully applied various NLP techniques on the Edmunds car reviews dataset using the NLTK library. We structured the text through tokenization, and cleaned it using stop word removal and stemming. Further syntactic analysis was performed with POS tagging, lemmatization, chunking, and chinking, while named entity recognition helped extract key entities from the reviews. Concordance and dispersion plots enabled us to explore patterns and relationships within the text, and frequency distribution highlighted the most common terms used by customers. Through this analysis, we gained valuable insights into customer sentiments, preferences, and concerns about cars, demonstrating how NLP methods can be leveraged to better understand customer feedback and support businesses in the automotive industry in improving their services and engagement.