INTRODUCTION

INTRODUCTION

Language is the basic way of communication that is required to exchange information between people. Each person acquires an ability, right from their childhood, to make use of this exchange of information either through sounds or gestures. As we progress, we tend to learn various languages to communicate among others. Language is not only limited to communicate thoughts and ideas, but it also builds friendship, career, strengthens the economic relationships of businesses and provides an understanding of diverse cultures scattered across the globe. Thousands of languages exist around the world which are used in everyday life. Most of the time, during visit to other places, a person would not be able to understand or communicate in their native language. So, here comes the role of Language Identification which determines the language of the written content given to it and help the user to communicate effectively.

Today, the world is moving towards Artificial Intelligence in every aspect of life. A generation of computers that are capable of depicting the thoughts and actions of human. One of the wonders of the world is the human visual system, which we are attempting to imitate. Any activity that a system undertakes must first be known to it in order for it to begin. Humans have ears, eyes, and a brain that they use to receive information and act on it. In a similar way, speech recognition and word recognition on their own would serve as a computer's eyes and ears, respectively. The most recent technology is used to recognize sign language and other languages.

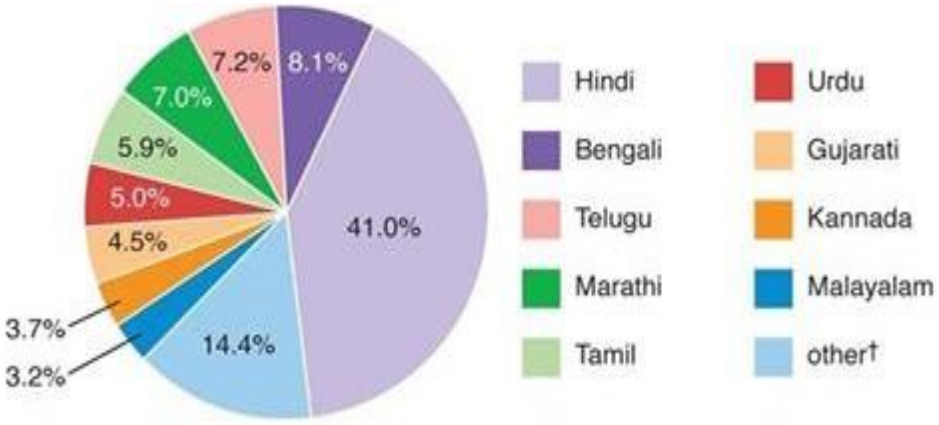
Language, as a medium of communication, stands as a cornerstone of human interaction, facilitating the exchange of ideas, cultures, and emotions across diverse communities worldwide. With thousands of languages spoken and written globally, each reflecting unique linguistic structures, vocabularies, and cultural nuances, the complexity of language diversity presents both a challenge and an opportunity in today's interconnected world. In this digital age characterized by rapid globalization and technological advancements, the need for automated language understanding and identification has become increasingly paramount. Language detection, a fundamental task within the realm of Natural Language Processing (NLP), addresses this need by enabling machines to automatically determine the language of a given text or document. This project embarks on a journey to explore and develop a sophisticated language detection system, leveraging state-of-the-art machine learning algorithms, advanced NLP techniques, and a rich dataset of text samples spanning a multitude of languages.

The primary aim of this project is to construct a robust and accurate language detection system capable of effectively identifying the language of diverse text samples with precision and reliability. At its core, the system harnesses the power of machine learning, specifically employing the Multinomial Naïve Bayes classifier—a probabilistic model renowned for its simplicity, scalability, and efficiency in text classification tasks. By leveraging a large and diverse dataset comprising labeled text samples, the system endeavors to learn and generalize language patterns and characteristics inherent to each linguistic domain. Through iterative training and optimization processes, the classifier evolves to become adept at distinguishing between languages based on the distribution of words, linguistic features, and structural elements unique to each language.

The journey towards developing a robust language detection system commences with data acquisition and preprocessing—a crucial phase aimed at sourcing, cleaning, and organizing a diverse corpus of text data. The dataset undergoes rigorous preprocessing steps to ensure data quality, consistency, and compatibility with machine learning algorithms. Text data is tokenized, normalized, and transformed into a numerical representation suitable for consumption by the classifier, often utilizing techniques such as tokenization, stemming, and vectorization. Furthermore, feature engineering may be employed to extract and select pertinent linguistic features from the text data, enhancing the model's discriminative power and generalization capabilities across languages.

Once the dataset is suitably prepared, the system proceeds to the model training phase, where the Multinomial Naïve Bayes classifier is trained on the preprocessed text data. During training, the classifier learns to estimate the probability distribution of observing a particular set of words given the language class, effectively modeling the linguistic characteristics and patterns specific to each language. Through the process of parameter estimation and model optimization, the classifier refines its probabilistic model, striving to maximize classification accuracy while minimizing errors and uncertainties. The training process is iterative, involving cross-validation techniques and hyperparameter tuning to ensure robustness and generalization across diverse language domains.

Following model training, the system undergoes comprehensive evaluation to assess its performance, reliability, and generalization capabilities across a wide range of languages. Various evaluation metrics, including accuracy, precision, recall, and F1-score, are computed to quantitatively measure the system's performance on both training and test datasets.



Additionally, visualization techniques such as confusion matrices and ROC curves may be employed to gain insights into the system's strengths and weaknesses, identifying potential areas for improvement and refinement. Through rigorous evaluation and analysis, the system strives to achieve high accuracy and reliability in language detection, paving the way for practical applications in real-world scenarios.

Beyond model evaluation, the project explores practical applications and user interactions, aiming to develop a user-friendly interface that allows users to interactively input text and receive real-time language predictions. This interactive feature enhances usability and accessibility, enabling users from diverse backgrounds to leverage the language detection system for various applications, including multilingual content management, language identification in social media platforms, and cross-lingual information retrieval. By bridging the gap between machine and human communication, the system seeks to empower users with the ability to navigate and interact with multilingual content seamlessly, fostering cross-cultural communication and understanding in the digital age.Looking ahead, the project envisions a future of continuous innovation and refinement, with ongoing efforts focused on expanding language coverage, enhancing model performance, and exploring emerging technologies in the field of NLP. By embracing advancements in deep learning, neural networks, and transfer learning, the project aims to push the boundaries of language detection, enabling the system to tackle complex language identification tasks with unprecedented accuracy and efficiency. Additionally, efforts will be directed towards integrating the language detection system into web and mobile applications, IoT devices, and virtual assistants, thereby extending its reach and impact to a broader audience worldwide.

In conclusion, this project represents a journey of exploration, innovation, and collaboration in the realm of language detection and understanding. By harnessing the power of machine learning, advanced NLP techniques, and a diverse dataset of text samples, the project endeavors to develop a robust and versatile language detection system capable of bridging linguistic barriers and fostering cross-cultural communication in the digital age. As language continues to evolve and adapt to the changing landscape of human interaction, the significance of automated language understanding and identification cannot be overstated, making this project a testament to the transformative potential of technology in shaping the future of communication and connectivity.

METHODS OF DATA ANALYSIS

In a language detection project, the methods of data analysis play a crucial role in training, evaluating, and fine-tuning the language detection models. Here are some key methods of data analysis commonly used in such projects:

1. Data Preprocessing:

- Removing any irrelevant characters, symbols, or punctuation marks from the text data.
- Breaking down the text data into smaller units such as words, phrases, or characters.
- Converting text to lowercase, removing accents, or expanding contractions to standardize the text format.
- Eliminating common words (e.g., "and", "the") that do not carry significant meaning for language detection.
- Reducing words to their root form to handle variations (e.g., "running" to "run").

2. Feature Extraction:

- Representing text data as a matrix of word frequencies, where each row corresponds to a document and each column corresponds to a word in the vocabulary.
- Calculating the importance of a word in a document relative to its frequency across all documents.
- Extracting sequences of adjacent words or characters as features to capture local context and dependencies.
- Representing text data at the character level to capture morphological and phonetic information.

3. Model Selection:

- Multinomial Naïve Bayes, Gaussian Naïve Bayes, or Bernoulli Naïve Bayes classifiers are commonly used for text classification tasks due to their simplicity and efficiency.
- Linear or kernel-based SVM classifiers can learn decision boundaries to separate different language classes in feature space.
- Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), or Transformer models may be employed for language detection, leveraging their ability to capture complex patterns and dependencies in text data.

4. Model Training and Evaluation:

- Splitting the dataset into training and validation sets using techniques such as k-fold cross-validation to assess the model's performance.
- Optimizing model parameters (e.g., regularization strength, learning rate) using techniques such as grid search or random search to improve performance.
- Using metrics such as accuracy, precision, recall, F1-score, and confusion matrices to evaluate the model's performance on both training and validation sets.
- Plotting learning curves to visualize the model's training and validation performance over different iterations or epochs.

5. Error Analysis and Model Interpretability:

- Identifying and analyzing misclassified samples to understand common sources of errors and areas for improvement.
- Examining the importance of features (e.g., words, n-grams) in the model's decision-making process to gain insights into language-specific patterns and characteristics.
- Visualizing decision boundaries, feature importance, or embeddings (e.g., t-SNE) to interpret the model's behavior and predictions.

6. Post-processing and Deployment:

- Setting decision thresholds or confidence levels to control the trade-off between precision and recall in language detection.
- Integrating the trained language detection model into production systems or applications through APIs, web services, or containerization for real-time inference on new text data.
- By employing these methods of data analysis, language detection projects can develop accurate, efficient, and interpretable models for identifying the language of text data across various applications and domains.

FEASIBILITY STUDY

FEASIBILITY STUDY

A feasibility study is a crucial step in the initiation phase of any project, including a language detection project. It involves evaluating the practicality, viability, and potential success of the project before committing resources and efforts to its implementation. In the context of a language detection project, a feasibility study assesses various factors such as technical, economic, operational, and legal considerations to determine whether the project is feasible and worth pursuing. Let's explore each aspect of the feasibility study in more detail:

1. Technical Feasibility:

Hardware and Software infrastructure:

- Assess the hardware and software infrastructure needed to develop and deploy the language detection system.
- Evaluate the availability of computing resources, such as servers, GPUs, and memory, to support model training and inference.
- Ensure compatibility with existing software frameworks, libraries, and tools commonly used in natural language processing and machine learning.

Data Availability and Quality:

- Determine the availability of labeled text data in multiple languages for training and evaluation.
- Evaluate the quality, quantity, and diversity of the data to ensure representative coverage of language variations, dialects, and linguistic characteristics.
- Identify potential sources for acquiring additional data and establish data acquisition pipelines to maintain a continuous supply of relevant data.

Algorithm Selection and Development:

- Research and evaluate different machine learning algorithms and techniques suitable for language detection, considering factors such as accuracy, scalability, and computational efficiency.
- Develop prototypes or proof-of-concept models to assess the feasibility and performance of selected algorithms on the available data.
- Iterate on model development and experimentation to refine and optimize the language detection system for real-world applications.

2. Economic Feasibility:

Cost-Benefit Analysis:

- Estimate the initial and ongoing costs associated with developing, deploying, and maintaining the language detection system.
- Identify potential cost savings, efficiencies, or revenue opportunities that the project may generate over its lifecycle.
- Conduct a cost-benefit analysis to compare the projected benefits of the project against its anticipated costs and assess its overall economic viability.

Return on Investment (ROI):

- Calculate the expected ROI of the language detection project based on projected revenue generation, cost savings, or other tangible benefits.
- Evaluate the payback period and breakeven point to determine the time it will take for the project to recoup its initial investment and start generating positive returns.
- Consider alternative investment opportunities and potential risks to assess the relative attractiveness of the language detection project from an economic standpoint.

3. Operational Feasibility:

Resource Availability and Expertise:

- Assess the availability of skilled personnel with expertise in natural language processing, machine learning, and software development.
- Identify any gaps in knowledge or skills that may need to be addressed through training, hiring, or collaboration with external partners.
- Ensure adequate support and collaboration from stakeholders, domain experts, and end users throughout the project lifecycle.

Integration with Existing Systems:

- Evaluate the compatibility and integration requirements of the language detection system with existing software, data pipelines, and business processes.
- Identify potential dependencies, constraints, or conflicts that may arise during system integration and deployment.
- Develop strategies for seamless integration and interoperability with other systems and applications to maximize the project's operational impact.

4. Legal and Regulatory Feasibility:

Data Privacy and Security:

- Ensure compliance with relevant data privacy regulations, such as GDPR, CCPA, and HIPAA, when collecting, storing, and processing personal or sensitive data.
- Implement appropriate data protection measures, encryption techniques, and access controls to safeguard against unauthorized access, data breaches, or privacy violations.
- Establish data governance policies and procedures to ensure transparency, accountability, and ethical use of data throughout the project lifecycle.

Intellectual Property Rights:

- Assess the intellectual property rights associated with the language detection system, including patents, copyrights, and trademarks.
- Identify any potential legal risks or infringement issues related to the use of third-party libraries, datasets, or proprietary algorithms.
- Secure necessary licenses, permissions, or agreements to mitigate legal liabilities and protect the project's intellectual property assets.

Conclusion:

In conclusion, a thorough feasibility study is essential for evaluating the viability and potential success of a language detection project. By assessing technical, economic, operational, and legal considerations, stakeholders can make informed decisions about whether to proceed with the project and allocate resources effectively. A well-executed feasibility study lays the foundation for a successful project implementation, ensuring that the language detection system meets the needs of its users, delivers val

REQUIREMENT SPECIFICATIONS

REQUIREMENT SPECIFICATIONS

The language detection project aims to develop a robust and efficient system capable of automatically identifying the language of text data. The system will be designed to process text from various sources and accurately determine the language of each input text sample. This document outlines the detailed functional and non-functional requirements of the language detection system to guide its design, development, and implementation.

HARDWARE REQUIREMENTS:

1. Computing Resources:

Processor:

- A multicore processor with support for parallel computing is recommended to accelerate model training and inference tasks.

Memory (RAM):

- A minimum of 8GB RAM is recommended for handling large datasets and running memory-intensive machine learning algorithms.

Storage:

- Sufficient storage space for storing datasets, model checkpoints, and other project-related files. SSD storage is preferred for faster read/write speeds.

2. Graphics Processing Unit (GPU) (Optional):

NVIDIA GPU:

- For accelerated model training and inference, an NVIDIA GPU with CUDA support is recommended. GPUs such as NVIDIA GeForce GTX or RTX series or NVIDIA Tesla series are suitable for deep learning tasks.

SOFTWARE REQUIREMENTS:

1. Operating System:

Linux-based OS:

- Ubuntu, CentOS, or Debian are commonly used operating systems for machine learning and data science projects due to their stability, compatibility, and availability of software packages.
- Windows and macOS can also be used, but Linux is preferred for its better support for deep learning frameworks and libraries.

2. Python and Libraries:

Python 3:

- Python 3.x is the programming language used for developing the language detection system.

Anaconda Distribution:

- Anaconda provides a comprehensive Python distribution with popular data science libraries pre-installed, including NumPy, pandas, scikit-learn, and TensorFlow.

Libraries:

- Install necessary Python libraries such as NumPy, pandas, scikit-learn, TensorFlow, Keras, PyTorch, NLTK, spaCy, and Matplotlib for data preprocessing, model development, and visualization tasks.

3. Development Tools:

Integrated Development Environment (IDE):

- Use an IDE such as PyCharm, Jupyter Notebook, or VS Code for code development, debugging, and experimentation.

Version Control:

- Utilize version control systems like Git for managing code repositories, tracking changes, and collaborating with team members.

4. Machine Learning Frameworks:

TensorFlow:

- TensorFlow is a popular deep learning framework used for building and training neural network models.

scikit-learn:

- scikit-learn provides a wide range of machine learning algorithms and tools for classification, regression, clustering, and model evaluation.

PyTorch:

- PyTorch is another deep learning framework known for its dynamic computational graph and ease of use.

5. Other Tools and Utilities:

Text Editors:

- Use text editors such as Vim or Nano for editing configuration files, scripts, and documentation.

Terminal Emulator:

- Access the command-line interface for running commands, executing scripts, and managing project directories.

FUNCTIONAL REQUIREMENTS:

1. Data Acquisition and Preprocessing:

Data Collection:

- The system should be able to collect text data from diverse sources, including web pages, social media platforms, news articles, and public repositories.

Data Preprocessing:

- Text data should undergo preprocessing steps such as cleaning, tokenization, normalization, and stopword removal to prepare it for feature extraction and model training.

2. Model Development and Training:

Algorithm Selection:

- Choose appropriate machine learning algorithms for language detection, including Naïve Bayes classifiers, Support Vector Machines (SVM), or neural networks.

Model Training:

- Train the language detection model using labeled text data, optimizing performance metrics such as accuracy, precision, recall, and F1-score.

3. Model Evaluation and Validation:

Cross-Validation:

- Evaluate the performance of the language detection model using cross-validation techniques such as k-fold cross-validation.

Evaluation Metrics:

- Compute evaluation metrics including accuracy, precision, recall, F1-score, and confusion matrices to assess model performance.

4. Integration and Deployment:

API Development:

- Develop APIs or web services for integrating the language detection model into existing systems, applications, or workflows.

Deployment:

- Deploy the language detection system on appropriate infrastructure, ensuring scalability, reliability, and availability.

NON-FUNCTIONAL REQUIREMENTS:

1. Performance:

Scalability:

- The system should be able to handle large volumes of text data and scale horizontally to accommodate growing user demand.

Latency:

- Response times for language detection requests should be minimal, ensuring real-time or near-real-time performance.

2. Reliability:

Fault Tolerance:

- The system should be resilient to failures, with built-in mechanisms for error handling, recovery, and fault tolerance.

Availability:

- Ensure high availability of the language detection service, minimizing downtime and service disruptions.

3. Security:

Data Privacy:

- Implement data encryption, access controls, and privacy safeguards to protect sensitive information and ensure compliance with data privacy regulations.

Authentication and Authorization:

- Implement authentication and authorization mechanisms to restrict access to authorizedusers and prevent unauthorized use of the system.

4.Usability:

User Interface:

- Develop user-friendly interfaces for interacting with the language detection system,catering to both technical and non-technical users.

Documentation:

- Provide comprehensive documentation, tutorials, and user guides to assist users in understanding and utilizing the system effectively.

5.Maintainability:

Modularity:

- Design the system with a modular architecture, allowing for easy maintenance, updates,and enhancements.

Code Quality:

- Adhere to coding standards, best practices, and version control to ensure code maintainability and readability.

OVER VIEW OF SQL AND MACHINE LEARNING

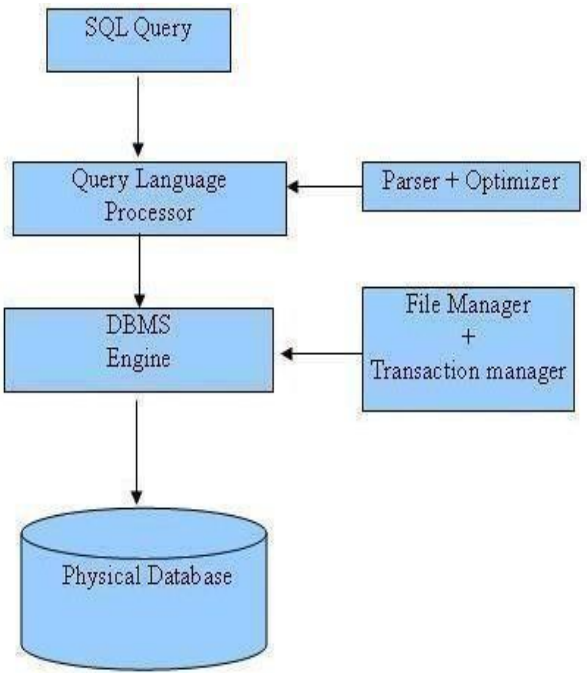
SQL(STRUCTURE QUERY LANGUAGE):

Microsoft SQL Server is a relational database management system(RDBMS)developed by Microsoft. It is widely used in the industry for storing, managing, and retrieving large amounts of structured data. SQL Server supports various enterprise-level applications and provides a range of features for data management, business intelligence ,and development.

Data Storage and Management: SQL Server allows for the storage and management of structured data in relational databases .It supports the SQL (Structured Query Language) for creating ,modifying ,and querying the data.

Development Tools and Languages: SQL Server is compatible with multiple programming languages, and Microsoft provides development tools like SQL Server Management Studio (SSMS) and SQL Server Data Tools(SSDT) for database development ,administration ,and debugging.

SQL Process:



When you are executing an SQL command for any RDBMS, the system determines the best way to carry out your request and SQL engine figures out how to interpret the task. There are various components included in the process .These components are Query Dispatcher, Optimization Engines, Classic Query Engine and SQL Query Engine ,etc. Classic query engine handles all non-SQL queries but SQL query engine won't handle logical file.

SQL Commands:

The standard SQL commands to interact with relational databases are CREATE ,SELECT ,INSERT ,UPDATE ,DELETE and DROP .These Commands can be classified into groups based on their nature:

DDL-Data Definition Language:

Command	Description
CREATE	Creates a new table, a view of a table, or other object in database
ALTER	Modifies an existing database object ,such as a table.
DROP	Deletes an entire table ,a view of a table or other object in the database.

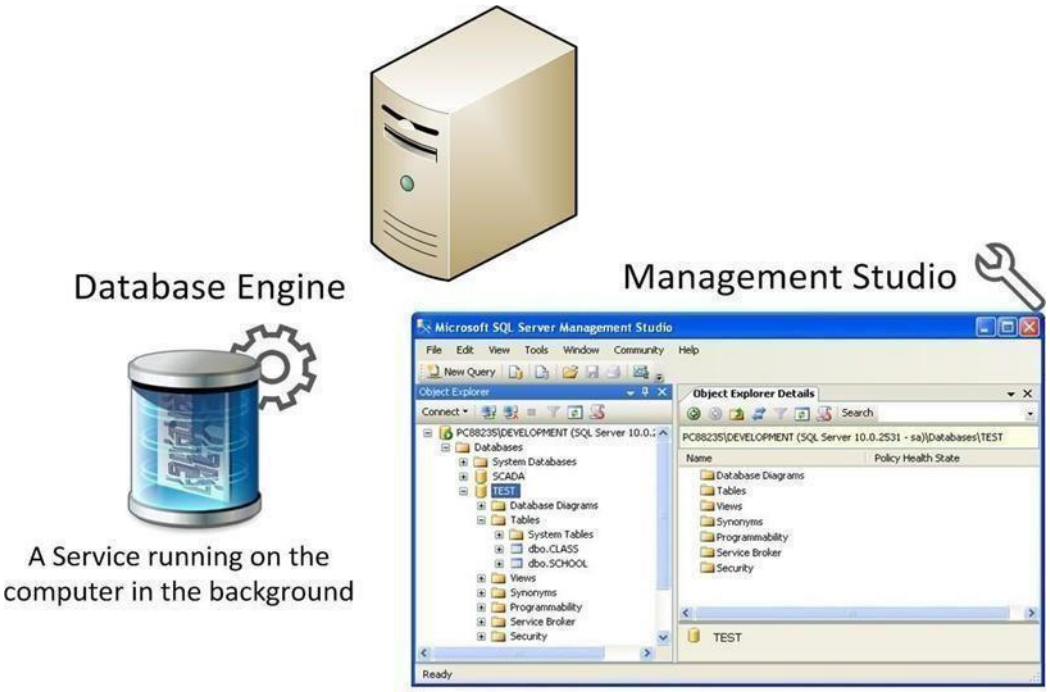
DML-Data Manipulation Language:

Command	Description
SELECT	Retrieves certain records from one or more tables
INSERT	Creates a record
UPDATE	Modifies records
DELETE	Deletes records

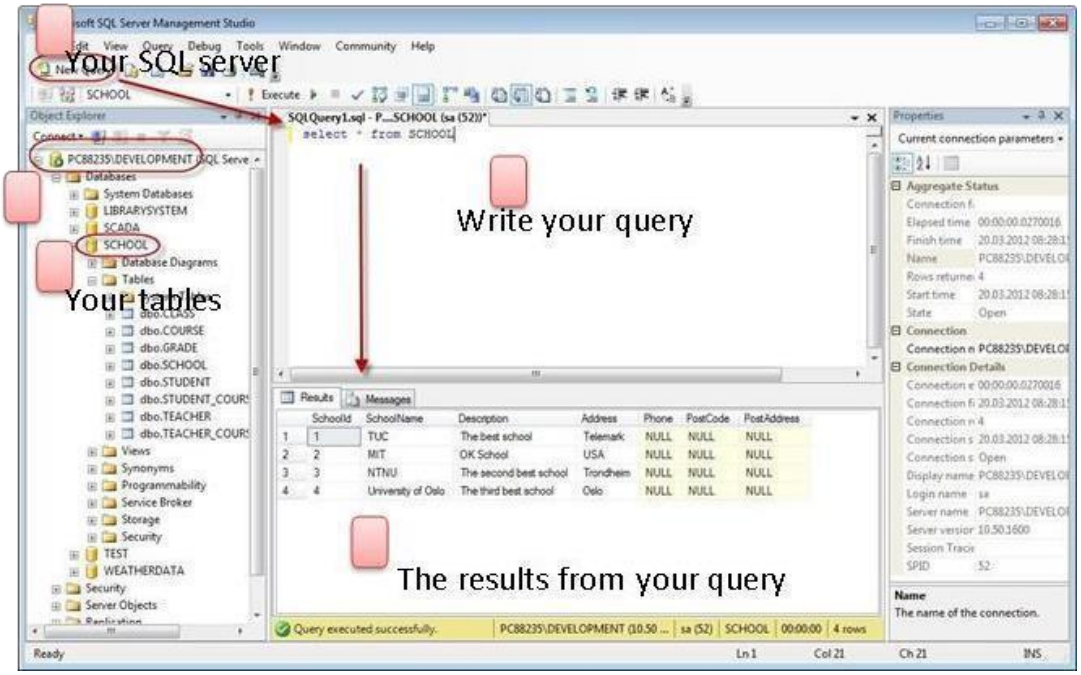
Microsoft SQL Server:

SQL Server consists of a Database Engine and a Management Studio .The Database Engine has no graphical interface - it is just a service running in the background of your computer(preferable on the server).The Management Studio is graphical tool for configuring and viewing the information in the database. It can be installed on the server or on the client(or both)

.

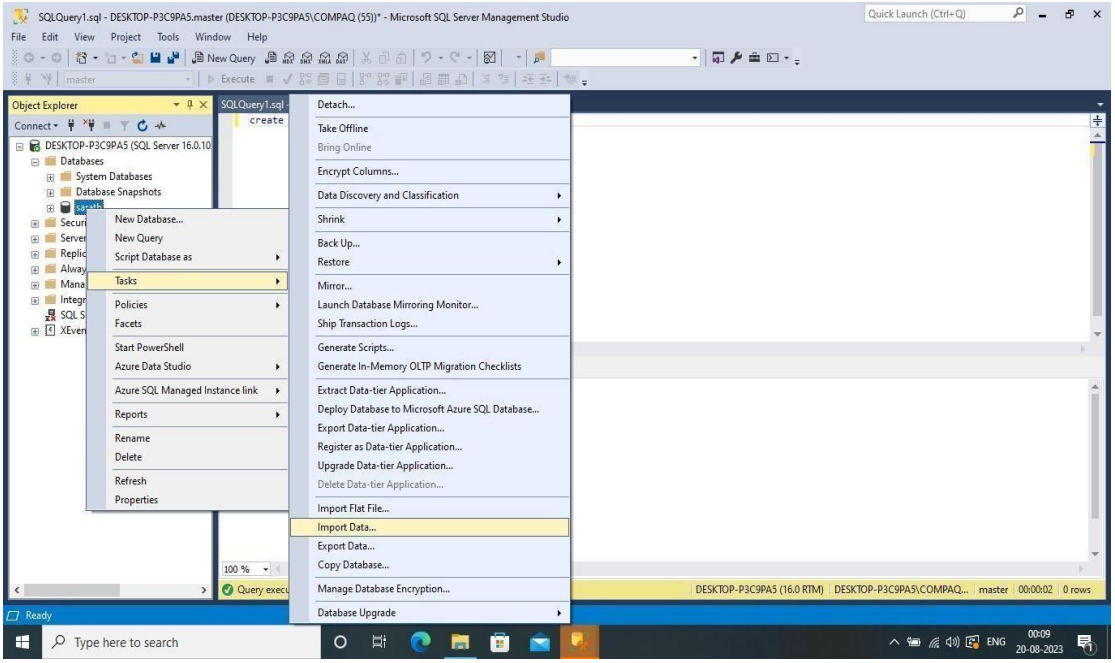


SQL Server Management Studio(SSMS):

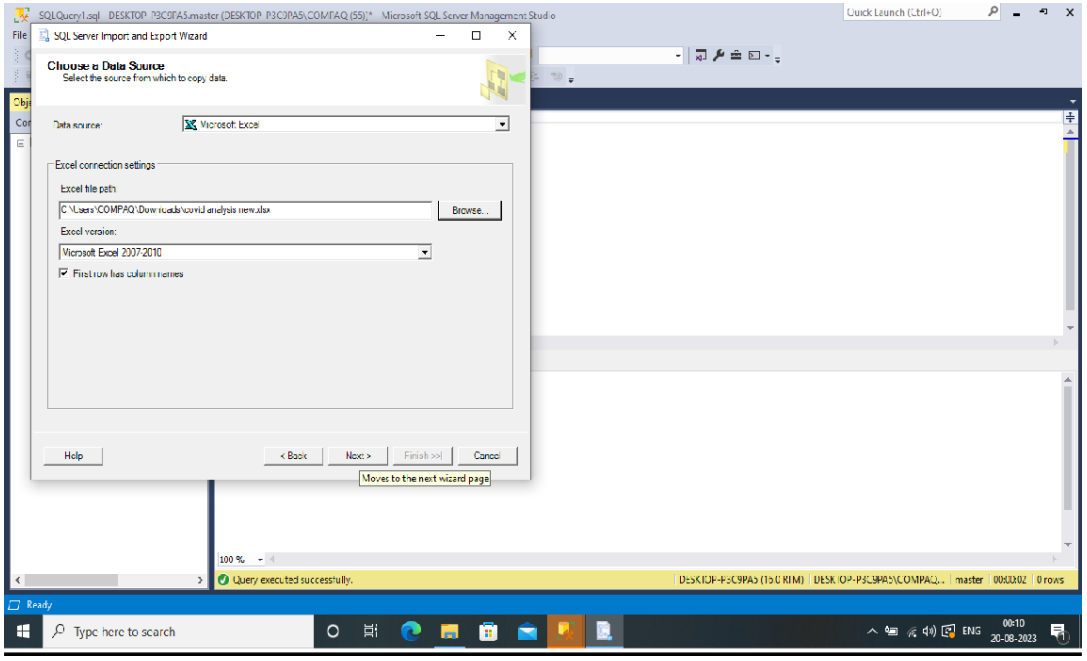


Inserting excel sheets into SQL Server:

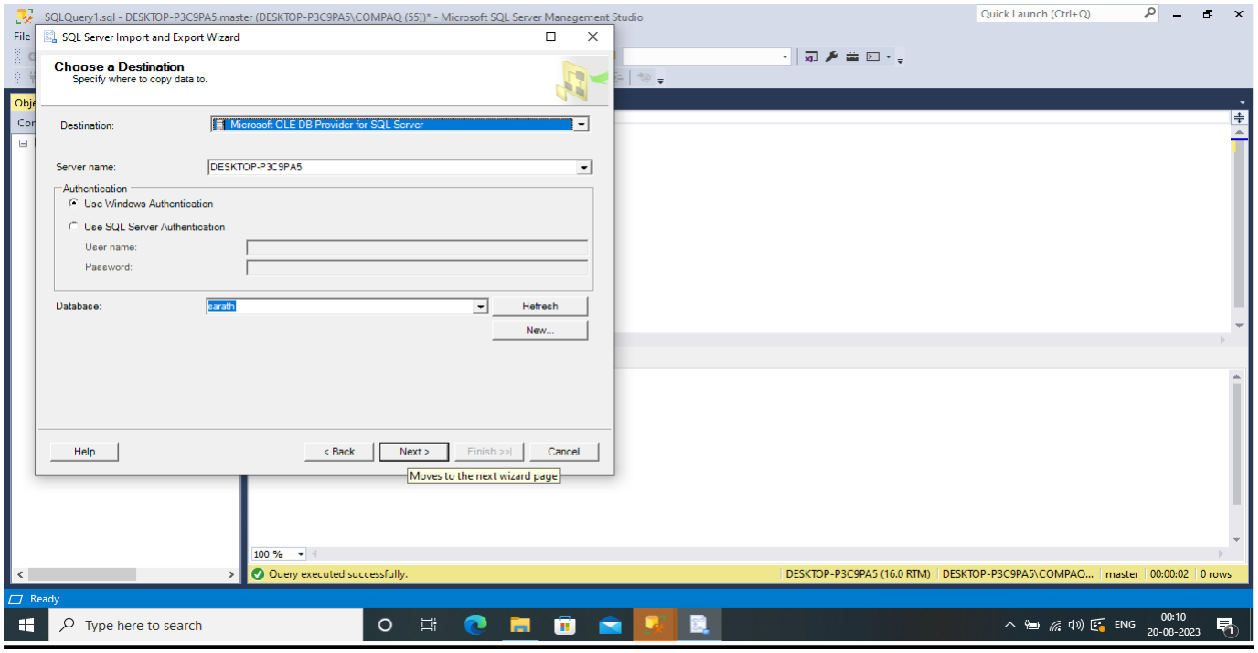
STEP-1: Click on the task button and after click on the import data



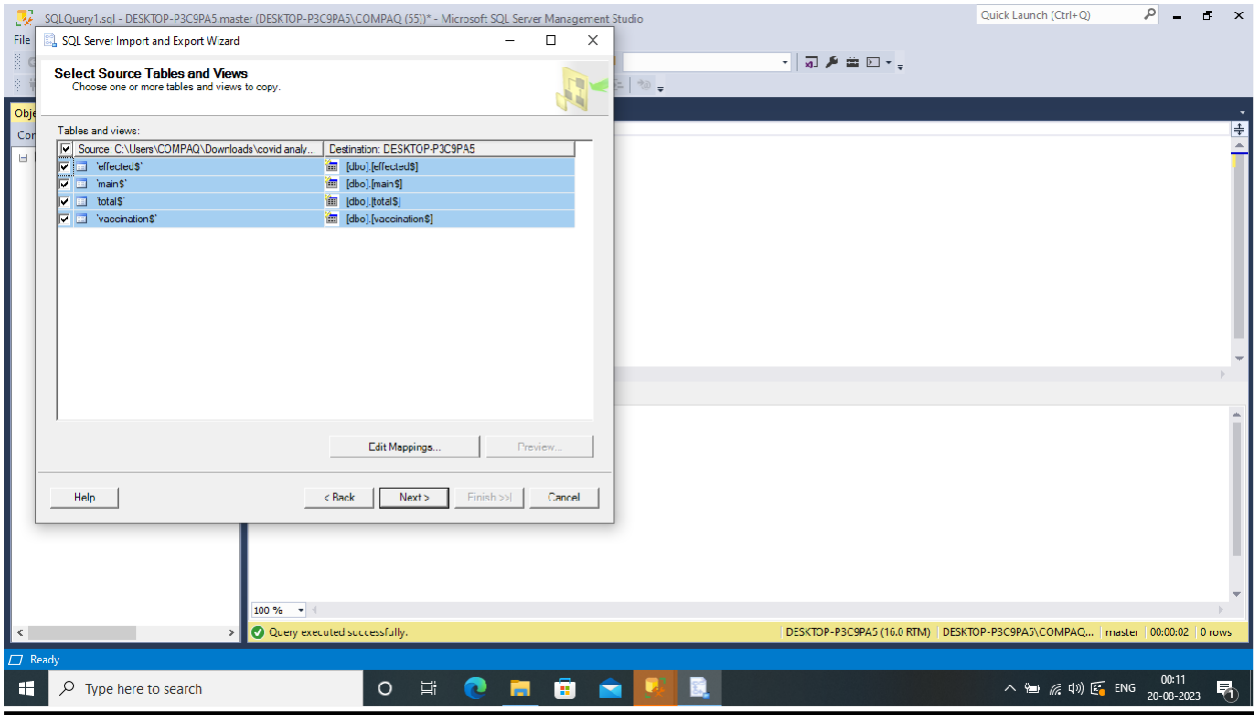
Step-2: After that a data source dialogue box is opened and select the data source path and click next



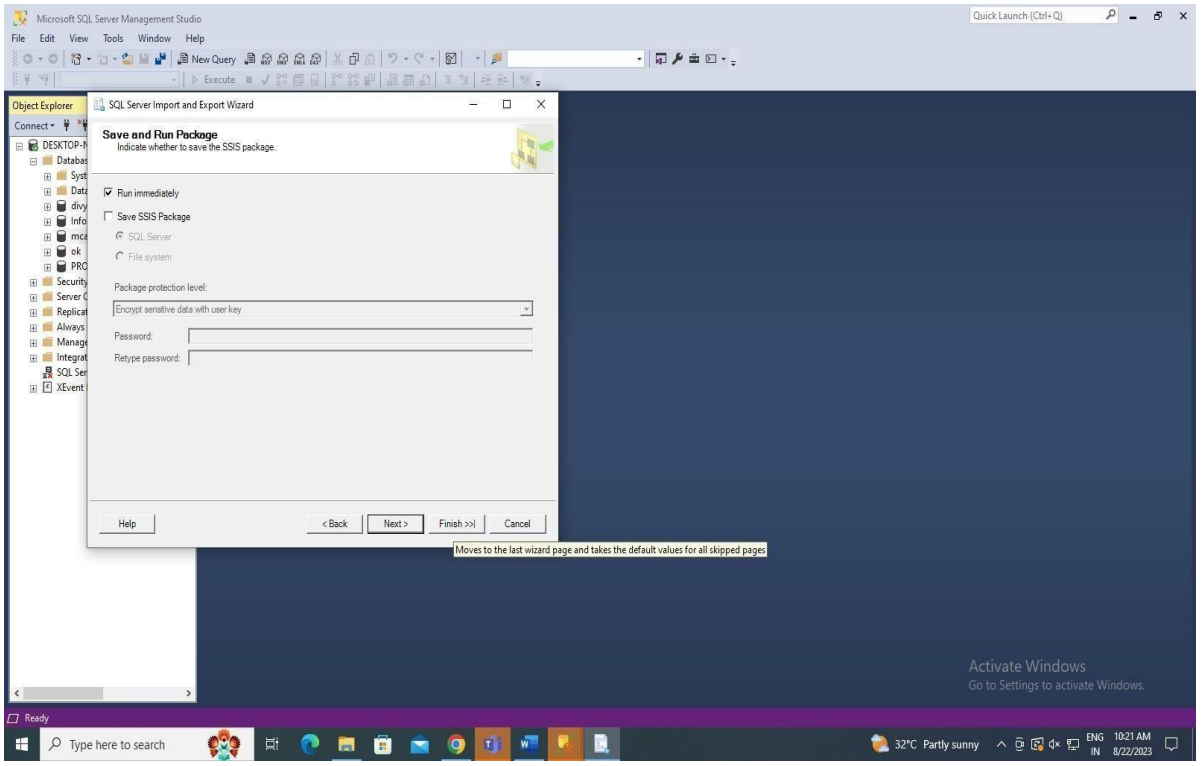
Step-3: After that a destination dialogue box is opened and select the destination path ,database name and click next



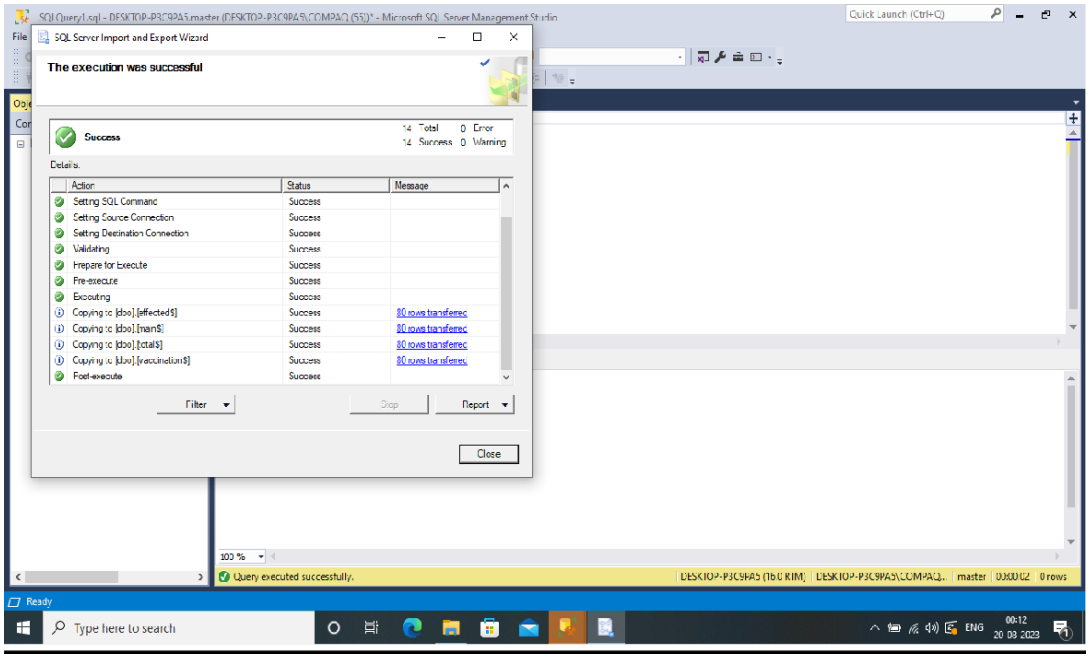
Step4:After that source table and views dialogue box is opened at that rename the sheets as table name and click



Step-5: save and run dialogue box is open click on the run immediately and click on finish



Step-6:In this step all the excel sheets are transformed into tables and stored in the named data base.



Machine Learning

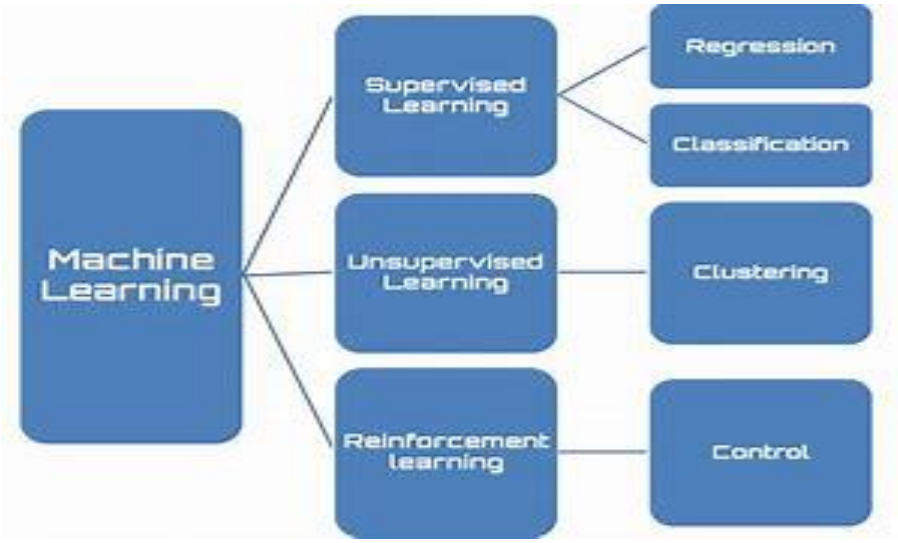
Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.
Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve.

Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes.
In this tutorial, we'll look into the common machine learning methods of supervised and unsupervised learning, and common algorithmic approaches in machine learning, including the k-nearest neighbor algorithm, decision tree learning, and deep learning. We'll explore which programming languages are most used in machine learning, providing y o u with some of the positive and negative attributes of each. Additionally, we'll discuss biases that are perpetuated by

machine learning algorithms, and consider what can be kept in mind to prevent these biases when building algorithms.



Machine Learning Methods

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data. Let's explore these methods in more detail.

Supervised Learning

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data.

For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

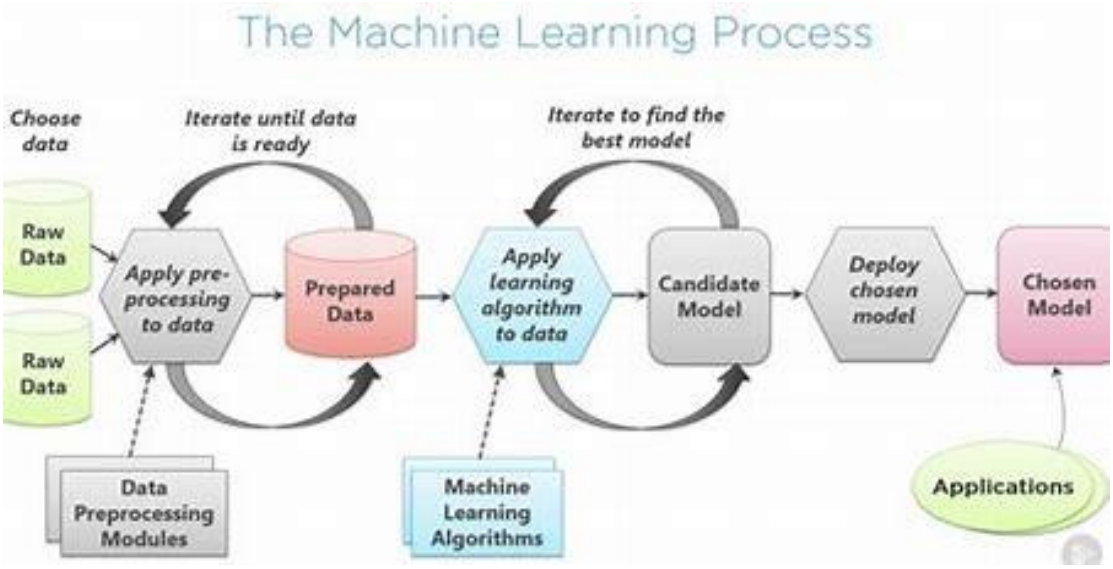
Unsupervised Learning

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

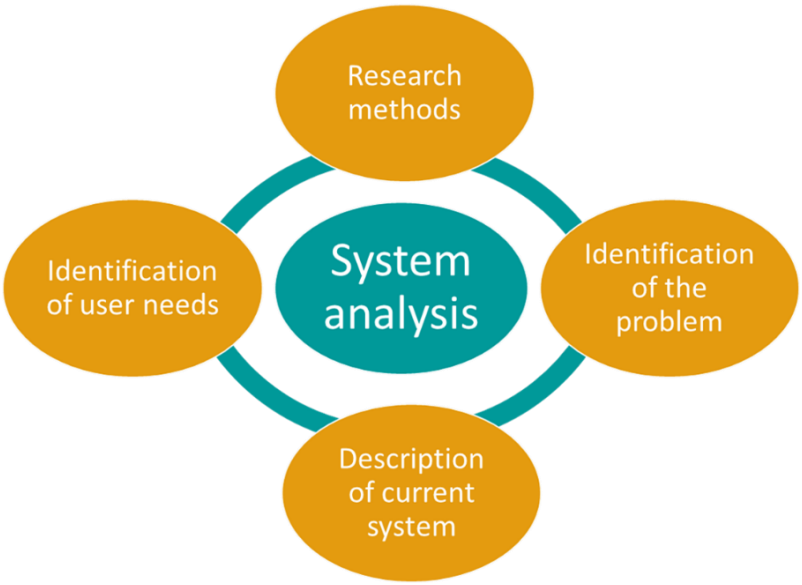
Without being told a “correct” answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.



SYSTEM ANALYSIS

SYSTEM ANALYSIS

System analysisof a language detection model involves examining its various components, interactions, and performance characteristics to ensure its effectiveness and efficiency. Here's a breakdown of the system analysis for a language detection model:



1. Overview:

System Purpose: Identify the language of input text data accurately and efficiently.

Components: Data acquisition, preprocessing, feature extraction, model development, evaluation, integration, and deployment.

2. Data Acquisition and Preprocessing:

Input: Text data from diverse sources (web pages, social media, etc.).

Acquisition: Collect text data from sources using web scraping, APIs, or datasets.

Preprocessing: Clean, tokenize, normalize, and remove noise from text data to prepare it for analysis.

3. Feature Extraction:

Techniques: Bag-of-Words (BoW), TF-IDF, N-grams, character-level features.

Representation: Convert text data into numerical features for model input.

Language-specific Patterns: Extract features to capture linguistic characteristics and variations.

4. Model Development:

Algorithms: Naïve Bayes, Support Vector Machines (SVM), neural networks.

Training: Train the model on labeled data using appropriate algorithms.

Parameter Tuning: Optimize model parameters for improved performance.

Validation: Evaluate model performance using techniques like k-fold cross-validation.

5. Evaluation and Validation:

Metrics: Accuracy, precision, recall, F1-score, confusion matrix.

Performance Analysis: Assess model performance on training and validation datasets.

Error Analysis: Identify and analyze misclassified samples to understand model weaknesses.

6. Integration and Deployment:

API Development: Develop APIs or web services for integrating the model into applications.

Deployment: Deploy the model on scalable infrastructure for real-time inference.

Scalability: Ensure the system can handle increasing volumes of text data and user requests.

7. Performance and Optimization:

Latency: Minimize response times for language detection requests.

Throughput: Handle multiple requests concurrently without compromising performance.

Optimization: Optimize algorithms, data pipelines, and infrastructure for efficiency.

8. Monitoring and Maintenance:

Monitoring: Monitor system performance, resource utilization, and error rates.

Maintenance: Update models, fix bugs, and address performance issues as needed.

Feedback Loop: Incorporate user feedback and data updates to improve model accuracy over time.

Conclusion

System analysis of a language detection model involves examining its data processing pipeline, model development, evaluation, integration, and performance optimization. By analyzing each component and its interactions, stakeholders can ensure the effectiveness, efficiency, and reliability of the language detection system. Ongoing monitoring and maintenance are essential to address evolving requirements and maintain optimal performance.

IMPLEMENTATION

IMPLEMENTATION

The implementation of a language detection model involves translating the design and requirements into a functional system. Here's a step-by-step guide to implementing a language detection model:

STEP - 1 First step is importing all the libraries
Pandas (pd) for data manipulation.
NumPy (np) for numerical operations.
Matplotlib (plt) for plotting.
Seaborn (sns) for statistical data visualization.
Various modules from scikit-learn for feature extraction, model selection, Naïve Bayes classifier, and evaluation metrics.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report, confusion_matrix
```

STEP:2 This line reads the dataset from a CSV file named "dataset.csv" using Pandas **read_csv()** function and stores it in a DataFrame called data.

```
# Load dataset
data = pd.read_csv("dataset.csv")
```

STEP:3 These lines print basic information about the dataset using the info() method of the DataFrame. It displays the data types, number of non-null values, and memory usage.

```
# Display basic information about the dataset
print("Dataset information:")
print(data.info())
```

STEP:4 This line prints the first few rows of the dataset using the head() method of the DataFrame.

```
# Display first few rows of the dataset
print("First few rows of the dataset:")
data.head()
```

STEP:5 These lines check for null values in the dataset using the isnull() method of the DataFrame, followed by sum() to count the null values in each column.

```
# Check for null values
print("Null values check:")
print(data.isnull().sum())
```

STEP:6 These lines count the occurrences of each language in the dataset using the value_counts() method applied to the "language" column.

```
# Check languages present in dataset
print(" present Languages in dataset:")
print(data["language"].value_counts())
```

STEP:7 These lines visualize the distribution of languages in the dataset using Seaborn's countplot() function.

```
# Visualize distribution of languages
plt.figure(figsize=(10, 6))
sns.countplot(data=data, x="language")
plt.title("Distribution of Languages")
plt.xlabel("Language")
plt.ylabel("Count")
plt.xticks(rotation=45)
plt.show()
```

STEP:8 These lines calculate the number of unique languages present in the dataset using the `nunique()` method applied to the "language" column

```
unique_languages = data["language"].nunique()
print("Number of unique languages:")
print(unique_languages)
```

STEP:9 These lines extract the features (x) and labels (y) from the DataFrame and convert them into NumPy arrays.

```
# Split data into features and labels
x = np.array(data["Text"])
y = np.array(data["language"])
```

STEP:10 These lines initialize a CountVectorizer object **cv** and use it to convert the text data (x) into numerical format (X) suitable for ma

```
# Convert text data into numerical format
cv = CountVectorizer()
X = cv.fit_transform(x)
```

STEP:11 This line splits the data into training and test sets using train_test_split() function from scikit-learn. The test_size parameter specifies the proportion of the dataset to include in the test split, and random_state ensures reproducibility.

```
# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

STEP:12 These lines initialize a Multinomial Naïve Bayes model model and train it using the training data (X_train, y_train) with the fit() method.

```
# Train Multinomial Naïve Bayes model
model = MultinomialNB()
model.fit(X_train, y_train)
```

STEP:13 These lines generate a classification report using the classification_report() function from scikit-learn, which includes precision, recall, F1-score, and support for each class.

```
# Generate classification report
y_pred = model.predict(X_test)
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

STEP:14 This line calculates and prints the accuracy of the model on the test data using the score() method of the model.

```
# Evaluate model
accuracy = model.score(X_test, y_test)
print("Model Accuracy:", accuracy)
```

STEP:15 These lines generate a confusion matrix to evaluate the performance of the classifier and visualize it using Seaborn's heatmap() function.

```
# Generate confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(12, 8))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", xticklabels=model.classes_, yticklabels=model.classes_)
plt.title("Confusion Matrix")
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.show()
```

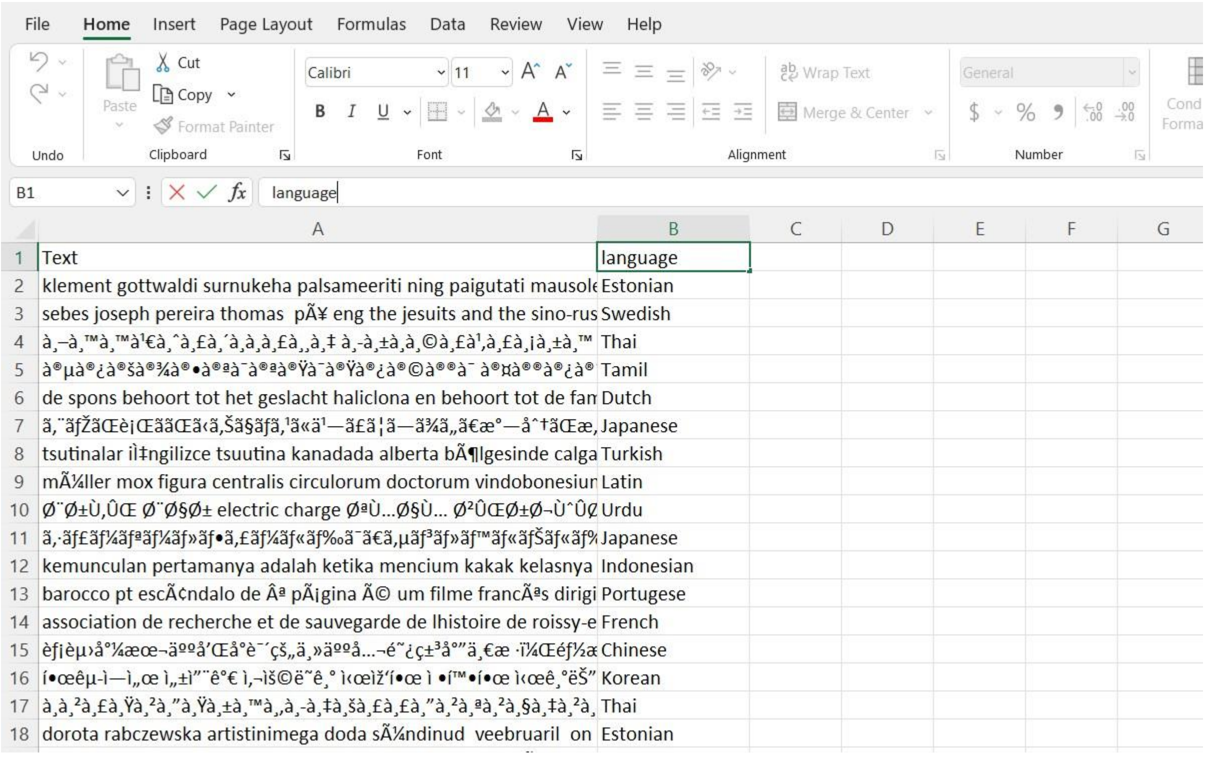
STEP:16 These lines prompt the user to enter a text and predict its language using the trained model. It converts the user input into numerical format using the transform() method of the CountVectorizer and then predicts the language using the predict() method of the model. Finally, it prints the predicted language.

```
# Language detection from user input
user_input = input("Enter a Text: ")
user_data = cv.transform([user_input]).toarray()
output = model.predict(user_data)
print("Predicted Language:", output)
```

DATA SET SCREENS

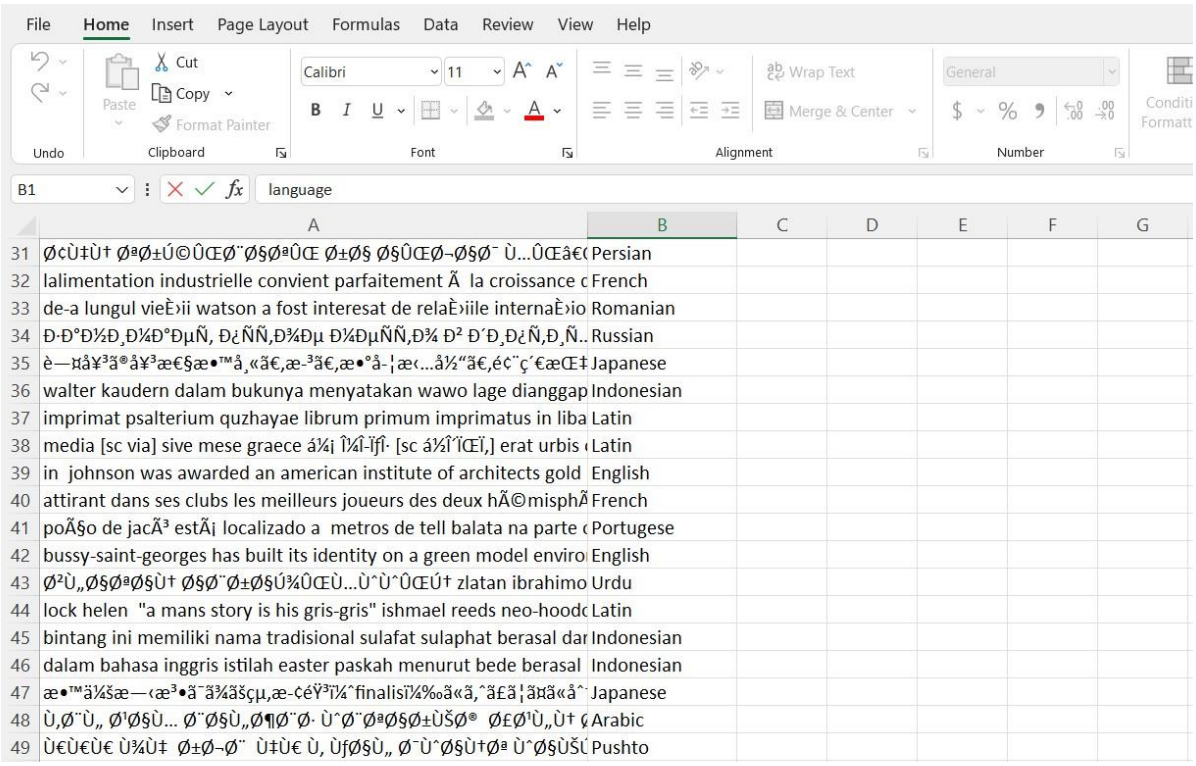
DATA SET SCREENS

Data tables containing information about different languages.



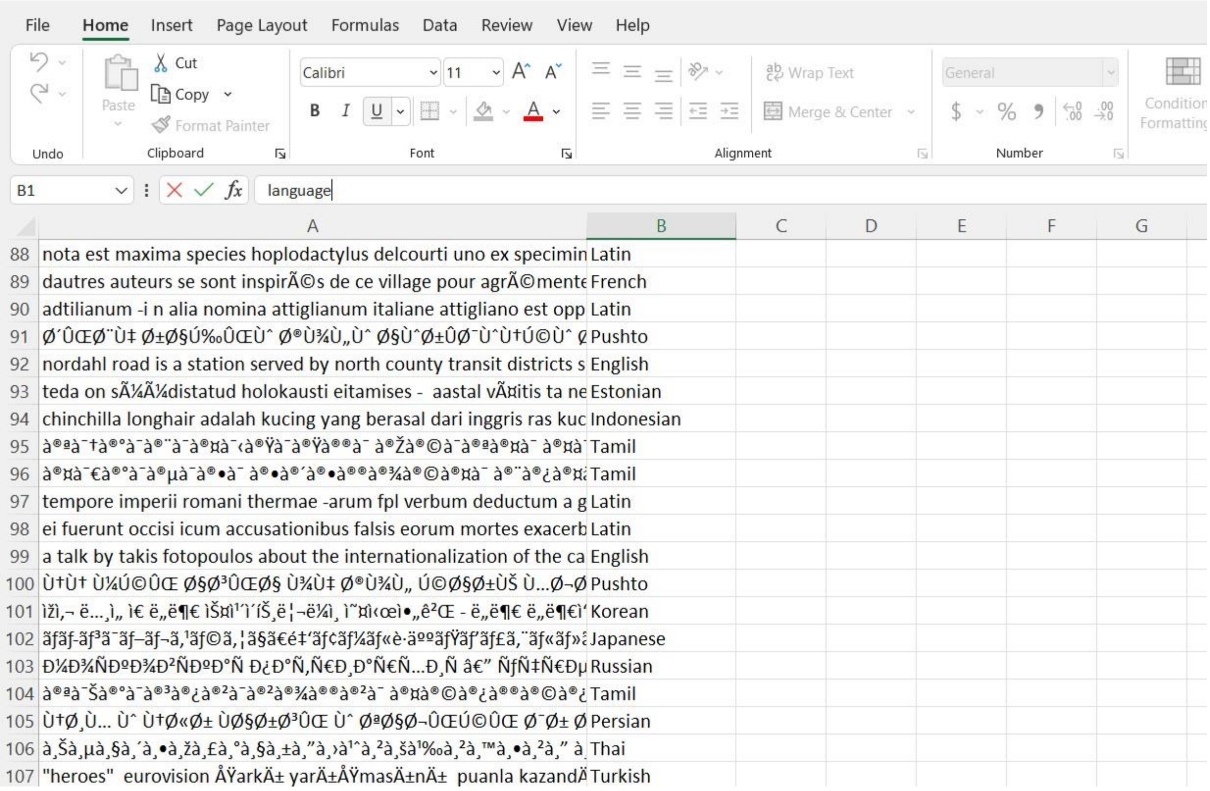
	Text	language
2	klement gottwaldi surnukeha palsameeriti ning paigutati mausole	Estonian
3	sebes joseph pereira thomas pÃ¥ eng the jesuits and the sino-rus	Swedish
4	à¸–à¸™à¸™à¹€à¸ˆà¸£à¸´à¸à¸à¸£à¸¸à¸‡ à¸à¸±à¸à¸©à¸£à¹‚à¸£à¸¡à¸±à¸™	Thai
5	à®µà®¿à®šà®¾à®•à®ªà¯à®ªà®Ÿà¯à®Ÿà®¿à®©à®®à¯ à®¤à®®à®¿à®	Tamil
6	de spons behoort tot het geslacht haliclona en behoort tot de fan	Dutch
7	ã‚¨ãƒŽãƒè¡ŒãƒãŒã‹ã‚Šã§ãƒã‚¹ã«ä¹—ã£ã¦ã—ã¾ã„ã€æ°—å^†ãŒæ,	Japanese
8	tsutinalar iÌ‡ngilizce tsuutina kanadada alberta bÃ¶lgesinde calga	Turkish
9	mÃ¼ller mox figura centralis circulorum doctorum vindobonesiur	Latin
10	Ø¨Ø±Ù‚ÛŒ Ø¨Ø§Ø± electric charge ØªÙ…Ø§Ù… Ø²ÛŒØ±Ø¬Ù^Û¢	Urdu
11	ã‚·ãƒ£ãƒ¼ãƒªãƒ¼ãƒ»ãƒ•ã‚£ãƒ¼ãƒ«ãƒ‰ã¯ã€ã‚µãƒ³ãƒ»ãƒ™ãƒ«ãƒŠãƒ«ãƒ%	Japanese
12	kemunculan pertamanya adalah ketika mencium kakak kelasnya	Indonesian
13	barocco pt escÃ¢ndalo de Âª pÃ¡gina Ã© um filme francÃªs dirigi	Portugese
14	association de recherche et de sauvegarde de lhistoire de roissy-e	French
15	èƒ¡èµ›å°¼æœ¬äººå'Œå°è¯´çš„ä¸»äººå…¬é˜¿ç±³å°”ä¸€æ ·ï¼Œéƒ½æ	Chinese
16	í•œêµì—ì„œ ì„±ì”¨ê°€ ì‚¬ìš©ë˜ê¸° ì‹œìž'í•œ ì •í™•í•œ ì‹œê¸°ëŠ”	Korean
17	à¸à¸²à¸£à¸Ÿà¸²à¸"à¸Ÿà¸±à¸™à¸‚à¸à¸‡à¸šà¸£à¸£à¸"à¸²à¸ªà¸²à¸§à¸‡à¸²à¸	Thai
18	dorota rabczewska artistinimega doda sÃ¼ndinud veebruaril on	Estonian

Data table containing information about different languages.



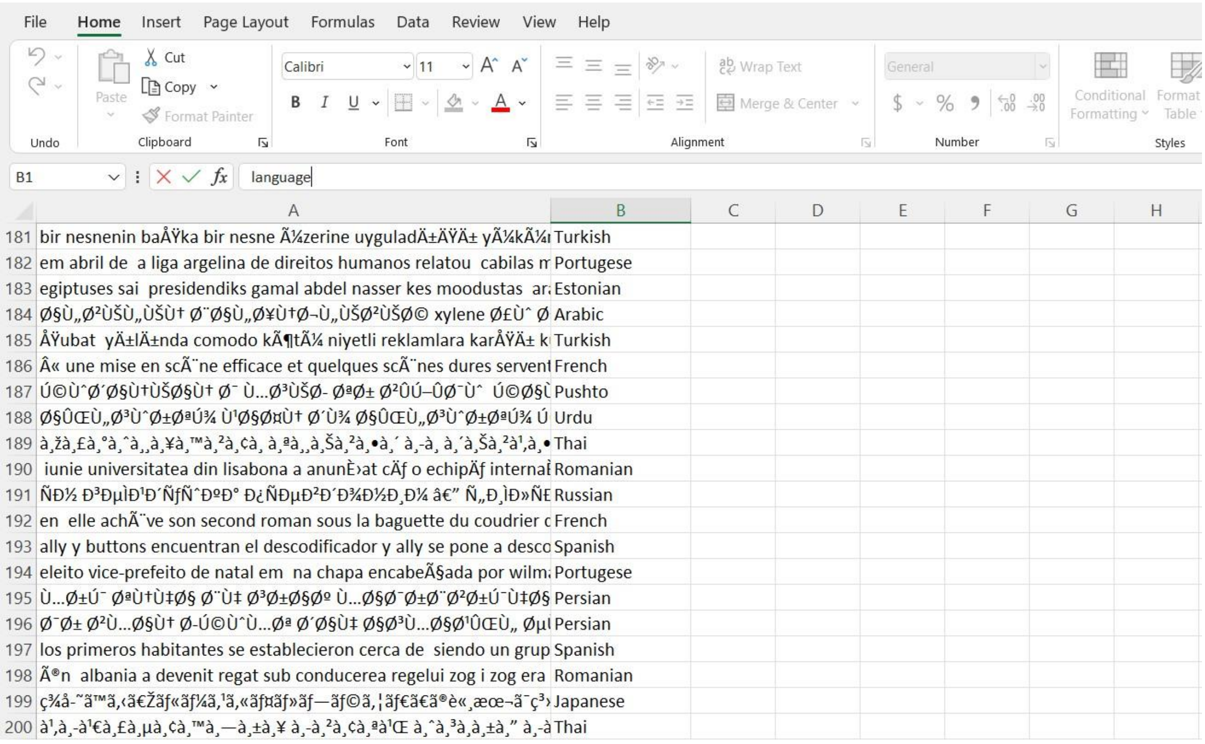
	A	B
31	Ø¢Ù‡Ù† ØªØ±Ú©ÛŒØ¨Ø§ØªÛŒ Ø±Ø§ Ø§ÛŒØ¬Ø§Ø¯ Ù…ÛŒâ€(	Persian
32	lalimentation industrielle convient parfaitement Ã la croissance c	French
33	de-a lungul vieÈ›ii watson a fost interesat de relaÈ›iile internaÈ›io	Romanian
34	Ð·Ð°Ð½Ð¸Ð¼Ð°ÐµÑ‚ Ð¿ÑÑ‚Ð¾Ðµ Ð¼ÐµÑÑ‚Ð¾ Ð² Ð´Ð¸Ð¿Ñ‚Ð¸Ñ..	Russian
35	è—¤å¥³ã®å¥³æ€§æ•™å¸«ã€æ-³ã€æ•°å-¦æ‹…å½“ã€é¢¨ç´€æŒ‡	Japanese
36	walter kaudern dalam bukunya menyatakan wawo lage dianggap	Indonesian
37	imprimat psalterium quzhayae librum primum imprimatus in liba	Latin
38	media [sc via] sive mese graece á¼¡ Î¼ÎÏƒÎ· [sc á½Î´ÏŒÏ‚] erat urbis (	Latin
39	in johnson was awarded an american institute of architects gold	English
40	attirant dans ses clubs les meilleurs joueurs des deux hÃ©misphÃ	French
41	poÃ§o de jacÃ³ estÃ¡ localizado a metros de tell balata na parte (	Portugese
42	bussy-saint-georges has built its identity on a green model enviro	English
43	Ø²Ù„Ø§ØªØ§Ù† Ø§Ø¨Ø±Ø§Ú¾ÛŒÙ…Ù^Ù^ÛŒÚ† zlatan ibrahimo	Urdu
44	lock helen "a mans story is his gris-gris" ishmael reeds neo-hood(	Latin
45	bintang ini memiliki nama tradisional sulafat sulaphat berasal dar	Indonesian
46	dalam bahasa inggris istilah easter paskah menurut bede berasal	Indonesian
47	æ•™ä¼šæ—‹æ³•ã¯ã¾ãšçµ‚æ-¢éŸ³ï¼ˆfinalisï¼‰ã«ã‚ˆã£ã¦ã¤ã«å^	Japanese
48	Ù‚Ø¨Ù„ Ø¹Ø§Ù… Ø¨Ø§Ù„Ø¶Ø¨Ø· Ù^Ø¨ØªØ§Ø±ÙŠØ® Ø£Ø¹Ù„Ù† (	Arabic
49	Ù€Ù€Ù€ Ù¾Ù‡ Ø±Ø¬Ø¨ Ù‡Ù€ Ù, ÙƒØ§Ù„ Ø¯Ù^Ø§Ù†Øª Ù^Ø§ÙŠÙ	Pushto

Data table containing information about different languages.



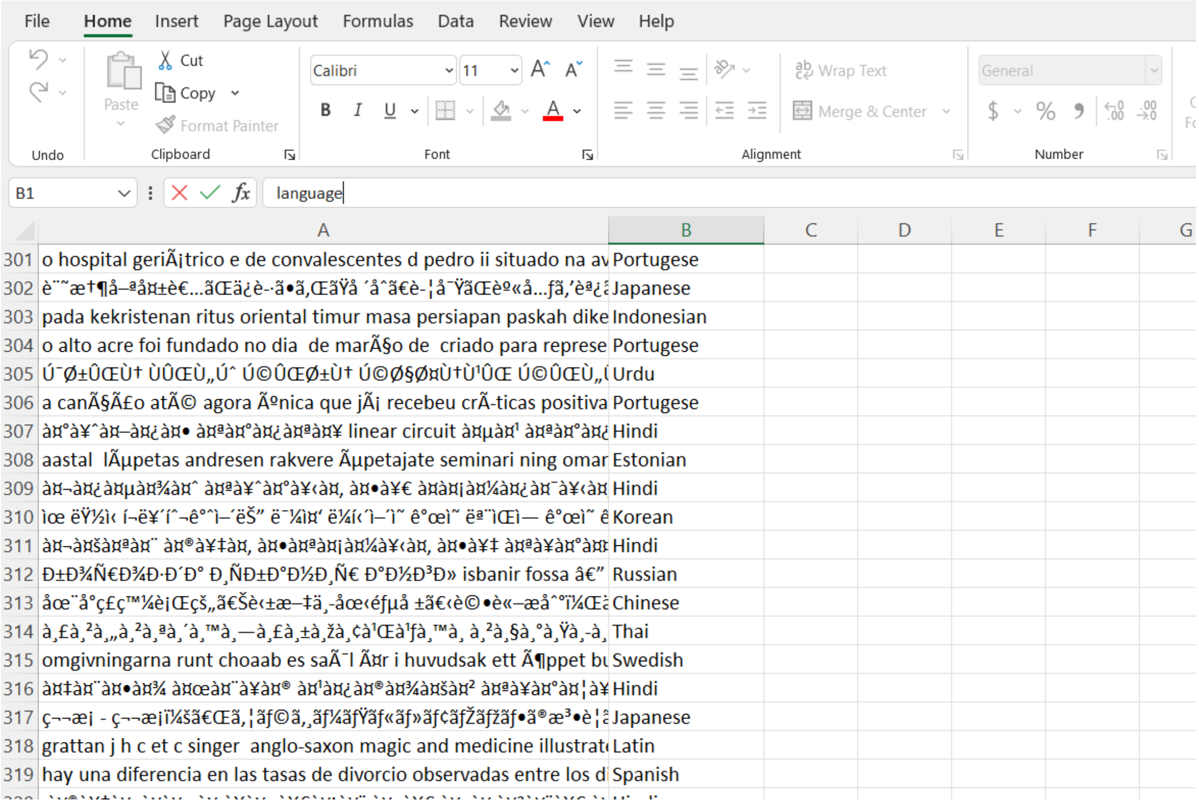
	A	B
88	nota est maxima species hoplodactylus delcourti uno ex specimin	Latin
89	dautres auteurs se sont inspirÃ©s de ce village pour agrÃ©mente	French
90	adtilianum -i n alia nomina attiglianum italiane attigliano est opp	Latin
91	Ø´ÛŒØ¨Ù‡ Ø±Ø§Ú‰ÛŒÙˆ Ø®Ù¾Ù„Ùˆ Ø§ÙˆØ±ÛØ¯ÙˆÙ†Ú©Ùˆ Ø	Pushto
92	nordahl road is a station served by north county transit districts s	English
93	teda on sÃ¼Ã¼distatud holokausti eitamises - aastal vÃ¤itis ta ne	Estonian
94	chinchilla longhair adalah kucing yang berasal dari inggris ras kuc	Indonesian
95	à®ªà¯†à®°à¯à®¨à¯à®¤à¯‹à®Ÿà¯à®Ÿà®®à¯ à®Žà®©à¯à®ªà®¤à¯ à®¤à	Tamil
96	à®¤à¯€à®°à¯à®µà¯à®•à¯ à®•à®´à®•à®®à®¾à®©à®¤à¯ à®¨à®¿à®¤	Tamil
97	tempore imperii romani thermae -arum fpl verbum deductum a g	Latin
98	ei fuerunt occisi icum accusationibus falsis eorum mortes exacerb	Latin
99	a talk by takis fotopoulos about the internationalization of the ca	English
100	Ù†Ù† Ù¼Ú©ÛŒ Ø§Ø³ÛŒØ§ Ù¾Ù‡ Ø®Ù¾Ù„ Ú©Ø§Ø±ÙŠ Ù…Ø¬Ø	Pushto
101	ìžì‚¬ ë…¸ì„ ì€ ë„ë¶€ ìŠ¤ì¹´ì´íŠ¸ë¦¬ë¼ì¸ ì˜¤ì‹œì•„ê²Œ - ë„ë¶€ ë„ë¶€ì'	Korean
102	ãƒãƒ³ãƒ‡ãƒ–ãƒ¬ã‚¹ãƒ©ã‚¦ã§ã€é‡‘ãƒ¢ãƒ¼ãƒ«è·äººãƒŸãƒ'ãƒ£ã‚¨ãƒ«ãƒ»ã	Japanese
103	Ð¼Ð¾ÑÐºÐ¾Ð²ÑÐºÐ°Ñ Ð¿Ð°Ñ‚Ñ€Ð¸Ð°Ñ€Ñ…Ð¸Ñ â€" ÑƒÑ‡Ñ€Ðµ	Russian
104	à®ªà¯Šà®°à¯à®³à®¿à®²à¯à®²à®¾à®®à®²à¯ à®¤à®©à®¿à®®à®©à®¿	Tamil
105	Ù†Ø¸Ù… Ùˆ Ù†Ø«Ø± ÙØ§Ø±Ø³ÛŒ Ùˆ ØªØ§Ø¬ÛŒÚ©ÛŒ Ø¯Ø± Ø	Persian
106	à¸Šà¸µà¸§à¸´à¸•à¸žà¸£à¸°à¸§à¸±à¸"à¸›à¹ˆà¸²à¸šà¹‰à¸²à¸™à¸•à¸²à¸" à¸	Thai
107	"heroes" eurovision ÅŸarkÄ± yarÄ±ÅŸmasÄ±nÄ± puanla kazandÄ	Turkish

Data table containing information about different languages.



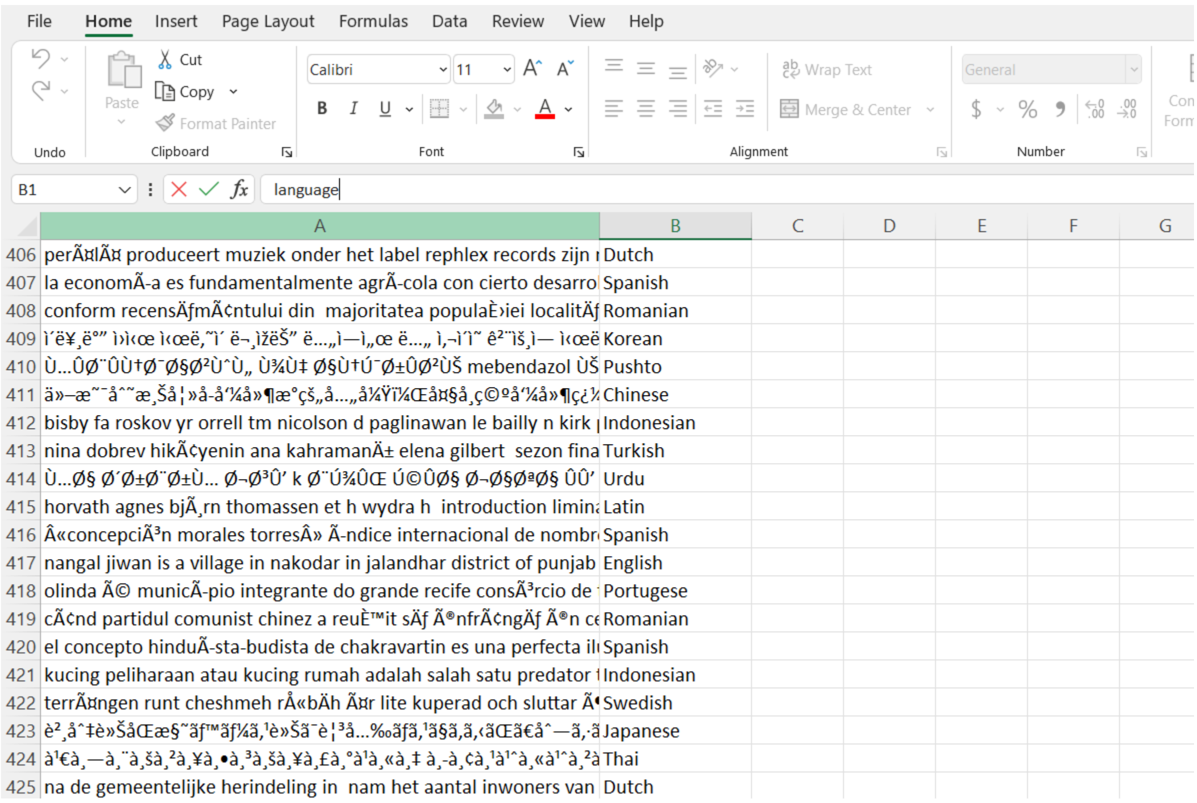
	A	B
181	bir nesnenin baÅŸka bir nesne Ã¼zerine uyguladÄ±ÄŸÄ± yÃ¼kÃ¼	Turkish
182	em abril de a liga argelina de direitos humanos relatou cabilas m	Portugese
183	egiptuses sai presidendiks gamal abdel nasser kes moodustas ar	Estonian
184	Ø§Ù„Ø²ÙŠÙ„ÙŠÙ† Ø¨Ø§Ù„Ø¥Ù†Ø¬Ù„ÙŠØ²ÙŠØ© xylene Ø£Ùˆ Ø	Arabic
185	ÅŸubat yÄ±lÄ±nda comodo kÃ¶tÃ¼ niyetli reklamlara karÅŸÄ± k	Turkish
186	Â« une mise en scÃ¨ne efficace et quelques scÃ¨nes dures servent	French
187	Ú©Ùˆ Ø´ Ø§Ù†ÙŠØ§Ù† Ø¯ Ù…Ø³ÙŠØ- ØªØ± Ø²ÛÚ–ÛØ¯Ùˆ Ú©Ø§Ù	Pushto
188	Ø§ÛŒÙ„Ø³ÙˆØ±ØªÚ¾ Ù¹Ø§Ø¤Ù† Ø´Ù¾ Ø§ÛŒÙ„Ø³ÙˆØ±ØªÚ¾ Ú	Urdu
189	à¸žà¸£à¸°à¸ˆà¸¸à¸¥à¸™à¸²à¸¢à¸ à¸ªà¸¸à¸Šà¸²à¸•à¸´ à¸-à¸ à¸´à¸Šà¸²à¹‚à¸•	Thai
190	iunie universitatea din lisabona a anunÈ›at cÄƒ o echipÄƒ internaÈ	Romanian
191	ÑÐ½ Ð³ÐµÌÐ¹Ð´ÑƒÑˆÐºÐ° Ð¿ÑÐµÐ²Ð´Ð¾Ð½Ð¸Ð¼ â€" Ñ„Ð¸ÌÐ»ÑŒ	Russian
192	en elle achÃ¨ve son second roman sous la baguette du coudrier c	French
193	ally y buttons encuentran el descodificador y ally se pone a desco	Spanish
194	eleito vice-prefeito de natal em na chapa encabeÃ§ada por wilm	Portugese
195	Ù…Ø±Ú¯ ØªÙ†Ù‡Ø§ Ø¨Ù‡ Ø³Ø±Ø§Øº Ù…Ø§Ø¯Ø±Ø¨Ø²Ø±Ú¯Ù‡Ø§	Persian
196	Ø¯Ø± Ø²Ù…Ø§Ù† ØÚ©ÙˆÙ…Øª Ø´Ø§Ù‡ Ø§Ø³Ù…Ø§Ø¹ÛŒÙ„ Øµ	Persian
197	los primeros habitantes se establecieron cerca de siendo un grup	Spanish
198	Ã®n albania a devenit regat sub conducerea regelui zog i zog era	Romanian
199	ç¾å˜ã™ã‚‹ã€Žãƒ«ãƒ¼ã‚¹ã‚«ãƒ¤ãƒ»ãƒ—ãƒ©ã‚¦ãƒ€ã€ã®è«¸æœ¬ã¯ç³›	Japanese
200	à¹‚à¸-à¹€à¸£à¸µà¸¢à¸™à¸—à¸±à¸¥ à¸-à¸²à¸¢à¸ªà¹Œ à¸ˆà¸³à¸à¸±à¸" à¸-à	Thai

Data table containing information about different languages.



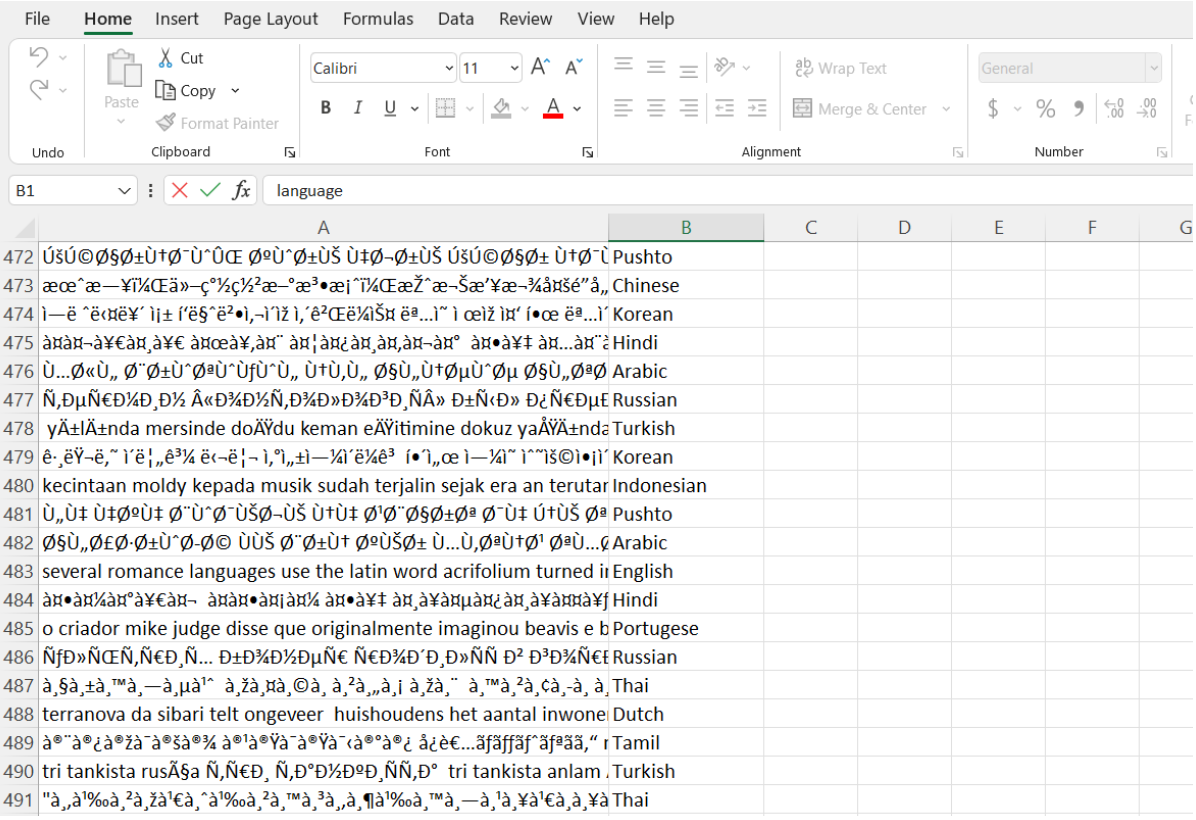
	A	B
301	o hospital geriÃ¡trico e de convalescentes d pedro ii situado na av	Portugese
302	è¨˜æ†¶å–ªå¤±è€…ãŒä¿è-·ã•ã‚ŒãŸå ´å^ã€è-¦å¯ŸãŒèº«å…ƒã'èª¿ã	Japanese
303	pada kekristenan ritus oriental timur masa persiapan paskah dike	Indonesian
304	o alto acre foi fundado no dia de marÃ§o de criado para represe	Portugese
305	Ú¯Ø±ÛŒÙ† ÙÛŒÙ„Ú^ Ú©ÛŒØ±Ù† Ú©Ø§Ø¤Ù†Ù¹ÛŒ Ú©ÛŒÙ„Ú	Urdu
306	a canÃ§Ã£o atÃ© agora Ãºnica que jÃ¡ recebeu crÃ-ticas positiva	Portugese
307	à¤°à¥ˆà¤–à¤¿à¤• à¤ªà¤°à¤¿à¤ªà¤¥ linear circuit à¤µà¤¹ à¤ªà¤°à¤¿	Hindi
308	aastal lÃµpetas andresen rakvere Ãµpetajate seminari ning omar	Estonian
309	à¤¬à¤¿à¤µà¤¾à¤ˆ à¤ªà¥ˆà¤°à¥‹à¤‚ à¤•à¥€ à¤à¤¡à¤¼à¤¿à¤¯à¥‹à¤	Hindi
310	ìœ ëŸ½ì‹ í¬ë¥´íˆ¬ê°ˆì–´ëŠ” ë¯¼ì¤' ë¼í‹´ì–´ì˜ ê°œì˜ ëª¨ìŒì— ê°œì˜ ê	Korean
311	à¤¬à¤šà¤ªà¤¨ à¤®à¥‡à¤‚ à¤•à¤ªà¤¡à¤¼à¥‹à¤‚ à¤•à¥‡ à¤ªà¥à¤°à¤	Hindi
312	Ð±Ð¾Ñ€Ð¾Ð·Ð´Ð° Ð¸ÑÐ±Ð°Ð½Ð¸Ñ€ Ð°Ð½Ð³Ð» isbanir fossa â€"	Russian
313	åœ¨å°ç£ç™¼è¡Œçš„ã€Šè‹±æ–‡ä¸åœ‹éƒµå ±ã€‹è©•è«–æå^°ï¼Œä	Chinese
314	à¸£à¸²à¸„à¸²à¸ªà¸´à¸™à¸—à¸£à¸±à¸žà¸¢à¹Œà¹ƒà¸™à¸ à¸²à¸§à¸°à¸Ÿà¸à¸	Thai
315	omgivningarna runt choaab es saÃ¯l Ã¤r i huvudsak ett Ã¶ppet bu	Swedish
316	à¤‡à¤¨à¤•à¤¾ à¤œà¤¨à¥à¤® à¤¹à¤¿à¤®à¤¾à¤šà¤² à¤ªà¥à¤°à¤¦à¥	Hindi
317	ç¬¬æ¡ - ç¬¬æ¡ï¼šã€Œã‚¦ãƒ©ã‚¸ãƒ¼ãƒŸãƒ«ãƒ»ãƒ¢ãƒŽãƒžãƒ•ã®æ³•è¦ã	Japanese
318	grattan j h c et c singer anglo-saxon magic and medicine illustrat	Latin
319	hay una diferencia en las tasas de divorcio observadas entre los d	Spanish

Data table containing information about different languages.



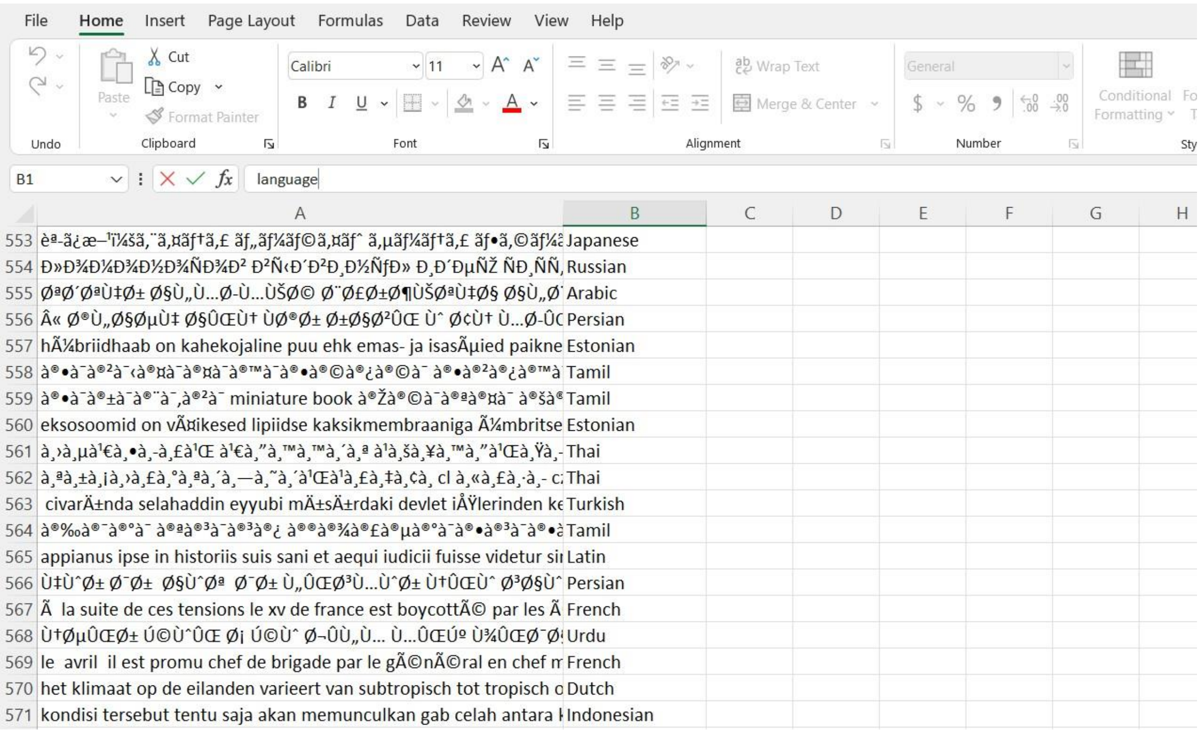
	A	B
406	perÃ¤lÃ¤ produceert muziek onder het label rephlex records zijn	Dutch
407	la economÃ-a es fundamentalmente agrÃ-cola con cierto desarro	Spanish
408	conform recensÄƒmÃ¢ntului din majoritatea populaÈ›iei localitÄƒ	Romanian
409	ì´ë¥¸ë°" ì›ì‹œ ì‹œë,˜ì´ ë¬¸ìžëŠ" ë…„ì—ì„œ ë…„ ì,¬ì´ì˜ ê²¨ìš¸ì— ì‹œë	Korean
410	Ù…ÛØ¨ÛÙ†Ø¯Ø§Ø²Ù^Ù„ Ù¾Ù‡ Ø§Ù†Ú¯Ø±ÛØ²ÙŠ mebendazol ÙŠ	Pushto
411	ä»–æ˜¯å^˜æ¸Šå¦»å-å'¼å»¶æ°çš„å…„å¼Ÿï¼Œå¤§å¸ç©ºå'¼å»¶ç¿	Chinese
412	bisby fa roskov yr orrell tm nicolson d paglinawan le bailly n kirk	Indonesian
413	nina dobrev hikÃ¢yenin ana kahramanÄ± elena gilbert sezon fina	Turkish
414	Ù…Ø§ Ø´Ø±Ø¨Ø±Ù… Ø¬Ø³Û' k Ø¨Ú¾ÛŒ Ú©ÛØ§ Ø¬Ø§ØªØ§ ÛÛ'	Urdu
415	horvath agnes bjÃ¸rn thomassen et h wydra h introduction limin	Latin
416	Â«concepciÃ³n morales torresÂ» Ã-ndice internacional de nombr	Spanish
417	nangal jiwan is a village in nakodar in jalandhar district of punjab	English
418	olinda Ã© municÃ-pio integrante do grande recife consÃ³rcio de	Portugese
419	cÃ¢nd partidul comunist chinez a reuÈ™it sÄƒ Ã®nfrÃ¢ngÄƒ Ã®n c	Romanian
420	el concepto hinduÃ-sta-budista de chakravartin es una perfecta il	Spanish
421	kucing peliharaan atau kucing rumah adalah salah satu predator	Indonesian
422	terrÃ¤ngen runt cheshmeh rÅ«bÄh Ã¤r lite kuperad och sluttar Ã	Swedish
423	è²¸å^‡è»ŠåŒæ§˜ãƒ™ãƒ¼ã‚¹è»Šã¯è¦³å…‰ãƒã‚¹ã§ã‚ã‚‹ãŒã€å^—ã,·ã	Japanese
424	à¹€à¸—à¸¨à¸šà¸²à¸¥à¸•à¸³à¸šà¸¥à¸£à¸°à¹à¸«à¸‡ à¸-à¸¢à¸¹à¹ˆà¸«à¹ˆà¸²à	Thai
425	na de gemeentelijke herindeling in nam het aantal inwoners van	Dutch

Data table containing information about different languages.



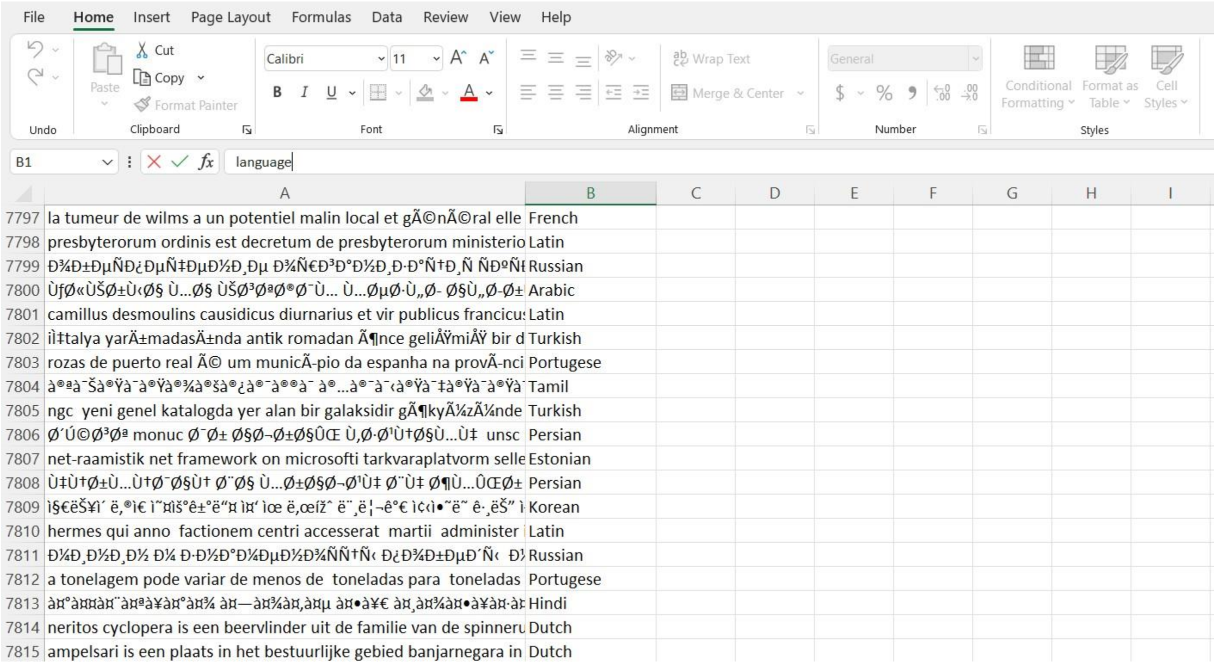
	A	B
472	ÚšÚ©Ø§Ø±Ù†Ø¯Ù ÛŒ ØºÙˆØ±ÙŠ Ù‡Ø¬Ø±ÙŠ ÚšÚ©Ø§Ø± Ù†Ø¯Ù	Pushto
473	æœˆæ—¥ï¼Œä»–ç°½ç½²æ–°æ³•æ¡ˆï¼ŒæŽˆæ¬Šæ'¥æ¬¾å¤šé"å„	Chinese
474	ì—ë ˆë‹¤ë¥´ ì¡± í'ë§ˆë²•ì‚¬ì´ìž ì‚´ê²Œë¼ìŠ¤ 모ì˜ ì œìž ì¤' 한 모ì	Korean
475	à¤à¤¬à¥€à¤¸à¥€ à¤œà¥‚à¤¨ à¤¦à¤¿à¤¸à¤‚à¤¬à¤° à¤•à¥‡ à¤…à¤¨	Hindi
476	Ù…Ø«Ù„ Ø¨Ø±Ù ˆØªÙˆÙƒÙˆÙ„ Ù†Ù‚Ù„ Ø§Ù„Ù†ØµÙˆØµ Ø§Ù„ØªØ	Arabic
477	Ñ‚ÐµÑ€Ð¼Ð¸Ð½ Â«Ð¾Ð½Ñ‚Ð¾Ð»Ð¾Ð³Ð¸ÑÂ» Ð±Ñ‹Ð» Ð¿Ñ€ÐµÐ	Russian
478	yÄ±lÄ±nda mersinde doÄŸdu keman eÄŸitimine dokuz yaÅŸÄ±nda	Turkish
479	ê·¸ëŸ¬ë‚˜ ì´ë¦„ê³¼ ë‹¬ë¦¬ ì‚°ì„±ì—¼ì´ë¼ê³ í•´ì„œ ì—¼ì˜ ìš©ì•¡ì	Korean
480	kecintaan moldy kepada musik sudah terjalin sejak era an terutar	Indonesian
481	Ù„Ù‡ Ù‡ØºÙ‡ Ø¨Ù ˆØ¯ÙŠØ¬ÙŠ Ù†Ù‡ Ø¹Ø¨Ø§Ø±Øª Ø¯Ù‡ Ú†ÙŠ Øª	Pushto
482	Ø§Ù„Ø£Ø·Ø±ÙˆØØ© ÙÙŠ Ø¨Ø±Ù† ØºÙŠØ± Ù…Ù‚ØªÙ†Ø¹ ØªÙ…	Arabic
483	several romance languages use the latin word acrifolium turned ir	English
484	à¤•à¤¼à¤°à¥€à¤¬ à¤à¤•à¤¡à¤¼ à¤•à¥‡ à¤¸à¥à¤µà¤¿à¤¸à¥à¤¤à¥ƒ	Hindi
485	o criador mike judge disse que originalmente imaginou beavis e b	Portugese
486	ÑƒÐ»ÑŒÑ‚Ñ€Ð¸Ñ… Ð±Ð¾Ð½ÐµÑ€ Ñ€Ð¾Ð´Ð¸Ð»ÑÑ Ð² Ð³Ð¾Ñ€Ð	Russian
487	à¸§à¸±à¸™à¸—à¸µà¹ˆ à¸žà¸¤à¸©à¸ à¸²à¸„à¸¡ à¸žà¸¨ à¸™à¸²à¸¢à¸à¸ à¸	Thai
488	terranova da sibari telt ongeveer huishoudens het aantal inwoner	Dutch
489	à®¨à®¿à®žà¯à®šà®¾ à®¹à®Ÿà¯à®Ÿà¯‹à®°à®¿ å¿è€…ãƒãƒƒãƒˆãƒªãã, " r	Tamil
490	tri tankista rusÃ§a Ñ‚Ñ€Ð¸ Ñ‚Ð°Ð½ÐºÐ¸ÑÑ‚Ð° tri tankista anlam	Turkish
491	"à¸‚à¹‰à¸²à¸žà¹€à¸ˆà¹‰à¸²à¸™à¸³à¸‚à¸¶à¹‰à¸™à¸—à¸¹à¸¥à¹€à¸à¸¥à	Thai

Data table containing information about different languages.

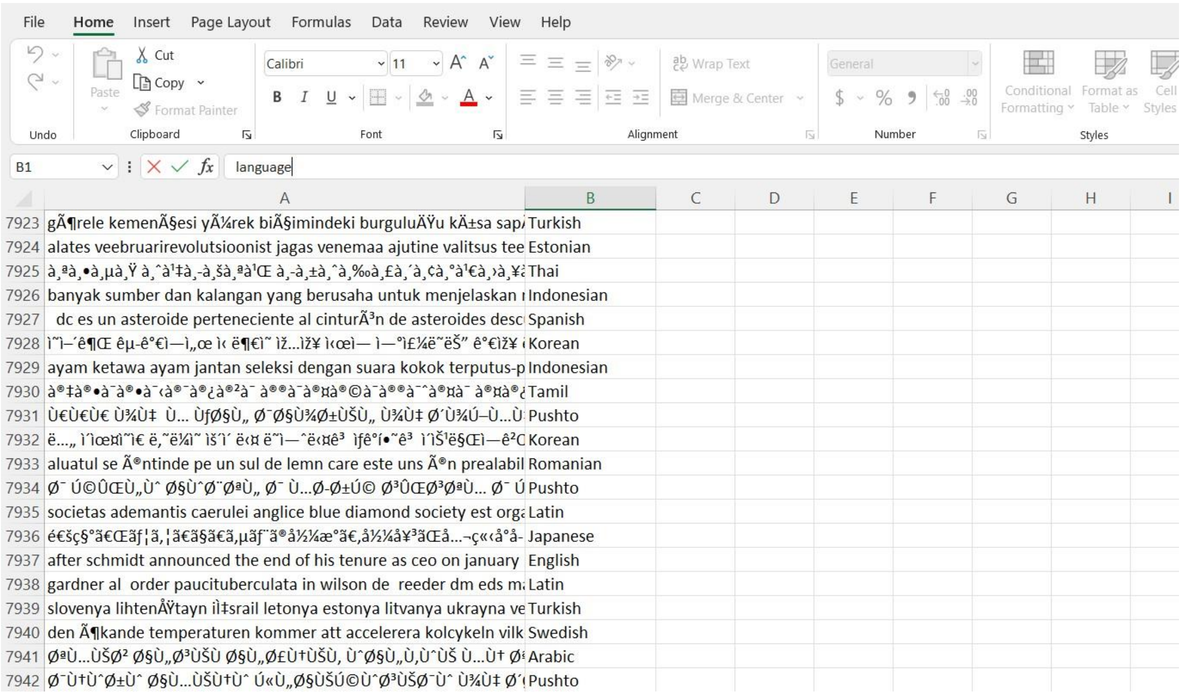


	A	B
553	èªã¿æ–¹ï¼šã,¨ã,¤ãƒ†ã,£ ãƒ„ãƒ¼ãƒ©ã,¤ãƒˆ ã,µãƒ¼ãƒ†ã,£ ãƒ•ã,©ãƒ¼ã	Japanese
554	Ð»Ð¾Ð¼Ð¾Ð½Ð¾ÑÐ¾Ð² Ð²Ñ‹Ð´Ð²Ð¸Ð½ÑƒÐ» Ð¸Ð´ÐµÑŽ ÑÐ¸ÑÑ‚	Russian
555	ØªØ´ØªÙ‡Ø± Ø§Ù„Ù…ØÙ…ÙŠØ© Ø¨Ø£Ø±Ø¶ÙŠØªÙ‡Ø§ Ø§Ù„Ø	Arabic
556	Â« Ø®Ù„Ø§ØµÙ‡ Ø§ÛŒÙ† ÙØ®Ø± Ø±Ø§Ø²ÛŒ Ùˆ Ø¢Ù† Ù…ØÛŒ	Persian
557	hÃ¼briidhaab on kahekojaline puu ehk emas- ja isasÃµied paikne	Estonian
558	à®•à¯à®²à¯‹à®¤à¯à®¤à¯à®™à¯à®•à®©à®¿à®©à¯ à®•à®²à®¿à®™à	Tamil
559	à®•à¯à®±à¯à®¨à¯,à®²à¯ miniature book à®Žà®©à¯à®ªà®¤à¯ à®šà®	Tamil
560	eksosoomid on vÃ¤ikesed lipiidse kaksikmembraaniga Ã¼mbritse	Estonian
561	à¸›à¸µà¹€à¸•à¸à¸£à¹Œ à¹€à¸"à¸™à¸™à¸´à¸ª à¹à¸šà¸¥à¸™à¸"à¹Œà¸Ÿà¸-	Thai
562	à¸ªà¸±à¸¡à¸›à¸£à¸°à¸ªà¸´à¸—à¸˜à¸´à¹Œà¹à¸£à¸‡à¸¢à¸ cl à¸«à¸£à¸·à¸- c	Thai
563	civarÄ±nda selahaddin eyyubi mÄ±sÄ±rdaki devlet iÅŸlerinden ke	Turkish
564	à®‰à®¯à®°à¯ à®ªà®³à¯à®³à®¿ à®®à®¾à®£à®µà®°à¯à®•à®³à¯à®•à	Tamil
565	appianus ipse in historiis suis sani et aequi iudicii fuisse videtur sir	Latin
566	Ù‡Ù ˆØ± Ø¯Ø± Ø§Ù ˆØª Ø¯Ø± Ù„ÛŒØ³Ù…Ù ˆØ± Ù†ÛŒÙ ˆ Ø³Ø§Ù ˆ	Persian
567	Ã la suite de ces tensions le xv de france est boycottÃ© par les Ã	French
568	Ù†ØµÛŒØ± Ú©Ù ˆÛŒ Ø¡ Ú©Ù ˆ Ø¬ÛÙ„Ù… Ù…ÛŒÚº Ù¾ÛŒØ¯Ø	Urdu
569	le avril il est promu chef de brigade par le gÃ©nÃ©ral en chef m	French
570	het klimaat op de eilanden varieert van subtropisch tot tropisch o	Dutch
571	kondisi tersebut tentu saja akan memunculkan gab celah antara k	Indonesian

Data table containing information about different languages.



Data table containing information about different languages.



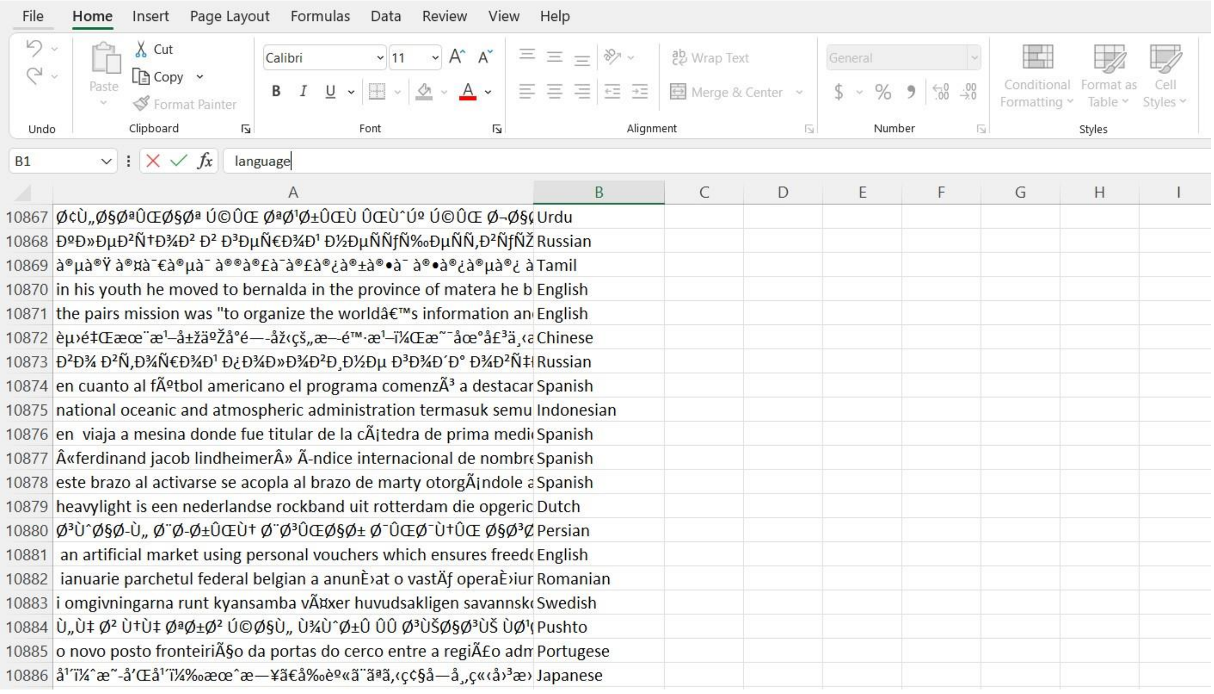
Data table containing information about different languages.

	A	B
9350	imposibilitatea teoreticĂƒ de a realiza filtre perfecte nu este un m	Romanian
9351	hoppstĂ¤dten-weiersbach is een plaats in de duitse deelstaat rijnl	Dutch
9352	robert aubĂ©ry fut conseiller au grand conseil au parlement de	French
9353	le sillon de bretagne est une rĂ©alisation de la sociĂ©tĂ© anon	French
9354	ë˜í•œ ì• ë‹ˆì—ì„ ì „í•™ ì˜¨ ìœ„ë„ˆë¥¼ ì²˜ìŒ ì—” ê·¸ë¦¬ ì¢‹ì•„í•˜ì§	Korean
9355	tipula pterelachisus margarita is een tweevleugelige uit de familie	Dutch
9356	existem evidĂªncias arqueolĂ³gicas que sugerem que povos dedic	Portugese
9357	the columbia high school student council was named an "njasc hc	English
9358	loomadest kahjustavad haaba nĂ¤rilised hiired ning jĂ¤nesed sĂ¼	Estonian
9359	à®®à®•à¯à®Ÿ à®…à®Ÿà®µà¯à®•à®³à¯ à®ªà®°à®¤à®¨à®¾à®Ÿà¯à®Ÿà	Tamil
9360	ç¾åœ¨ã®å®®æ®¿ã®ä¸»è¦éƒ¨åˆ†ã¯ã€å¹´ã‹ã‚‰å¹´ã«ã‹ã‘ã¦ã€æœ‰	Japanese
9361	Ð°Ð»ÐµÐºÑÐµÐµÐ² Ð¸ Ð¸ Ð²Ð·Ð²Ð¸Ð»Ð¾ÑÑŒ Ð·Ð½Ð°Ð¼Ñ Ð½Ð	Russian
9362	à¸§à¸±à¸™à¸ˆà¸±à¸™à¸—à¸£à¹Œà¸«à¸¥à¹‡à¸à¹€à¸›à¹‡à¸™à¸à¸²à¸£à¸«à¸¡à¸¸à¸™	Thai
9363	â€¢ staf ahlinarasumber dalam rapat dengar pendapat umum di c	Indonesian
9364	aemilius gabba editionem suam libri quinti bellorum civilium pret	Latin
9365	Ñ…ÑƒÐ´Ð¾Ð¶Ð½Ð¸Ðº Ð¿Ñ€ÐµÐ´Ð¿Ñ€Ð¸Ð½ÑÐ» Ð½ÐµÑÐºÐ¾Ð›	Russian
9366	ÙØ§Ù„Ø² Ø³ÛŒØªÛŒ Ø¨Ù‡ Ø§Ù†Ú¯Ù„ÛŒØ³ÛŒ falls city Ø´Ù‡Ø	Persian
9367	à¤°à¤¾à¤œà¤¸à¥à¤¥à¤¾à¤¨ à¤®à¥‡à¤‚ à¤šà¥à¤°à¥‚ à¤œà¤¿à¤²à¥	Hindi
9368	ì „ì²´ ì¸ êµ•ì˜ë¯¼ì€ ì ëŒ ìµœê³ ê¸°ë¡ê¸°ì˜¨ì´ Ëšc ìµœì €ê¸°ë¡ê¸°ì˜¨ì	Korean
9369	Ð´Ð»Ñ Ð¼Ð¾Ð½Ñ‚Ð°Ð¶Ð° Ð½Ð° ÑÐ°Ñƒ Ð²Ñ‹Ð¿ÑƒÑÐºÐ°Ð»ÑÑ Ð	Russian
9370	Ù¾Ø§Ú©Ø³ØªØ§Ù† Ù…ÛŒÚº Ù…Ø³ÛŒØÛŒ ØªÙ…Ø§Ù… Ø§Ù‚Ù	Urdu
9371	à¤à¤• à¤ªà¤¾à¤°à¤‚à¤ªà¤°à¤¿à¤• à¤ à¥‹à¤¸-à¤…à¤µà¤¸à¥à¤¥à¤¾	Hindi

Data table containing information about different languages.

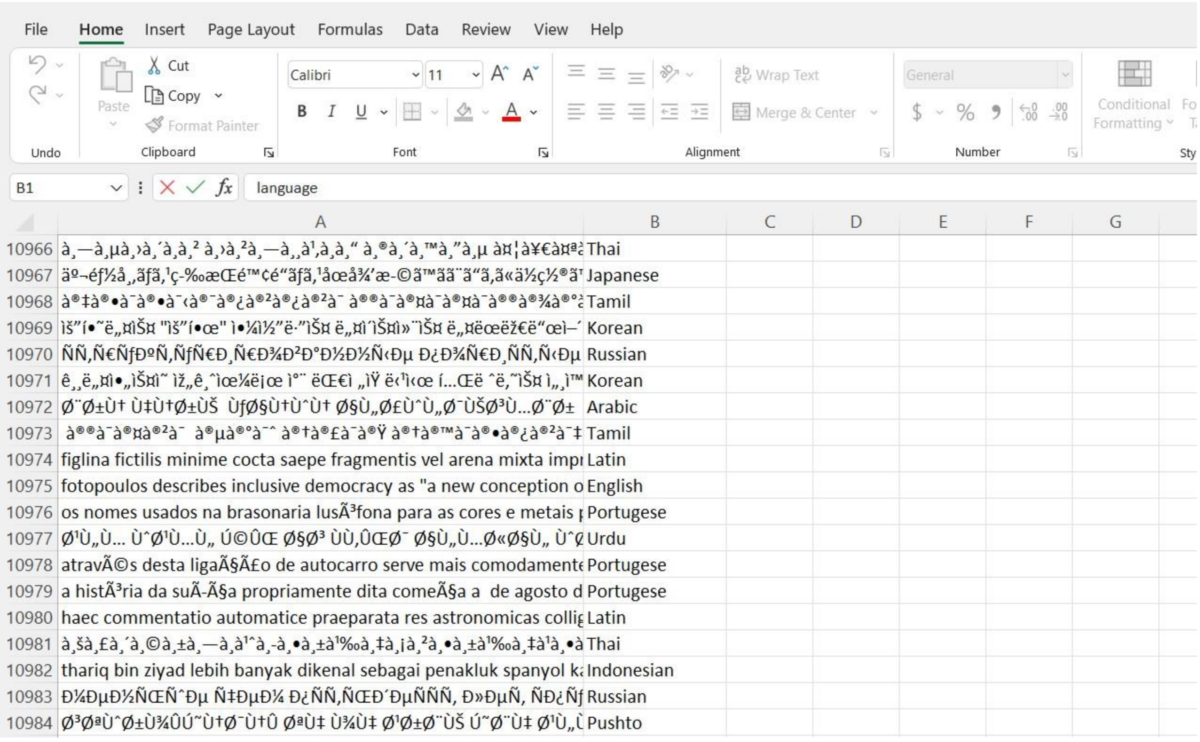
	A	B
9533	Ø¯Ø§Ø³Û Ú©Ø³Ø§Ù† Ù†Ù‡ ÙŠÙˆØ§Ø²Û Ø®Ù¾Ù„Ù‡ Ù…ÙˆØ±Ù†Û	Pushto
9534	Ú¯Ø±Ø¨Ù‡ Ø³ÛŒØ§Ù‡ Ø¨Ù‡ Ø§Ù†Ú¯Ù„ÛŒØ³ÛŒ black cat ÛŒÚ	Persian
9535	Ă®n general receptoarele gps sunt compuse dintr-o antenĂƒ incor	Romanian
9536	[villon franĂ§ois] oĂ¹ sont les neiges dantan? met een houtgravu	Dutch
9537	untuk mendukung itu akan diterapkan asean law asean currency	Indonesian
9538	ÙˆØ¹Ù‚Ø¯Øª Ø¨Ø±Ù† ÙÙŠ ÙƒØ±ÙŠÙÙŠÙ„Ø¯ Ø«Ù… ÙÙŠ Ù‡ÙˆÙ„Ø	Arabic
9539	à®†à®®à¯ à®¨à¯‚à®±à¯à®±à®¾à®£à¯à®Ÿà¯à®•à®³à®¿à®²à¯ à®®à®²	Tamil
9540	æ–¯å¤§æž—æœ¬äººå¸¦æœ‰é²œæ˜Žçš„åçŠ¹ä¸»ä¹‰å€¾å'ï¼Œæ	Chinese
9541	kasaba yĂ¼zyÄ±lda bĂ¶lgedeki birĂ§ok yerle beraber osmanlÄ±	Turkish
9542	à¸¡à¸µà¹‚à¸„à¸£à¸‡à¸à¸²à¸£à¸£à¸´à¹€à¸£à¸´à¹ˆà¸¡à¸„à¸¥à¹‰à¸²à¸¢à¸à¸	Thai
9543	noĂ«l godin surnommĂ© georges le gloupier â€” en rĂ©fĂ©ren	French
9544	ao ato de desenhar um brasĂ£o dĂ¡-se o nome de brasonar para	Portugese
9545	Ù†ÛŒØ±ÙˆÛŒ Ø¯Ø±ÛŒØ§ÛŒÛŒ Ø§Ù…Ù¾Ø±Ø§ØªÙˆØ±ÛŒ Ø¢	Persian
9546	å¾žæœˆæ—¥èµ·è¨±æ°¸ç”Ÿè‡æœ´æ”¿ç‰å…±åŒä¸»æŒé››å°sb	Chinese
9547	Ø¯Ø± Ø³Ø§Ù„ Ù‡Ù€Ù‚ Ø³Ù¾Ø§Ù‡ÛŒØ§Ù† Ù†Ø§Ø¯Ø±Ø´Ø§Ù‡ Ø	Persian
9548	antonius gregorii filius rubinstein russice Ð°Ð½Ñ‚Ð¾Ð½ Ð³Ñ€Ð¸Ð³	Latin
9549	la Ă®nceputul anilor marina regalĂƒ britanicĂƒ dezvoltĂƒ Ă®n tim	Romanian
9550	-Ð¹ Ð¸ÑÑ‚Ñ€ÐµÐ±Ð¸Ñ‚ÐµÐ»ÑŒÐ½Ñ‹Ð¹ Ð°Ð²Ð¸Ð°Ñ†Ð¸Ð¾Ð½Ð½Ñ	Russian
9551	le territoire de la commune est traversĂ© en son centre par lautc	French
9552	Ù‚Ø§Ù… ØºØ±ÙŠÙ†ÙŠÙ„ Ø¨ØªÙˆØ«ÙŠÙ‚ Ø§Ù„ØµÙŠØ¯ ØºÙŠØ± (	Arabic
9553	à®‡à®¨à¯à®¤à®¿à®¯ à®®à®•à¯à®•à®³à¯ à®¤à¯Šà®•à¯ˆ à®•à®£à®•	Tamil

Data table containing information about different languages.



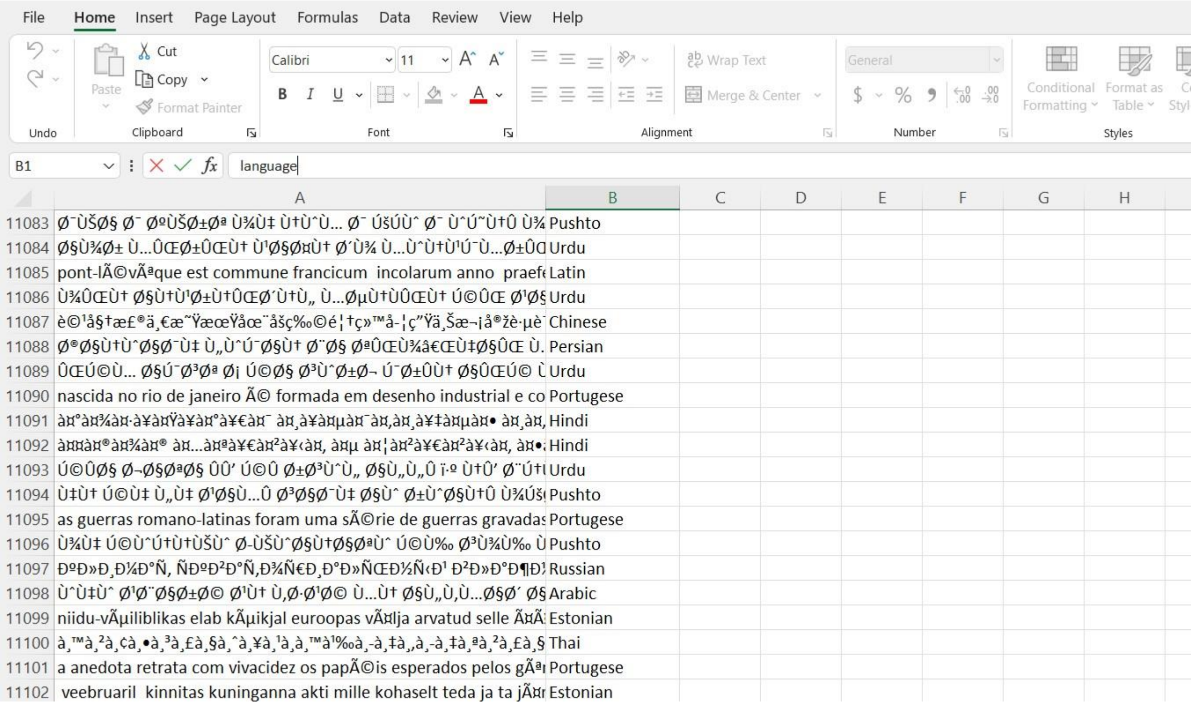
	A	B
10867	Ø¢Ù„Ø§Ø¡ØªÛŒØ§Øª Ú©ÛŒ Ø¡Ø¹Ø±ÛŒÙ ÛŒÙˆÚº Ú©ÛŒ Ø¬Ø§	Urdu
10868	ÐºÐ»ÐµÐ²Ñ†Ð¾Ð² Ð² Ð³ÐµÑ€Ð¾Ð¹ Ð½ÐµÑÑƒÑ‰ÐµÑ‚Ð²ÑƒÑŽ	Russian
10869	à®µà®Ÿ à®¤à¯€à®µà¯ à®®à®£à¯à®£à®¿à®±à®•à¯ à®•à®¿à®µà®¿ à	Tamil
10870	in his youth he moved to bernalda in the province of matera he b	English
10871	the pairs mission was "to organize the worldâ€™s information an	English
10872	èµ›é‡Œæœ¨æ¹–å±žäºŽå°é—åž‹ç„‚æ—é™æ¹–ï¼Œæ˜¯åœ°åŠ³ä¸‹	Chinese
10873	Ð²Ð¾ Ð²Ñ‚Ð¾Ñ€Ð¾Ð¹ Ð¿Ð¾Ð»Ð¾Ð²Ð¸Ð½Ðµ Ð³Ð¾Ð´Ð° Ð¾Ð²Ñ‡	Russian
10874	en cuanto al fÃºtbol americano el programa comenzÃ³ a destacar	Spanish
10875	national oceanic and atmospheric administration termasuk semu	Indonesian
10876	en viaja a mesina donde fue titular de la cÃ¡tedra de prima medi	Spanish
10877	Â«ferdinand jacob lindheimerÂ» Ãndice internacional de nombre	Spanish
10878	este brazo al activarse se acopla al brazo de marty otorgÃ¡ndole a	Spanish
10879	heavylight is een nederlandse rockband uit rotterdam die opgeric	Dutch
10880	Ø³Ù^Ø§ØÙ„ Ø¨ØØ±ÛŒÙ† Ø¨Ø³ÛŒØ§Ø± Ø¯ÛŒØ¯Ù†ÛŒ Ø§Ø³Ø	Persian
10881	an artificial market using personal vouchers which ensures freed	English
10882	ianuarie parchetul federal belgian a anunÈ›at o vastÄƒ operaÈ›iur	Romanian
10883	i omgivningarna runt kyansamba vÃ¤xer huvudsakligen savannsk	Swedish
10884	Ù„Ù‡ Ø² Ù†Ù‡ Ø¡Ø±Ø² Ú©Ø©Ø§Ù„ Ù¾Ù^Ø±Û Û Ø³ÙŠØ§Ø³ÙŠ ÙØ	Pushto
10885	o novo posto fronteiriÃ§o da portas do cerco entre a regiÃ£o adm	Portugese
10886	å¹³ï¼ˆæ˜å'Œå¹´ï¼‰æœˆæ—¥ãƒ€å‰ãƒ«ãƒ¼ãƒ³ã‚¯ç§å—å¸çœ‹ä¸æ	Japanese

Data table containing information about different languages.



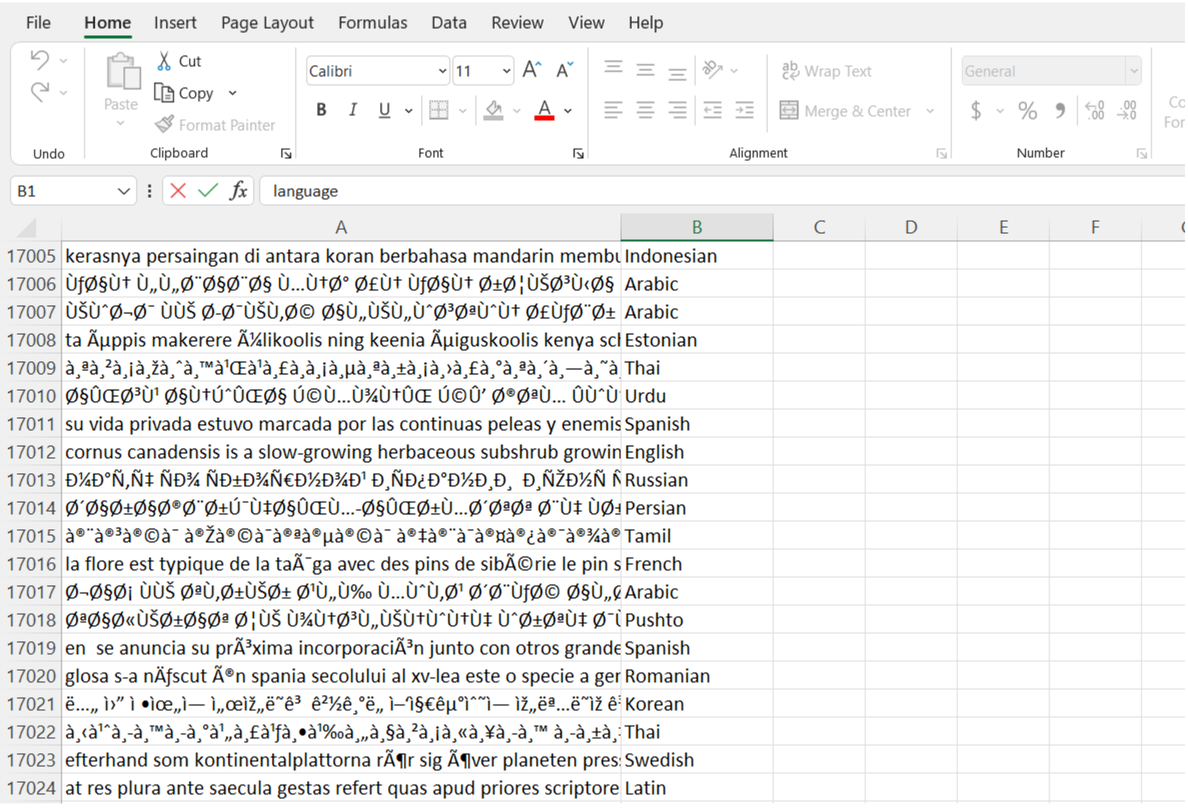
	A	B
10966	à¸—à¸µà¹ˆà¸ªà¸¸à¸” à¸ªà¸²à¸—à¸¸à¸¥à¸¸à¸" à¸®à¸´à¸™à¸"à¸µ à¤¦à¥€à¤ªà¤	Thai
10967	äº¬éƒ½å¸‚ãƒ¬ç-‰æŒ¢é™¢ãƒˆåœ¨å¾'æ-©ã™ãä¸"ã‚«ä½ç½®ã	Japanese
10968	à®‡à®•à¯à®•à¯à®¤à¯à®²à®¿à®²à¯ à®®à¯à®¤à®¤à¯à®®à®¾à®ª	Tamil
10969	ìš"í•˜ë„¤ìŠ¤ "ìš"í•œ" ì•¼ì½"ë·"ìŠ¤ ë„¤ì´ŠìŠ¤ ë„¤ëœëž€ë"œì–´	Korean
10970	ÑÑ‚Ñ€ÑƒÐºÑ‚ÑƒÑ€Ð° Ñ€Ð¾Ð²Ð°Ð½Ð½Ñ‹Ðµ Ð¿Ð¾Ñ€Ð¸ÑÑ‚Ñ‹Ðµ	Russian
10971	ê¸ë¬¸Ì•„ìŠ¤ì˜ ìž„ê¸°ìœ¼ëì œ ì°¨ ëŒ€€ „ìŸ ë‹¹ìœ¡ í…Œë ^ë"ˆìŠ¤ì„¸ì™	Korean
10972	Ø¨Ø±Ù† Ù‡Ù†Ø±ÙŠ ÙƒØ§Ù†Ù^Ù† Ø§Ù„Ø£Ù^Ù„Ø¨ÙŠØ³Ù…Ø¨Ø±	Arabic
10973	à®®à¯à®¤à®²à¯ à®µà®°à¯ˆ à®†à®£à¯à®Ÿà®†à®™à¯à®•à®¿à®²à¯‡	Tamil
10974	figlina fictilis minime cocta saepe fragmentis vel arena mixta impi	Latin
10975	fotopoulos describes inclusive democracy as "a new conception o	English
10976	os nomes usados na brasonaria lusÃ³fona para as cores e metais p	Portugese
10977	Ø¹Ù„Ù… Ù^Ø¹Ù…Ù„ Ú©ÛŒ Ø§Ø³ ÙÙ"ÛŒØ¯ Ø§Ù„Ù…Ø«Ø§Ù„ Ù^Ø	Urdu
10978	atravÃ©s desta ligaÃ§Ã£o de autocarro serve mais comodamente	Portugese
10979	a histÃ³ria da suÃÃ§a propriamente dita comeÃ§a a de agosto d	Portugese
10980	haec commentatio automatice praeparata res astronomicas collig	Latin
10981	à¸šà¸£à¸´à¸©à¸±à¸—à¸ˆà¸±à¸"à¸•à¸±à¹‰à¸‡à¸‚à¸¶à¹‰à¸™à¹€à¸¡à¸·à¹ˆà¸	Thai
10982	thariq bin ziyad lebih banyak dikenal sebagai penakluk spanyol k	Indonesian
10983	Ð¾Ñ‚Ð²ÐµÐ½ÑŒÑˆÐµ ÑÑ‚ÐµÐ¾ Ð¿Ñ‚ÑŒÐ´ÐµÑÑ, Ð»ÐµÑ‚, ÑÐ¿Ñ	Russian
10984	Ø³Ø¡Ù^Ø±Ù¾ÛŒÚ©Ù†Ø¯Ù‡ Ø¡Ù‡ Ù¾Ù‡ Ø¹Ø±Ø¨ÙŠ Ú˜Ø¨Ù‡ Ø¹Ù„Ù	Pushto

Data table containing information about different languages.



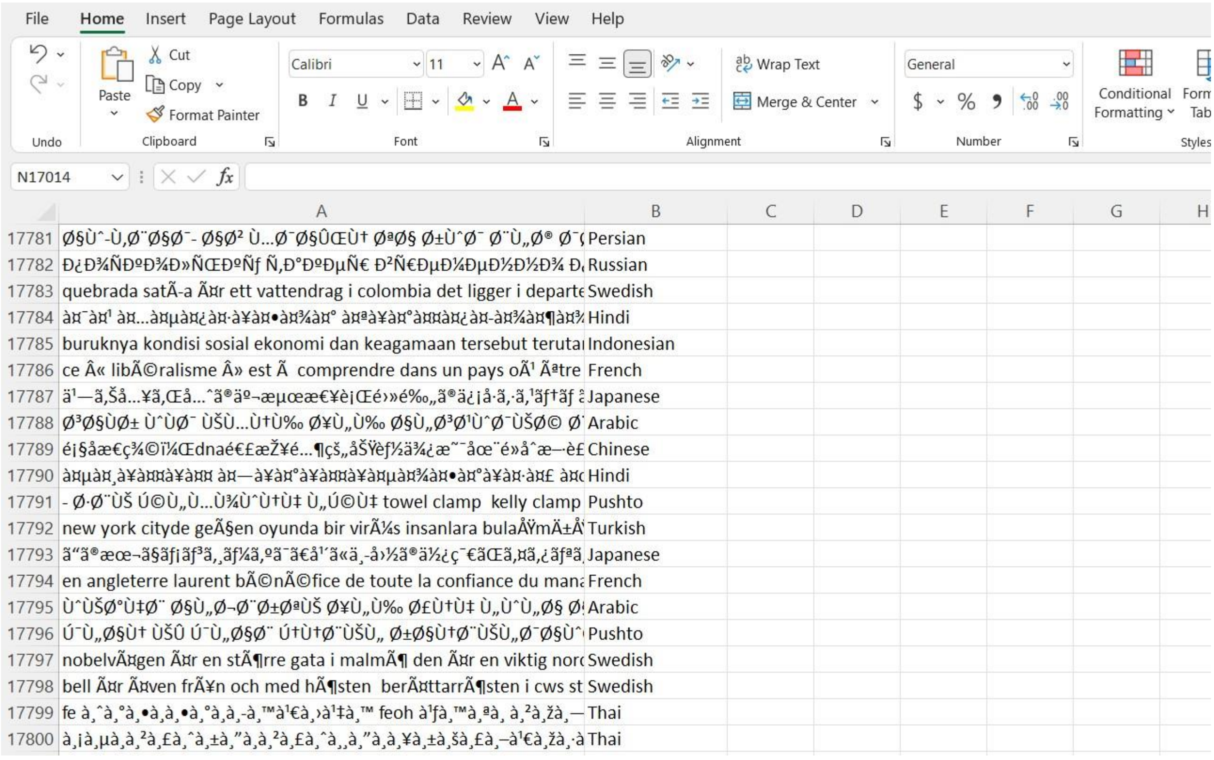
	A	B
11083	Ø¯ÙŠØ§ Ø¯ ØºÙŠØ±Øª Ù¾Ù‡ ÙƒÙˆÙ… Ø¯ ÚšÚÙˆ Ø¯ Ù¨Ú˜Ù†Û Ù¾	Pushto
11084	Ø§Ù¾Ø± Ù…ÛŒØ±ÛŒÙ† Ù¹ØµØ¤Ù† Ø´Ù¾ Ù…Ù¨Ù†Ù¹Ú¨Ù…Ø±Û	Urdu
11085	pont-lÃ©vÃªque est commune francicum incolarum anno praef	Latin
11086	Ù¾ÛŒÙ† Ø§ÙŠÙ¹Ø±ÙŠÛŒØ´ Ù†Ù„ Ù…ØµÙ†ÙÛŒÙ† Ú©ÛŒ Ø´Ø§	Urdu
11087	è©¹å§†æ£®ä¸€æ˜Ÿæœ€åŽæ‹ç‰©é¦†ç»™å-¦ç”Ÿä¸Šæ¬¡å®žè·µ	Chinese
11088	Ø®Ø§Ù†Ù˜Ø§Ø¯Ù‡ Ù„Ù˜Ú¯Ø§Ù† Ø¯Ø§ Ø¨Ø±ÛŒÙ¾Ø€ÛŒÙ‡Ø§Û Ù	Persian
11089	ÛŒÚ©Ù… Ø§Ú¯Ø³Øª Ø¡ Ú©Ø§ Ø³Ù˜Ø±Ø¬ Ú¯Ø±ÛÙ† Ø§ÛŒÚ© Ù	Urdu
11090	nascida no rio de janeiro Ã© formada em desenho industrial e co	Portugese
11091	à¤°à¤¾à¤·à¥à¤Ÿà¥à¤°à¥€à¤¯ à¤¸à¥à¤µà¤¯à¤‚à¤¸à¥‡à¤µà¤• à¤¸à¤‚	Hindi
11092	à¤¤à¤®à¤¾à¤® à¤…à¤ªà¥€à¤²à¥‹à¤‚ à¤µ à¤¦à¤²à¥€à¤²à¥‹à¤‚ à¤•	Hindi
11093	Ú©ÛÙ†Ø§ Ø¬Ø§Ø³Ø§ ÛÛ’ Ú©Û Ø±Ø³Ù˜Ù„ Ø§Ù„Ù„Û Ù†Û’ Ø¨Ú¾	Urdu
11094	Ù‡Ù† Ú©Ù‡ Ù„Ù‡ Ø¹Ø§Ù…Û Ø³Ø§Ø¯Ù‡ Ø§Ù˜ Ø±Ù˜Ø§Ù†Û Ù¾Ú	Pushto
11095	as guerras romano-latinas foram uma sÃ©rie de guerras gravadas	Portugese
11096	Ù¾Ù‡ Ú©Ù˜Ù†Ù†ÙŠÙ˜ Ø-ÙŠÙ˜Ø§Ù†Ø§Øª Ù˜ Ú©Ù‰ Ø³Ù¾Ù‰ Ù	Pushto
11097	Ð'Ð»Ð¸Ð½Ð°Ñ, ÑÐºÐ²Ð°Ñ‚Ð¾Ñ€Ð¸Ð°Ð»ÑŒÐ½Ñ‹Ð¹ Ð²Ð»Ð°Ð¶Ð½	Russian
11098	Ù˜Ù†Ù˜ Ø¹Ø¯Ø§Ø±Ø© Ø¹Ù† Ù‚Ø·Ø¹Ø© Ù…Ù† Ø§Ù„Ù…Ø§Ø´ Ø§	Arabic
11099	niidu-vÃµiliblikas elab kÃµikjal euroopas vÃ¤lja arvatud selle Ã¤Ã	Estonian
11100	à¸™à¸²à¸¢à¸•à¸³à¸£à¸§à¸ˆà¸¥à¸¹à¸à¸™à¸±à¸à¸‰à¸¸à¸-à¸—à¸ªà¸²à¸£à¸§à¸ˆ	Thai
11101	a anedota retrata com vivacidez os papÃ©is esperados pelos gÃª	Portugese
11102	veebruaril kinnitas kuninganna akti mille kohaselt teda ja ta jÃ¤r	Estonian

Data table containing information about different languages.



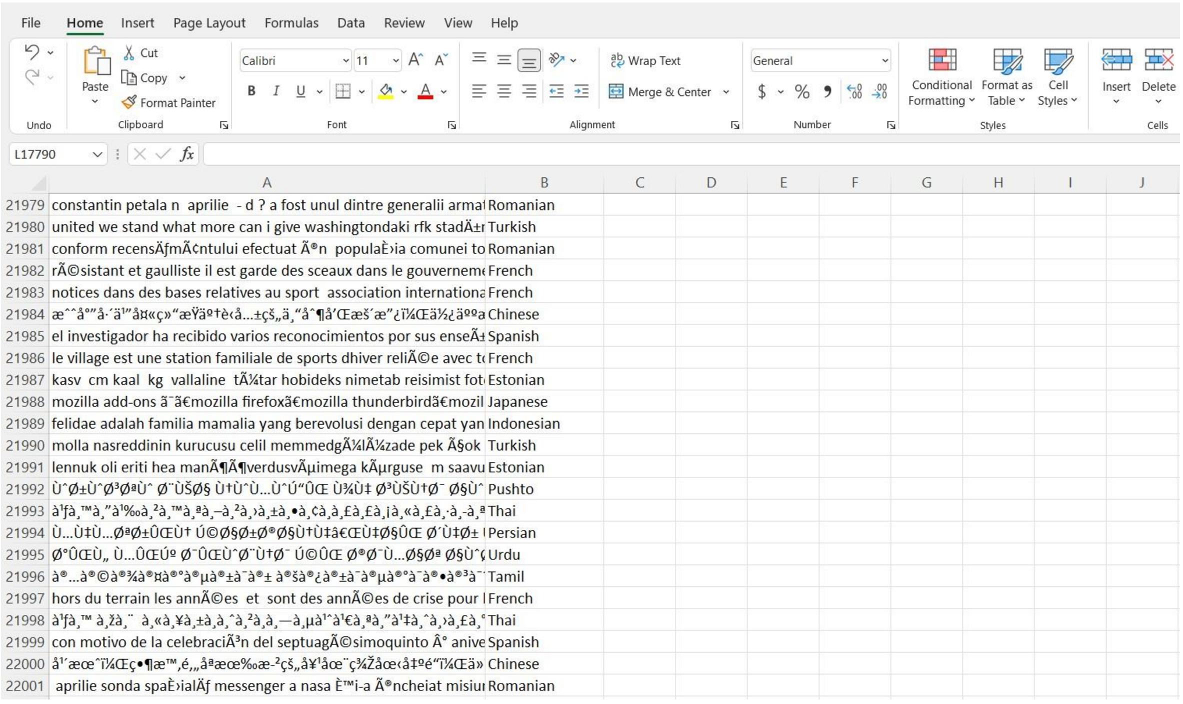
	A	B
17005	kerasnya persaingan di antara koran berbahasa mandarin membu	Indonesian
17006	ÙƒØ§Ù† Ù„Ù„ØºØ§Ø¨Ø§Øª Ù…Ù†Ø° Ø£Ù‚Ø¯Ù… Ø§Ù„Ø¹ØµÙˆØ±	Arabic
17007	ÙŠÙˆØ¬Ø¯ ÙŠÙŠ ØØ¯ÙŠÙ‚Ø© Ø§Ù„ÙŠÙˆØ³Ù†Ø¯Ø±Ø¨ Ø£ÙƒØ¨Ø±	Arabic
17008	ta Ãµppis makerere Ã¼likoolis ning keenia Ãµiguskoolis kenya sc	Estonian
17009	à¸ªà¸²à¸¡à¸žà¸±à¸™à¸˜à¹Œà¹€à¸£à¸·à¸à¸£à¸šà¸à¸²à¸£à¸±à¸›à¸£à¸°à¸ªà¸‡à¸„à¹Œà¸	Thai
17010	Ø§ÛŒØ³Ù¹ Ø§Ù†ÚˆÛŒØ§ Ú©Ù…Ù¾Ù†ÛŒ Ú©Û’ Ø®Ø·Ù… ÛÙˆÙ	Urdu
17011	su vida privada estuvo marcada por las continuas peleas y enemis	Spanish
17012	cornus canadensis is a slow-growing herbaceous subshrub growin	English
17013	Ð½Ð°Ñ‡ Ñ‚Ð¾ Ñ€Ð°Ð±Ð¾Ñ‚Ð°Ð» Ð¿Ñ€Ð¸Ð½Ð¸Ð¼Ð°Ð» Ñ	Russian
17014	Ø´Ø§Ø±Ø§Ø®Ø¨Ø±Ú¯Ø²Ø§Ø±ÛŒÙ…-Ø§ÛŒØ±Ù…Ø´ Ø®Ø·Ø± Ø¨Ù‡ Ù	Persian
17015	à®¨à®³à®©à¯ à®Žà®©à¯ à®ªà®µà®©à¯ à®ªà®•à¯	Tamil
17016	la flore est typique de la taÃ¯ga avec des pins de sibÃ©rie le pin s	French
17017	Ø¬Ø§Ø¡ ÙÙŠ Ø±ÙˆØ§ÙŠØ© Ø¹Ù„Ù‰ Ù…Ù˜Ù‚Ø¹ Ø´Ø¨ÙƒØ© Ø§Ù„	Arabic
17018	ØªØ§Ø«ÙŠØ±Ø§Øª Ø¦ÙŠ Ù¾Ù†Ø³Ù„ÙŠÙ†Ù˜Ù†Ù‡ Ù˜Ø±Ø±Ù‡ Ø¯	Pushto
17019	en se anuncia su prÃ³xima incorporaciÃ³n junto con otros grande	Spanish
17020	glosa s-a nÄƒscut Ã®n spania secolului al xv-lea este o specie a ger	Romanian
17021	ë…„ ì" ì •ìœ„ì— ì„œìž„ë˜ê³ ê²½ì¢…ì—ê²Œ ì—¬ì„¯ì—	Korean
17022	à¸‹à¹‰à¸³à¸™à¸°à¸¹à¸„à¸£à¹ƒà¸•à¹‰à¸„à¸¡à¸§à¸¥à¸«à¸¥à¸±à¸™ à¸à¸	Thai
17023	efterhand som kontinentalplattorna rÃ¶r sig Ã¶ver planeten pres	Swedish
17024	at res plura ante saecula gestas refert quas apud priores scriptore	Latin

Data table containing information about different languages.



	A	B
17781	Ø§Ù^-Ù,Ø¨Ø§Ø¯- Ø§Ø² Ù…Ø¯Ø§ÛŒÙ† ØªØ§ Ø±Ù^Ø¯ Ø¨Ù„Ø® Ø¯	Persian
17782	Ð¿Ð¾ÑÐºÐ¾Ð»ÑŒÐºÑƒ Ñ‚Ð°ÐºÐµÑ€ Ð²Ñ€ÐµÐ¼ÐµÐ½Ð½Ð¾ Ð	Russian
17783	quebrada satÃ-a Ã¤r ett vattendrag i colombia det ligger i departe	Swedish
17784	à¤¯à¤¹ à¤…à¤µà¤¿à¤·à¥à¤•à¤¾à¤° à¤ªà¥à¤°à¤¤à¤¿à¤à¤¾à¤¶à¤¾	Hindi
17785	buruknya kondisi sosial ekonomi dan keagamaan tersebut terutai	Indonesian
17786	ce Â« libÃ©ralisme Â» est Ã comprendre dans un pays oÃ¹ Ãªtre	French
17787	ä¹—ã‚Šå…¥ã‚Œå…ˆã®äº¬æµœæ€¥è¡Œé›»é‰„ã®ä¿¡å·ã‚·ã‚¹ãƒ†ãƒ ã	Japanese
17788	Ø³Ø§ÙØ± Ù^ÙØ¯ ÙŠÙ…Ù†Ù‰ Ø¥Ù„Ù‰ Ø§Ù„Ø³Ø¹Ù^Ø¯ÙŠØ© Ø	Arabic
17789	é¡§åæ€ç¾©ï¼Œdnaé€£æŽ¥é…¶çš„åŠŸèƒ½ä¾¿æ˜¯åœ¨é»å^æ–·è£	Chinese
17790	à¤µà¤¸à¥à¤¤à¥à¤¤ à¤—à¥à¤°à¥à¤¤à¥à¤µà¤¾à¤•à¤°à¥à¤·à¤£ à¤	Hindi
17791	- Ø·Ø¨ÙŠ Ú©Ù„Ù…Ù¾Ù^Ù†Ù‡ Ù„Ú©Ù‡ towel clamp kelly clamp	Pushto
17792	new york cityde geÃ§en oyunda bir virÃ¼s insanlara bulaÅŸmÄ±Å	Turkish
17793	ã“ã®æœ¬ã§ãƒ¡ãƒ³ã‚¸ãƒ¼ã‚ºã¯ã€å¹´ã«ä¸-å›½ã®ä½¿ç¯€ã‚Œã‚¤ã‚¿ãƒªã	Japanese
17794	en angleterre laurent bÃ©nÃ©fice de toute la confiance du mana	French
17795	Ù^ÙŠØ°Ù‡Ø¨ Ø§Ù„Ø¬Ø¨Ø±ØªÙŠ Ø¥Ù„Ù‰ Ø£Ù†Ù‡ Ù„Ù^Ù„Ø§ Ø	Arabic
17796	Ú¯Ù„Ø§Ù† ÙŠÛ Ú¯Ù„Ø§Ø¨ Ú†Ù†Ø¨ÙŠÙ„ Ø±Ø§Ù†Ø¨ÙŠÙ„Ø¯Ø§Ù^	Pushto
17797	nobelvÃ¤gen Ã¤r en stÃ¶rre gata i malmÃ¶ den Ã¤r en viktig nor	Swedish
17798	bell Ã¤r Ã¤ven frÃ¥n och med hÃ¶sten berÃ¤ttarrÃ¶sten i cws st	Swedish
17799	fe à¸ˆà¸°à¸•à¸à¸•à¸°à¸à¸-à¸™à¹€à¸›à¹‡à¸™ feoh à¹ƒà¸™à¸ªà¸ à¸²à¸žà¸—	Thai
17800	à¸¡à¸µà¸à¸²à¸£à¸ˆà¸±à¸”à¸à¸²à¸£à¸ˆà¸¸à¸”à¸à¸¥à¸±à¸šà¸£à¸–à¹€à¸žà¸·à	Thai

Data table containing information about different languages.



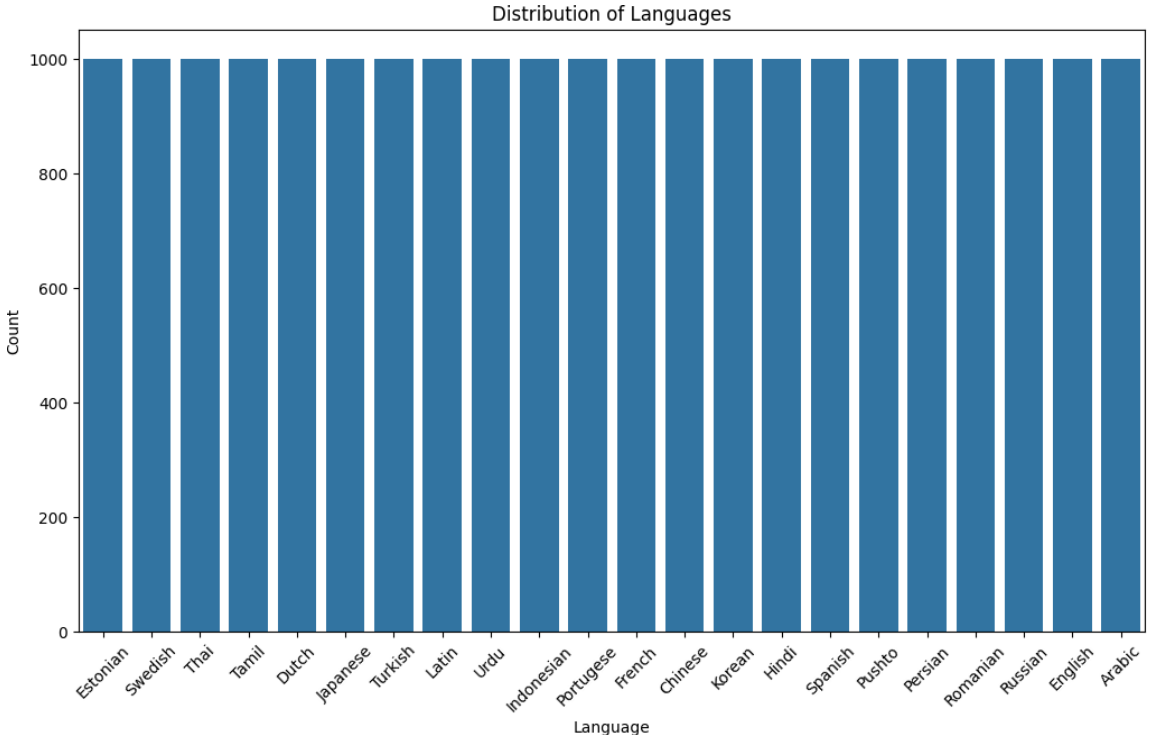
	A	B
21979	constantin petala n aprilie - d ? a fost unul dintre generalii armat	Romanian
21980	united we stand what more can i give washingtondaki rfk stadÄ±r	Turkish
21981	conform recensÄƒmÃ¢ntului efectuat Ã®n populaÈ›ia comunei to	Romanian
21982	rÃ©sistant et gaulliste il est garde des sceaux dans le gouverneme	French
21983	notices dans des bases relatives au sport association internationa	French
21984	æ˜¯åº”å·´ä¹”å¤«ç»“æŸäº†è‹å…±çš„ä¸“å^¶å'Œæš´æ”¿ï¼Œä½¿äººa	Chinese
21985	el investigador ha recibido varios reconocimientos por sus enseÃ±	Spanish
21986	le village est une station familiale de sports dhiver reliÃ©e avec t	French
21987	kasv cm kaal kg vallaline tÃ¼tar hobideks nimetab reisimist fot	Estonian
21988	mozilla add-ons ã¯ã€mozilla firefoxã€mozilla thunderbirdã€mozil	Japanese
21989	felidae adalah familia mamalia yang berevolusi dengan cepat yan	Indonesian
21990	molla nasreddinin kurucusu celil memmedgÃ¼lÃ¼zade pek Ã§ok	Turkish
21991	lennuk oli eriti hea manÃ¶Ã¶verdusvÃµimega kÃµrguse m saavu	Estonian
21992	Ù^Ø±Ù^Ø³ØªÙ^ Ø¨ÙŠØ§ Ù†Ù^Ù…Ù^Ú“ÛŒ Ù¾Ù‡ Ø³ÙŠÙ†Ø¯ Ø§Ù^	Pushto
21993	à¹ƒà¸™à¸”à¹‰à¸²à¸™à¸ªà¸–à¸²à¸›à¸±à¸•à¸¢à¸à¸£à¸£à¸¡à¸«à¸£à¸·à¸-à¸ª	Thai
21994	Ù…Ù‡Ù…ØªØ±ÛŒÙ† Ú©Ø§Ø±Ø®Ø§Ù†Ù‡â€ŒÙ‡Ø§ÛŒ Ø´Ù‡Ø±	Persian
21995	Ø°ÛŒÙ„ Ù…ÛŒÚº Ø¯ÛŒÙ^Ø¨Ù†Ø¯ Ú©ÛŒ Ø®Ø¯Ù…Ø§Øª Ø§Ù^	Urdu
21996	à®…à®©à®¾à®¤à®°à®µà®±à¯à®± à®šà®¿à®±à¯à®µà®°à¯à®•à®³à¯	Tamil
21997	hors du terrain les annÃ©es et sont des annÃ©es de crise pour l	French
21998	à¹ƒà¸™ à¸žà¸¨ à¸«à¸¥à¸±à¸à¸ˆà¸²à¸à¸—à¸µà¹ˆà¹€à¸ªà¸”à¹‡à¸ˆà¸›à¸£à¸	Thai
21999	con motivo de la celebraciÃ³n del septuagÃ©simoquinto Â° anive	Spanish
22000	å¹´æœˆï¼Œç•¶æ™‚é„å³æœ‰æ-²çš„å¥¹åœ¨ç¾Žåœ‹å‡ºé“ï¼Œä»	Chinese
22001	aprilie sonda spaÈ›ialÄƒ messenger a nasa È™i-a Ã®ncheiat misiu	Romanian

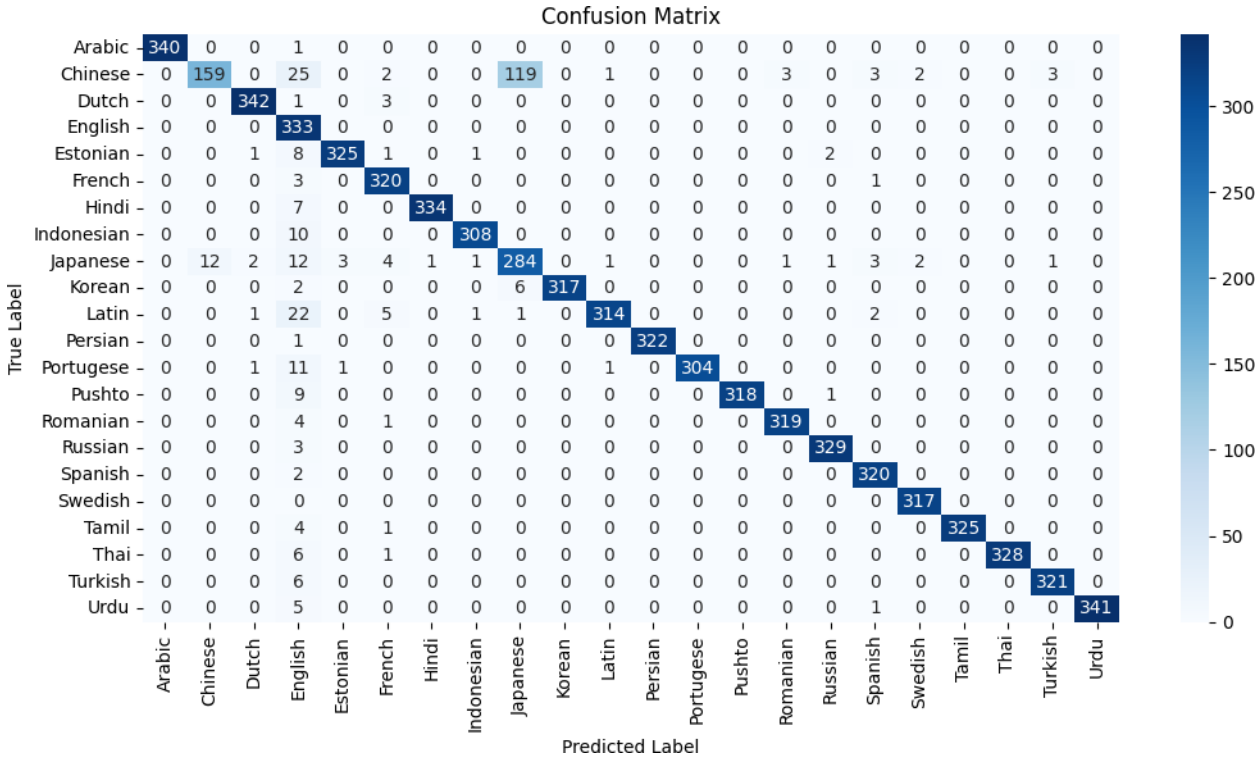
OUTPUT SCREENS

OUTPUT SCREENS

```
Dataset information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22000 entries, 0 to 21999
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Text      22000 non-null  object
 1   language  22000 non-null  object
dtypes: object(2)
memory usage: 343.9+ KB
None
First few rows of the dataset:
Null values check:
Text        0
language    0
dtype: int64
```

```
 present Languages in dataset:
language
Estonian      1000
Swedish       1000
English       1000
Russian       1000
Romanian      1000
Persian       1000
Pushto        1000
Spanish       1000
Hindi         1000
Korean        1000
Chinese       1000
French        1000
Portugese     1000
Indonesian    1000
Urdu          1000
Latin         1000
Turkish       1000
Japanese      1000
Dutch         1000
Tamil         1000
Thai          1000
Arabic        1000
Name: count, dtype: int64
Number of unique languages:
22
```


Distribution of Languages
Count
Language
0
200
400
600
800
1000
Estonian
Swedish
Thai
Tamil
Dutch
Japanese
Turkish
Latin
Urdu
Indonesian
Portugese
French
Chinese
Korean
Hindi
Spanish
Pushto
Persian
Romanian
Russian
English
Arabic


Confusion Matrix
True Label
Predicted Label
Arabic 340 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Chinese 0 159 0 25 0 2 0 0 119 0 1 0 0 0 3 0 3 2 0 0 3 0
Dutch 0 0 342 1 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
English 0 0 0 333 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Estonian 0 0 1 8 325 1 0 1 0 0 0 0 0 0 0 2 0 0 0 0 0 0
French 0 0 0 3 0 320 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
Hindi 0 0 0 7 0 0 334 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Indonesian 0 0 0 10 0 0 0 308 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Japanese 0 12 2 12 3 4 1 1 284 0 1 0 0 0 1 1 3 2 0 0 1 0
Korean 0 0 0 2 0 0 0 0 6 317 0 0 0 0 0 0 0 0 0 0 0 0
Latin 0 0 1 22 0 5 0 1 1 0 314 0 0 0 0 0 2 0 0 0 0 0
Persian 0 0 0 1 0 0 0 0 0 0 0 322 0 0 0 0 0 0 0 0 0 0
Portugese 0 0 1 11 1 0 0 0 0 0 1 0 304 0 0 0 0 0 0 0 0 0
Pushto 0 0 0 9 0 0 0 0 0 0 0 0 0 318 0 1 0 0 0 0 0 0
Romanian 0 0 0 4 0 1 0 0 0 0 0 0 0 0 319 0 0 0 0 0 0 0
Russian 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 329 0 0 0 0 0 0
Spanish 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 320 0 0 0 0 0
Swedish 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 317 0 0 0 0
Tamil 0 0 0 4 0 1 0 0 0 0 0 0 0 0 0 0 0 0 325 0 0 0
Thai 0 0 0 6 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 328 0 0
Turkish 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 321 0
Urdu 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 341
Arabic
Chinese
Dutch
English
Estonian
French
Hindi
Indonesian
Japanese
Korean
Latin
Persian
Portugese
Pushto
Romanian
Russian
Spanish
Swedish
Tamil
Thai
Turkish
Urdu
0
50
100
150
200
250
300

```
Classification Report:
              precision    recall  f1-score   support

      Arabic       1.00      1.00      1.00       341
     Chinese       0.93      0.50      0.65       317
       Dutch       0.99      0.99      0.99       346
     English       0.70      1.00      0.82       333
    Estonian       0.99      0.96      0.97       338
      French       0.95      0.99      0.97       324
       Hindi       1.00      0.98      0.99       341
  Indonesian       0.99      0.97      0.98       318
    Japanese       0.69      0.87      0.77       328
      Korean       1.00      0.98      0.99       325
       Latin       0.99      0.91      0.95       346
     Persian       1.00      1.00      1.00       323
   Portugese       1.00      0.96      0.98       318
      Pushto       1.00      0.97      0.98       328
    Romanian       0.99      0.98      0.99       324
     Russian       0.99      0.99      0.99       332
     Spanish       0.97      0.99      0.98       322
     Swedish       0.99      1.00      0.99       317
       Tamil       1.00      0.98      0.99       330
        Thai       1.00      0.98      0.99       335
     Turkish       0.99      0.98      0.98       327
        Urdu       1.00      0.98      0.99       347

    accuracy                           0.95      7260
   macro avg       0.96      0.95      0.95      7260
weighted avg       0.96      0.95      0.95      7260

Model Accuracy: 0.953168044077135
Enter a Text: the
Predicted Language: ['English']
```

CONCLUSION

CONCLUSION

The rise of technology in the modern world has also given rise to increased requirements which justify the development taking place around us every day. Natural Language Processing and Language Detection, here, give rise to wider as well as broader scopes which can make tasks easier for human beings and can help them recognize texts in a much easier, better and systematic manner, hence, making technical work easier for them with the use of statistical methods.

Therefore, an attempt has been put forward by us for creating such a Language Detection model with the help of Natural Language Processing that can solve Language Detection problems and can help us in identifying text easily and aptly with the help of appropriate and efficient methods as it is very important and useful in today's world and justifies fairly the usage of words and linguistics in the body of the content provided in various documents.

BIBLIOGRAPHY

BIBLIOGRAPHY

- https://scikit-learn.org/stable/modules/naive_bayes.html
- https://machinelearningmastery.com/
- https://machinelearningmastery.com/
- https://huggingface.co/models?search=language%20detection
- https://ieeexplore.ieee.org/document/9225610
- https://aclanthology.org/P12-3005
- https://www.kaggle.com/code/prashant111/naive-bayes-classifier-in-python

REFERENCES

REFERENCES

1. Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2018. A Survey of the Usages of Deep Learning in Natural Language Processing. 1, 1 (July 2018), 35 pages.
2. ROBERT DALE. "The commercial NLP Landscape in 2017", Article in Natural Language Engineering, July 2017
3. ACL 2018: 56th Annual Meeting of Association for Computational Linguistics https://acl2018.org
4. Predictive Analytics Today: www.predictiveanalyticstoday.com[accessed in Dec 2018]
5. Ali Shatnawi, Ghadeer Al-Bdour, Raffi Al-Qurran and Mahmoud Al-Ayyoub 2018. A Comparative Study of Open Source Deep Learning Frameworks. 2018 9th International Conference on Information and Communication Systems (ICICS)
6. Intelligent automation: Making cognitive real Knowledge Series I Chapter 2. 2018, EY report.
7. Jacques Bughin, Eric Hazan, SreeRamaswamy, Michael Chui , TeraAllas, Peter Dahlström, Nicolaus Henke, Monica Trench, 2017. MGI ARTIFICIAL INTELLIGENCE THE NEXT DIGITAL FRONTIER? McKinsey & Company McKinsey & Company report July 2017
8. Svetlana Sicular, Kenneth Brant 2018, Hype Cycle for Artificial Intelligence, 2018 Gartner report July 2018.
9. Radiuk, Pavlo, Pavlova, Olga , Hrypynska, Nadiia .An ensemble machine learning approach for Twitter sentiment analysis. Issue Date: 17-Jul-2022.
10. Luca Barbaglia Sergio Consoli , Sebastiano Manzan , Luca Tiozzo Pezzoli, Elisa Tosetti. Sentiment Analysis of Economic Text: A Lexicon-based Approach. ,23 Pages Posted: 13 May 2022 , Date Written: May 11, 2022.