Python and Data Mining (INFO-501)

Project Title:
Article Topic Classification

Members:

Maneendra Burrannagari
Sathya Sai Ram Prabhala
Sumedh Kulkarni
Luoxi Tang

I have done this assignment completely on my own. I have not copied it, nor have I given my solution to anyone else. I understand that if I am involved in plagiarism or cheating, I will have to sign an official form that I have cheated and that this form will be stored in my official university record. I also understand that I will receive a grade of 0 for the involved assignment and my grade will be reduced by one level (e.g., from A to A- or from B+ to B) for my first offense, and that I will receive a grade of “F” for the course for any additional offense of any kind.

Acknowledgement

We extend our sincere gratitude to Dr. Professor Sriniwas Pandey, whose guidance and expertise have been invaluable throughout the duration of the course. Dr. Professor Sriniwas Pandey’s commitment to fostering a dynamic learning environment and his unwavering support have significantly enriched our understanding of tools for data science, culminating in the successful completion of the Quiz Application project.
Special appreciation is also extended to our Teaching Assistant, Danila Rozhevskii. Their dedication to assisting students, providing timely feedback, and addressing queries have played a crucial role in our academic journey. They have created a positive and collaborative atmosphere that facilitated our learning and project development.
The collaborative success of the Article Topic Classification project is owed to the exceptional contributions of our team Maneendra Burrannagari, Sumedh Kulkarni,Luoxi Tang, and Sathya Sai Ram Prabhala. Each team member played a pivotal and equal role, showcasing individual strengths that seamlessly merged into a cohesive and efficient unit. Maneendra's attention to detail, Sumedh's strategic problem-solving, Luoxi's creative design flair, and Sathya's systematic project management were instrumental in navigating the complexities of web application development. The team’s collective efforts and effective collaboration were paramount to the project's overall success.

Table of contents

Repository Link

https://github.com/ltang24/article-topic-classifer

Introduction

Background

Need for an Article Topic Classification

Article topic classification models find application in various fields, enhancing content organization and accessibility. They are crucial for search engines, ensuring relevant and accurate search results, thereby improving user experience. In online platforms, these models assist in personalized content recommendations, increasing user engagement. Furthermore, topic classification is vital for information retrieval, aiding researchers, students, and professionals in accessing specific content efficiently.

In business, effective topic classification supports marketing efforts by optimizing SEO strategies, driving web traffic. Ethically, these models address concerns by providing transparency and mitigating bias in content organization. Additionally, they contribute to the development of intelligent systems, such as chatbots or virtual assistants, by enabling them to understand and respond to diverse topics. Overall, article topic classification models play a pivotal role in streamlining information, fostering accessibility, and shaping the future of content organization across various domains.

Role of an Article Topic Classification Model

Article topic classification plays a pivotal role in organizing and enhancing content accessibility. By systematically categorizing articles based on subject matter, it streamlines information retrieval, improving user experience on websites and databases. This process is integral to effective search engine optimization (SEO), ensuring that content is accurately understood by search algorithms and boosting online visibility. Additionally, topic classification facilitates personalized content recommendations, tailoring user experiences based on individual preferences. It contributes to the development of machine learning and AI applications, enabling automated categorization and supporting ethical content organization. Overall, article topic classification is a fundamental mechanism that not only structures information but also plays a crucial role in optimizing search, personalization, and the ethical presentation of diverse content.

Significance of an Article Topic Classification Model

The significance of an article topic classification project lies in its ability to bring order and efficiency to the vast landscape of digital information. In an era of information overload, where content is continuously generated, proper classification serves as a navigational guide for users. It enhances content organization on websites and databases, ensuring that users can swiftly locate and access specific information without being overwhelmed.

Search engines heavily rely on topic classification to interpret and index content accurately. This significantly impacts Search Engine Optimization (SEO), as well-classified content improves a website's visibility in search results, ultimately driving traffic.

Furthermore, the project's importance extends to user experience. Effective topic classification enables personalized content recommendations, contributing to user engagement and satisfaction. By understanding user preferences and behavior, platforms can offer tailored suggestions, creating a more enjoyable and relevant browsing experience.

In the realm of artificial intelligence and machine learning, a well-executed article topic classification project becomes a cornerstone. It enables the training of models for automated content categorization, supporting applications like chatbots, sentiment analysis, and virtual assistants.

Ethically, the project addresses concerns related to biased content presentation. Transparent and fair classification mitigates the risk of reinforcing stereotypes or promoting misinformation, fostering an environment of trust and credibility.

In summary, an article topic classification project is significant for its impact on content organization, search engine visibility, user experience, AI applications, and ethical content presentation. It serves as a foundational element in managing and harnessing the wealth of digital information in a way that is both user-friendly and aligned with ethical standards.

Objectives

The objective of our project is to build an article classification model, which predicts under which category the article falls under based on existing categories such as Technology, Business, Politics, Entertainment,Sport

Scope

Data Collection

- Utilize the BBC dataset as the primary source for training and validating the article topic classification model.
- Explore additional datasets or sources for diverse news articles to enhance the model's generalization beyond the initial training data.

Model Training

- Cleanse and preprocess the data to handle challenges such as missing values, outliers, and text data complexities.
- Implement techniques like tokenization,lemmatization and removal of stop words to prepare the data for model training.

Data Mining Techniques

- Implement machine learning algorithms like Support Vector Machines (SVM), Logistic Regression,K-Nearest Neighbors (KNN), and Decision Tree, for article topic classification.
- Experiment with parameter tuning to optimize the performance of each algorithm.

Integration with Other News Sources

- Extend the model's capability to read and classify articles from diverse news sources beyond the BBC dataset.
- Ensure the model adapts well to different writing styles and effectively categorizes articles from varied domains.

Documentation

- Create comprehensive documentation detailing the model architecture, training process, and integration steps for each algorithm. Provide a README on how to interact with the system and understand the predictions.

Implementation

Key Features

Dataset Utilization

- Utilizing a rich and varied dataset sourced from BBC, our model was trained and validated to achieve a comprehensive understanding of news topics. The dataset covered a broad spectrum, encompassing politics, science, technology, business, and sports, ensuring the model's proficiency in classifying diverse content.

Machine Learning Techniques

- To enhance the model's accuracy, we deployed a combination of robust machine learning techniques, including K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVM), and Decision Trees.

Model Pickling

- The trained model was serialized and stored using pickling, streamlining the storage and retrieval processes. This approach facilitates efficient model deployment without the need for repeated training, ensuring quick and seamless usage, and contributing to a more responsive application.

Visualization Techniques

- Incorporating visualization techniques, our application provides users with a deeper understanding of the classified articles. Word clouds highlight the most prominent terms within articles, offering a visual summary of key themes. Bar graphs visually represent the distribution of articles across different categories, presenting a clear overview of the dataset's composition.

Continuous Improvement

- A robust mechanism for continuous model improvement has been established. This involves user feedback integration and adaptive measures to evolving language use and emerging news topics. This commitment ensures the application evolves over time, staying effective, relevant, and aligned with user expectations and the dynamic nature of news content.

Components

Tools Utilized

Scikit-learn (sklearn)

Leveraged the powerful machine learning library Scikit-learn for implementing and fine-tuning the machine learning algorithms (K-Nearest Neighbors, Logistic Regression, SVM, Decision Trees). Scikit-learn provided a comprehensive set of tools for model training, evaluation, and optimization.

NLTK

NLTK stands for Natural Language Toolkit. It is a powerful Python library for working with human language data, particularly in the field of natural language processing (NLP).

Command Line Interface (CLI)

Integrated a Command Line Interface for efficient execution of various tasks, such as data preprocessing, model training, and serialization. CLI offered flexibility and automation in managing the project components.

Anaconda Navigator

Utilized Anaconda Navigator as a comprehensive platform for package management, environment setup, and project organization. Anaconda streamlined the installation of libraries, ensuring compatibility and consistency across the development environment.

Spyder

Spyder is an open-source integrated development environment (IDE) for Python programming. It provides a user-friendly interface with powerful tools for code editing, debugging, and data exploration, making it popular among data scientists and engineers.

Jupyter Notebook

Implemented Jupyter Notebooks as an interactive and visual environment for data exploration, model development, and result visualization. Jupyter provided a collaborative and iterative workspace, facilitating seamless integration with Scikit-learn.

Integration and Workflow

Scikit-learn Integration

Integrated Scikit-learn seamlessly into the Jupyter Notebook environment for developing and training machine learning models. This integration allowed for interactive model development and quick experimentation.

CLI for Task Automation

Leveraged the Command Line Interface for automating repetitive tasks, such as data preprocessing and model training. CLI enhanced efficiency and provided a streamlined workflow for project management.

Anaconda Navigator for Environment Management

Utilized Anaconda Navigator to manage project dependencies, ensuring a consistent and reproducible environment. Anaconda's environment management capabilities facilitated smooth collaboration among team members.

Jupyter Notebook for Iterative Development

Incorporated Jupyter Notebooks for iterative development, enabling step-by-step code execution, visualizations, and easy collaboration. Jupyter's interactive interface facilitated data exploration and model fine-tuning.

Benefits

Efficiency and Consistency

The integration of Scikit-learn, CLI, Anaconda Navigator, and Jupyter Notebook provided an efficient and consistent development environment, fostering smooth collaboration and reducing potential compatibility issues.

Interactivity and Visualization

Jupyter Notebook's interactive nature allowed for real-time exploration and visualization of data and model outputs, enhancing the understanding of the machine learning process.

Automation and Streamlined Workflow

The Command Line Interface automated routine tasks, reducing manual effort and ensuring a streamlined workflow. This automation contributed to the efficiency and reproducibility of the project.

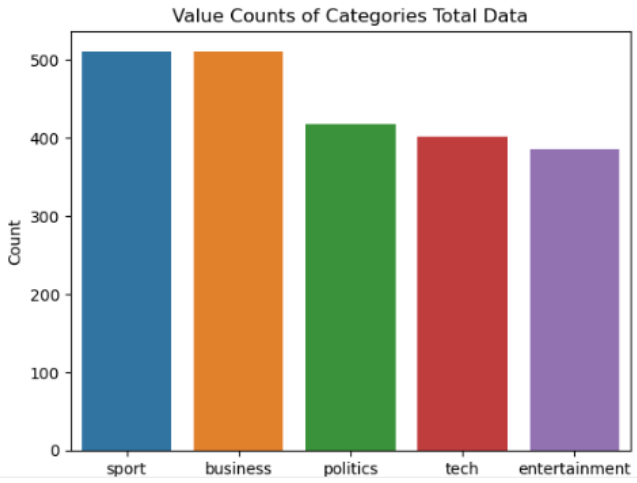
Environment Management

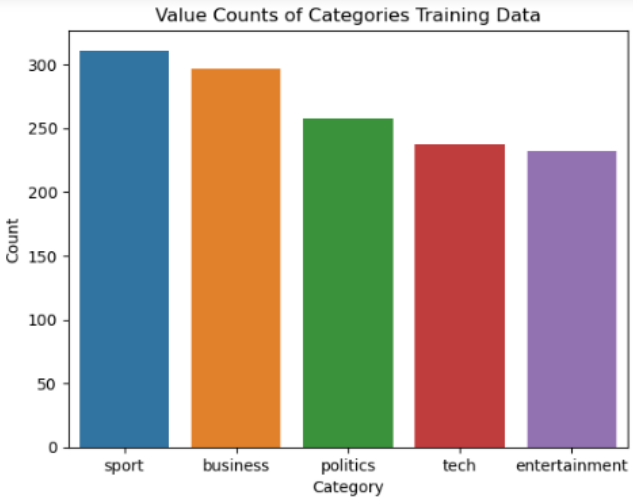
Anaconda Navigator's environment management capabilities ensured that the project's dependencies were well-managed, minimizing compatibility concerns and facilitating a cohesive development environment

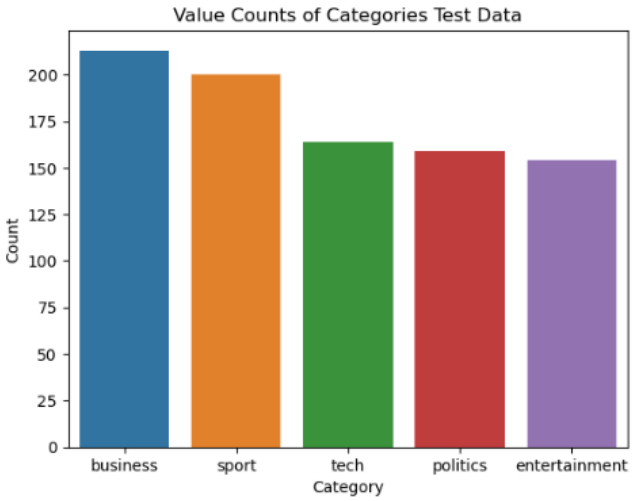
Snapshots

```
In [22]: 1 EDA(data, training_data, testing_data)

 Dataframe of the entire data=
 <bound method NDFrame.head of       Category   Id                              title  \
0     business    1  Ad sales boost Time Warner profit
1     business    2   Dollar gains on Greenspan speech
2     business    3  Yukos unit buyer faces loan claim
3     business    4  High fuel prices hit BA's profits
4     business    5  Pernod takeover talk lifts Domecq
...        ...  ...                                ...
2220      tech  397   BT program to beat dialler scams
2221      tech  398    Spam e-mails tempt net shoppers
2222      tech  399            Be careful how you code
2223      tech  400    US cyber security chief resigns
2224      tech  401   Losing yourself in online gaming

                                            Content
0     Quarterly profits at US media giant TimeWarner...
1     The dollar has hit its highest level against t...
2     The owners of embattled Russian oil giant Yuko...
3     British Airways has blamed high fuel prices fo...
4     Shares in UK drinks and food firm Allied Domec...
...                                                 ...
2220  BT is introducing two initiatives to help beat...
2221  Computer users across the world continue to ig...
2222  A new European directive could put software wr...
2223  The man making sure US computer networks are s...
2224  Online role playing games are time-consuming, ...

[2225 rows x 4 columns]>
 Dataframe of training data=
 <bound method NDFrame.head of           Category   Id                              title  \
41         business   42   UK Coal plunges into deeper loss
1387          sport   75    Gardener wins double in Glasgow
1861           tech   38  Movie body hits peer-to-peer nets
1448          sport  136         Iranian misses Israel match
2015           tech  192   Mobile gaming takes off in India
...             ...  ...                                ...
1638          sport  326           Robinson answers critics
1095       politics  200   Mallon wades into NE vote battle
1130       politics  235    Lib Dems' new election PR chief
1294       politics  399   Tories reject rethink on axed MP
860   entertainment  351   Women in film 'are earning less'

                                            Content
41    share uk coal fallen mining group reported los...
1387  britain jason gardener enjoyed double success ...
1861  movie industry struck file sharing network ano...
1448  iranian striker vahid hashemian travel israel ...
2015  gaming move one fastest growing activity among...
...                                                 ...
1638  england captain jason robinson rubbished sugge...
1095  middlesbrough mayor ray mallon drafted boost y...
```

```
[1335 rows x 4 columns]>
 Dataframe of testing data=
 <bound method NDFrame.head of       Category   Id                                title  \
414   business  415     UK house prices dip in November
420   business  421   LSE 'sets date for takeover deal'
1644     sport  332     Harinordoquy suffers France axe
416   business  417   Barclays shares up on merger talk
1232  politics  337    Campaign 'cold calls' questioned
...        ...  ...                                 ...
1509     sport  197             FA probes crowd trouble
1961      tech  138  Firefox browser takes on Microsoft
2090      tech  267     The year search became personal
1477     sport  165    Unclear future for striker Baros
1888      tech   65    Yahoo celebrates a decade online

                                                Content
414   uk house price dipped slightly november office...
420   london stock exchange lse planning announce pr...
1644  number eight imanol harinordoquy dropped franc...
416   share uk banking group barclays risen monday f...
1232  labour conservative still telephoning million ...
...                                                 ...
1509  fa take action trouble marred wednesday carlin...
1961  microsoft internet explorer serious rival long...
2090  odds fire browser go straight favourite search...
1477  liverpool forward milan baros uncertain whethe...
1888  yahoo one net iconic company celebrating th an...

[890 rows x 4 columns]>
```







Category

```
Missing Value Count :
Category    0
Id          0
title       0
Content     0
dtype: int64

Mean word count:
business: 324.34313725490193
entertainment: 325.96891191709847
politics: 449.84892086330933
sport: 325.2172211350294
tech: 498.2942643391521
```

```
vectorizer = TfidfVectorizer(stop_words='english')
```

```
results = building_results(vectorizer, training_data, testing_data)
```

```
{'accuracy': 0.9393258426966292, 'Precision': 0.9405689595906287}
{'accuracy': 0.9775280898876404, 'Precision': 0.9784881277716613}
{'accuracy': 0.9775280898876404, 'Precision': 0.9775079390061823}
{'accuracy': 0.8078651685393259, 'Precision': 0.8089782697851354}
```

```
save_best_model(results)
```

```
Best Model: Logistic Regression
```

```
best_trained_model = load_best_model()
```

```
user_interaction(best_trained_model, vectorizer, data)
```

```
Enter some text: Neeraj Chopra on why youngsters should opt for patience and process over shortcuts to fleeting success, his fa
vourite fast bowler, why he likes anonymity when not competing and his need to speak up for fellow athletes. This Idea Exchange
was moderated by Associate Editor Nihal Koshie.  Nihal Koshie: This year, you celebrated Diwali with your family. Were there ch
eat meals? Neeraj Chopra: It was very good to spend Diwali with the family. Cousins studying in Dehradun had come down. We are
never together. But this year we had meals together, spent hours chatting. It was free time spent with family after many years.
Usually, the family sleeps by 8-9 pm. I would keep them awake till midnight. So they were a little hassled, but it was fun.
I've controlled a lot this time (eating sweets). Last time, I was eating everything and my weight had increased. Now, I can con
trol a lot even if I'm not training. If I eat a bit more in one meal, I ensure I skip the next. I'm also training a little. Las
t time, I made the mistake of eating to my heart's content three times a day.
Predicted category: sport
Is this prediction correct? (yes/no): no
You can only select these: ['business' 'entertainment' 'politics' 'sport' 'tech']
Enter the correct category: entertainment
{'accuracy': 0.9405162738496072, 'Precision': 0.941681851387967}
{'accuracy': 0.9730639730639731, 'Precision': 0.9736079992320874}
{'accuracy': 0.9775533108866442, 'Precision': 0.9775561917792899}
{'accuracy': 0.8170594837261503, 'Precision': 0.8184577254021699}
Best Model: Support Vector Machine
Enter some text: Nihal Koshie: You were at the cricket World Cup final. How was the experience as a fan? Also, your thoughts on
Australia as the big match-winners. Neeraj Chopra: This was the first time I watched a match fully. When I was on the flight, I
ndia had lost three wickets already. Virat (Kohli) bhai and KL Rahul were batting when I reached. There are some technical thin
gs that I don't understand. Batting in the daytime wasn't very easy. In the evening, I think, batting became easy. But our guys
tried. Sometimes, it's just not our day. But, frankly, everyone had a great tournament. Maybe, somewhere mentally, the Australi
an team held an edge at the start. When they bowled, I found they had a strong mindset. In the end, they had completely flipped
it over. They were confident about their game.
Predicted category: sport
Is this prediction correct? (yes/no): yes
```

How to use the application

Access the User Interface

Open the application's user interface through a web browser or a dedicated application interface.

Input Article

Choose the method for inputting articles:

- Direct Text Input: Type or paste the article text directly into the provided input field.
- External Source Integration: If available, enter the URL or source for fetching articles from external sources.

Submit and Wait for Prediction

- After inputting the text data, submit via the command prompt to initiate the topic classification process.
- The application will process the input and display the predicted category or topic for each submitted article.

Review Predictions

- Examine the results presented on the user interface, showcasing the predicted topics for the submitted articles.
- View additional details such as precision metrics and accuracy estimates to gauge the model's certainty.

Explore Visualizations

The application incorporates visualizations such as word clouds or bar graphs, explore these features to gain additional insights into the content distribution and key themes.

Continuous Improvement

- Understand that the application is designed for continuous improvement.
- User feedback and ongoing model updates ensure that the system evolves over time to enhance accuracy and relevance.
- The user has the option to be able to provide their feedback on whether they are satisfied with the output. In case, the model predicted an unsatisfied output for the user, the model is trained based on the feedback and pickled following.

Conclusions

Challenges Faced

Adaptability to Diverse News Sources

Implement transfer learning techniques to fine-tune the model on a broader dataset representative of various news sources. Regularly update the training data to expose the model to evolving language patterns.

Ensuring Continuous Improvement

Establish a systematic feedback loop for user input. Regularly update the model based on user feedback, emerging language trends, and shifts in news reporting styles.

Effective Preprocessing of Diverse Data

Develop custom preprocessing pipelines that are adaptive to diverse data structures. Leverage text processing libraries and techniques to handle intricacies, ensuring consistent data representation.

Interpreting Metric Scores

Provide documentation and education on interpreting the metric scores such as precision and accuracy.

Future Improvements

Diverse Categories

Expand the range of categories to capture a broader spectrum of topics and themes. This ensures that the model can accurately classify articles from a more extensive array of subjects, making it more versatile and applicable to various domains.

Enhanced Dataset

Curate a more comprehensive and high-quality dataset to train the model. This involves ensuring a balanced representation of each category, minimizing biases, and including a diverse set of articles. A well-constructed dataset is crucial for the model to learn robust features and generalize effectively.

Web Scraping for Data Collection

Implement web scraping techniques to gather the latest and most relevant articles from the web. This dynamic approach ensures that the model stays updated with current trends and emerging topics. Utilizing web scraping also allows for the inclusion of real-time data, improving the model's accuracy and relevance.

Predictive Analysis for Article Titles

Extend the model's capabilities to predict not only article categories but also generate potential article titles. This involves incorporating natural language generation techniques to create engaging and contextually relevant titles based on the content of the articles. This enhancement adds value by automating the title creation process and saving time for content creators.

GitHub repo link for code

https://github.com/ltang24/article-topic-classifer