DAILY WORK REPORT TR-02

INFOWIZ

26 JUNE 2024

Day 19: Text Classification with NLTK or spaCy

Summary: Today, we focused on text classification, a fundamental task in Natural Language Processing (NLP), using libraries like NLTK (Natural Language Toolkit) or spaCy. We explored the workflow for text classification, preprocessing techniques, feature extraction methods, and model evaluation metrics.

Key Learnings:

1. **Text Classification Basics:**
    - **Definition and Applications:** Defined text classification as the process of assigning predefined categories or labels to text documents based on their content.
    - **Applications:** Discussed applications such as sentiment analysis, spam detection, topic categorization, and language identification, showcasing the practical relevance of text classification in various domains.
2. **Workflow for Text Classification:**
    - **Data Preparation:** Preprocessed text data by removing stopwords, performing tokenization, and applying stemming or lemmatization to normalize text.
    - **Feature Extraction:** Utilized techniques like Bag-of-Words (BoW) model or TF-IDF (Term Frequency-Inverse Document Frequency) to convert text data into numerical feature vectors suitable for machine learning algorithms.
3. **Model Selection and Training:**
    - **Classification Algorithms:** Explored different classification algorithms such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression for text classification tasks.
    - **Implementation:** Implemented text classification models using NLTK or spaCy libraries, configuring classifiers and training them on labeled datasets.
4. **Evaluation Metrics:**
    - **Accuracy and Confusion Matrix:** Evaluated model performance using metrics like accuracy, precision, recall, and F1-score to measure classification accuracy and effectiveness.
    - **Confusion Matrix:** Visualized model predictions versus actual labels to analyze true positives, true negatives, false positives, and false negatives.
5. **Practical Application:**
    - Applied text classification techniques to a practical example, such as sentiment analysis of movie reviews or topic categorization of news articles.
    - Experimented with different preprocessing steps, feature extraction methods, and classification algorithms to optimize model performance and accuracy.

Today's session deepened our understanding of text classification techniques, equipping us with practical skills to preprocess text data, build classification models, and evaluate their performance using relevant metrics. The hands-on experience with NLTK or spaCy facilitated a comprehensive exploration of NLP applications in machine learning.