Module 11: Natural Language Processing

Project 1: NewsBot – Headline Classifier for InfoStream News

InfoStream News is a fast-growing digital news platform publishing hundreds of headlines daily. To improve user experience and automate categorization, You, as a data science intern, have been tasked to build NewsBot, an NLP-based Chabot that predicts the category of news headlines in real time.

Dataset;
https://www.kaggle.com/datasets/rmisra/news-category-dataset/data

Load the news headlines and categories from a JSON file.

1. Keep only these categories: TECHNO, ENTERTAINMENT, POLITICS, BUSINESS.
2. Load spaCy's English model without unnecessary components.
3. Write a function to preprocess headlines: lowercase, remove stopwords, punctuation, and lemmatize
4. Apply preprocessing to all headlines.
5. Convert text into numeric vectors using CountVectorizer with unigrams and bigrams
6. Limit the vocabulary size to 5000 features; explain why
7. Create feature matrix X and label vector y.
8. Split data into training and test sets with balanced categories.
9. Train Logistic Regression model with enough iterations.
10. Test the model and report accuracy.
11. Build a function to predict categories for new headlines.
12. Create a Chabot that takes user input and predicts the category until user types 'quit' or 'exit'.
13. Also, Deploy Streamlit App.