NO MORE MESSED UP FOLDERS:

Domain Background:

Natural language processing and machine learning:

Natural language processing or NLP is a domain of AI which deals with the interaction of machine learning algorithms and human language such as text or speech. Generally the spoken words have different meanings and it depends on the perspective of the receiver. The ultimate objective of NLP is to decrypt the messages hidden in human language, understand and make sense so that a pattern or insights could be developed.

Advances in the field of NLP, ML and AI has let researchers deep dive into this domain and have valuable insights form the data. The data in NLP is generally unstructured in real life. The task and different researches in NLP has led to the point that now we can have pretrained models such as BERT or use RNN such as LSTM to know what is meant in the sentence.

Personal motivation:

NLP is a growing field and that's why I have opted for it to be in my personal portfolio and hands-on in it.

The state of the art machine learning clustering and classification algorithms has further added to the pitching of this idea.

Problem Statement:

We all have a folder which is messed up from different categories of text documents. The idea is to have a automatic segmentor for the folder which can segregate our text documents into categorical folders, given that we know what categories we want our data to be segregated in.

The idea is to make a document segmentor, which per se will help in segmenting documents based on the content present in it and segregate in out local system storage automatically from the main folder based on the type of content present in it.

Dataset and Inputs:

The dataset I will be using will be BBC news article documents data which contains txt documents of news articles from different domains such as Sports, Entertainment, Technical, Business and Politics.

The input will be train data containing 10 documents of each type. I will be using k-shot learning method to train the model and will test on the rest of the dataset. The original dataset contains 400+ documents of each category.

Solution Statement:

Since the documents in BBC dataset are structured, I will not be using much pre-processing here.

The solution would involve getting the features extracted from the training documents, train our model and perform classification for final segregation.

Benchmark model:

I will be using Random forest classifier, multinomial NB algorithm, and Logistic regression to compare the results.

Evaluation metrics:

The accuracy would be derived using actual and predicted answers, false negatives, false positives, precision and recall.

Project design:

The steps involved for the program flow will be:

1) Importing the target folder where all documents are present.

2) Mentioning the output folder where you want the segregated folders to be made.

3) On training data, applying vectorizer (TF-IDF or count, whichever suits) to get all the features and do fit_transform on it.

4) Extracting bi-grams and tri-grams from the vectorized data.

5) Using Scikit-learn library to train on classifier such as Random Forest or Logistic regressor or Multinomial NB classifier.

6) Do predictions on the test data and send the files in the output folder declared earlier based on the categories.

**THIS IS SUPERVISED LEARNING BASED SOLUTION