Title: Preprocessing Fraud Documents(13)

Description:

This script includes two functions: one for converting .docx files to plain text and another for cleaning the extracted text using various utilities from the NLTK library.

Scope of improvement:

Currently, the stopwords resource is downloaded and imported each time to remove commonly used words that contribute little to the meaning of the text. This approach requires downloading the stopwords list every time, which can be inefficient. Instead, a predefined list of stopwords can be created. By utilizing a predefined list of stopwords for text cleaning, the process avoids repeated downloads of the stopwords resource, thereby improving the efficiency of the preprocessing workflow. Also, we can use Lemmatization in text perprocessing .