APPROACH FOR THE ‘BLACK COFFER’ TASK

1. First step was to importing the libraries i.e., ‘BeautifulSoup’, ‘nltk’, ‘re’, ‘requests’, ‘pyphen’, ‘numpy’ and ‘pandas’ in which ‘BeautifulSoup’, ‘requests’ was used to scrap the useful information from all the URLs provided, ‘pyphen’ was used to get the number of exact syllables in a word which was useful in calculating the keywords related to complex words. ‘nltk’ and ‘re’ was used for tokenizing and finding particular pattern in a word or sentence.
2. Next was the reading of all the text files provided in ‘StopWords’ folder and combining them all to make a set out of those words, next was the creating of list of positive and negative words using the text files provided in ‘MasterDictionary’ folder.
3. Next we create a function which is ‘final\_output’ which includes loop which runs around each URL in the dataset and carry out only the ‘p’ and ‘title’ html tags according to requirement and clean it, tokenize it using ‘re’ and ‘nltk’ libraries, inside that ‘for loop’ all functions creation was done which was all the columns values required for output i.e. ‘total\_words\_after\_cleaning’, ‘total\_no\_of\_sentences’, ‘positive\_score’, ‘negative\_score’, ’polarity\_score’, ‘subjectivity\_score’, ‘average\_sentence\_length’,’ number\_of\_complex\_words’, ‘percentage\_of\_complex\_words’, ’fog\_index’, ’ avg\_no\_of\_words\_per\_sent’, ‘word\_count’, ‘syllable\_per\_word’, ‘personal\_pronouns’, ‘average\_word\_length’ and finally carrying out all the values in a list and then updating into the the existing Output file.
4. Finally we exported the output dataframe to csv file and results are all calculated according to formulas and methods provided in ‘Text Analysis.docx’ file.

From : Trashant Rathore email-id: [trashnt22@gmail.com](mailto:trashnt22@gmail.com)

Contact : 9716089076