UE20CS332 – ALGORITHMS FOR INTELLIGENCE WEB AND INFORMATION RETRIEVAL

HANDS-ON SESSION : 01

TITLE : NLP TEXT PREPROCESSING

GOAL :

The goal of this hands-on session is to familiarise yourself with the kaggle platform at a base level, to download datasets from kaggle and use it on a python notebook for further processing.

CONCEPTS COVERED AND KEY TAKEAWAYS :

- Working with Kaggle datasets
- Major steps involved in NLP text pre-processing

USEFUL POINTS AND LINKS :

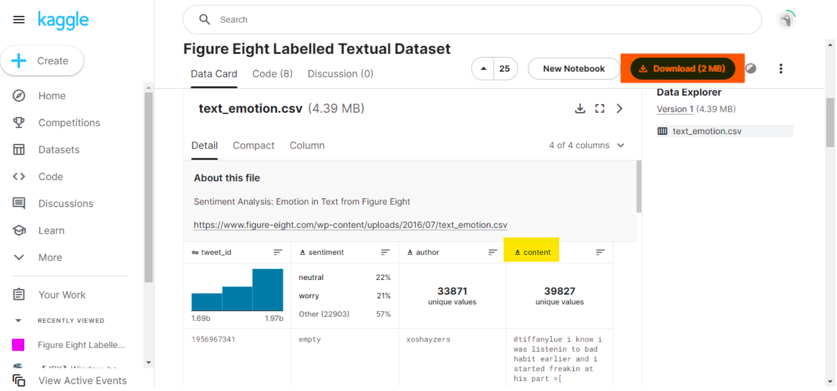
- Tokenization: Divide the text into individual words or phrases, called tokens.
- Case-folding: Convert all the characters in the text to lowercase to reduce the dimensionality of the data.
- Removing Stopwords: Eliminate commonly used words such as "the", "is", and "are" which do not provide much meaning to the text.
- Stemming or Lemmatization: Both techniques are used to reduce words to their base form, but they work in slightly different ways. Stemming uses heuristic rules to remove suffixes from words, while lemmatization uses a dictionary-based approach to find the base form of a word.
- Removing Punctuation and Special Characters: Remove any non-alphabetic characters such as punctuation marks or special characters.
- Converting numerical values to text: Remove any numerical values from the text as they may not be useful for certain NLP tasks.
- Removing HTML tags: Remove any HTML tags if the text is obtained from a

webpage.

- Removing Emoji and Emoticons: Remove any emoticons or emojis as they may not be useful for certain NLP tasks.
- Removing user mentions and hashtags: Remove any mention of specific users or hashtags as they may not be useful for certain NLP tasks.
- Reading material : Stemming and lemmatization

STEPS TO FOLLOW :

- Working with Kaggle :
  a. Follow the link Figure Eight Labelled Textual Dataset | Kaggle to download the dataset for this hands-on session.
  b. Download the csv file with the option highlighted red in the following image:



About this file

Sentiment Analysis: Emotion in Text from Figure Eight

https://www.figure-eight.com/wp-content/uploads/2016/07/text_emotion.csv

tweet_id	sentiment	author	content
	neutral 22% worry 21% Other (22903) 57%	33871 unique values	39827 unique values
1956967341	empty	xoshayzers	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[

The column highlighted in yellow is the focus column for this hands-on session. You may explore the page to learn more about kaggle datasets and how to use the same.

  c. To use the csv file as a dataframe, one of the following methods can be used :
    - Upload the csv file to the notebook working environment :

      df = pandas.read_csv(“<filename>.csv”) (OR)

    - Mention the path to access the csv file in the code :

      df = pandas.read_csv(“<path_to_file>/<filename>.csv”)

  d. The uploaded csv file is now in a dataframe format, ready to be used for the text processing tasks.
  e. The given dataset has 40,000 rows. Use the first 1000 rows for this hands-on

session to avoid resource limitation. One of these methods can be used :

  - df = df.iloc[:1000] (OR)
  - df.drop(df.tail(30000).index, inplace = True)

- On the dataframe created (named "df" in the above example), perform the following nlp text processing tasks on the text data in the column named "content" :
  a. Tokenization
  b. Case-folding
  c. Removal of punctuation marks, emoticons, HTML tags and links
  d. Convert numerical values to text (Ex: 10 -> "ten")
  e. Stopword removal
  f. Stemming
  g. Lemmatization

NOTE :

- We require you to appropriately document your code using the Markdown feature for each different text processing task.The first cell must be a markdown cell which contains the following details :
  a. UE20CS332 : Algorithms For Intelligence Web And Information Retrieval
  b. SRN : PES1UG20CSXXX
  c. Name : Tom Cruise
  d. Section : X

SUBMISSION LINK FOR HANDS ON - 01:

https://forms.gle/Zv8VfVKXRbKPPpXB6

Format for evaluation: PES1UG20XXX_HandsOn01.ipynb

- **DEADLINE : END OF DAY**
- **ANY SUBMISSION POST THE DEADLINE WON'T BE CONSIDERED**
- **ENSURE THE FIRST CELL OF THE NOTEBOOK IS AS MENTIONED ABOVE**

TA CONTACT DETAILS:

Abhay D A:	abhayda2001@gmail.com
Neha Angadi:	nehaangadi19@gmail.com
Priya S S:	sspriya@pesu.pes.edu