**Possible Issues/Userstories:**

1. Explore different processing techniques of given dataset using the packages learned in nltk chapters
2. Extracting features for different tasks to be used in the algorithm
3. Generation and modification algorithm based on Dissimilarity Measure
4. Run the baseline code provided in PAN20 website and run evaluation measures provided in pan20
5. Run algorithm generated with dissimilarity measure with evaluation code and run evaluation measures provided in pan20
6. Run and compare the baseline code and algorithm generated with dissimilarity measure with evaluation code along with tuning of the thresholding parameters to improve evaluation measures

**Possible Sprints:**

1. Explored different processing techniques of given dataset using the packages learned in nltk chapters
   1. Counting vocabulary, lexical diversity, Operations on different data types and basics statistics
   2. Word n-gram and Character n-gram Manipulation, Tokenization and file management
   3. Application of different possible Regular Expression and Normalisation text
   4. All possible operation using text data type, Comparative wordlists, and Exploring different functions in wordnet package
   5. Explore different types of functions and exploring different libraries
   6. Exploring different types of taggers and Combining taggers to be operated on texts
2. Extracting features for different tasks to be used in the algorithm
   1. Word n gram
   2. Character n-gram
   3. Punctuation count
   4. Part of Speech(POS) Frequency
   5. Vocabulary Strength
3. Generation and modification algorithm based on Dissimilarity Measure
   1. Creating the dissimilarity measure function
   2. Creating dataset with mean average of dissimilarity measure
   3. Creating the predicted dataset with
4. Run the baseline code provided in PAN20 website and run evaluation measures provided in pan20
5. Run algorithm generated with dissimilarity measure with evaluation code and run evaluation measures provided in pan20
6. Run and compare the baseline code and algorithm generated with dissimilarity measure with evaluation code along with tuning of the thresholding parameters to improve evaluation measures

**Possible Tasks in Sprints :**

**Issue 1: Explore different processing techniques of given dataset using the packages learned in nltk chapters**

1. Counting vocabulary, lexical diversity, Operations on different datatypes and basics statistics
   1. Vocabulary count and lexical diversity count of the text  (Nilesh)
   2. Frequency Distribution to count the unigrams (Nilesh)
   3. Counting the number of times digits used in the text (Nilesh)
   4. Converting tokens into lowercase for proper counting (Nilesh)
   5. Counting number of words with different length (Nilesh)
   6. Average word length of text (Nilesh)
2. Word n-gram and Character n-gram Manipulation, Tokenisation and file management
   1. Reading of given datasets in JSONL file and appropriate conversion into other datatypes(Smruti)
   2. Tokenize the text (Nilesh)
   3. Using biagram and trigram to count number of word bigram and trigram. (Nilesh)
   4. Making character n-gram from raw text and counting them using Frequency Distribution (Nilesh)
3. Application of different possible Regular Expression and Normalisation text
   1. Counting the punctuation marks (Nilesh)
4. All possible operation using text data type, Comparative wordlists, and Exploring different functions in wordnet package
   1. Performing operations like Lexical diversity on the text data type, (Rohan)
   2. exploring comparative wordlists in case of future requirement according to our need. (Rohan)
   3. Finding the most common words and the unusual words  from the text to have a statistical figure of it. (Rohan)
5. Explore different types of functions and exploring different libraries
   1. Making use of structured programming so that the programs are readable for every team members. (Rohan)
   2. Working on Structured programming includes operators, sequences, operating on sequence types, working on functions(input and output), Parameter Passing, Checking Parameter types.(Rohan)
   3. Exploring libraries in python like matplotlib. (Rohan)
6. Exploring different types of taggers and Combining taggers to be operated on texts
   1. Basic data processing on the text like finding concordance and similarity of words,collocation list
   2. frequency distribution and conditional frequency distribution of words and tags and plotting of graphs
   3. Usage of default dictionary for assigning unknown tags
   4. Creating sentences from text and applying taggers to
   5. Evaluation of Regex tagger, Lookup tagger, n gram taggers and Combining taggers