QUESTION:

Explanation for features Design for feat gen file.

ANSWER:

Feature Design:

Data Preprocessing:

- **Lexicon dictionary:**
  - Whenever a word would be found in the dictionary, the lexicon name would be added to its feature. This was done for all the lexicons except a few which were not included, since the categories had no connection to it.
  - All the words of the given lexicons were added in a one dictionary. The value of each word is a list of lexicon name.
  - All the words were lower cased.
  - Lowering the case of the word is another feature that helps us in the accuracy.

- **Removing punctuations:**
  - The punctuations were also removed from the data.
  - This data with the removal of punctuation was added in the dictionary.
  - The punctuation removal proved to increase the accuracies to some extent.

Features:

- **Capitalization**:

  - Since isupper() was already applied to the word, Capitalization was only checked for the first letter of every word.
  - How is it helpful?

    This helped because we know named entities start with a capital letter. So, they would get similar features.

- **POS Tags:**

  - All the words are assigned their POS tags, accordingly. The POS tags gave the form of the word, as per the sentence.
  - I used "nltk" library- POS tags to compute the POS tags of the words
  - How is it helpful?

    Like the above reason, POS tags of most of the named entities would be similar, probably NNP since they would be proper nouns. This would give same features to similar nouns.

- **Lexicon features:**

  - The lexicon dictionary that I made above, is used for this feature. This is also known as Gazetteer features.
  - In other words, a gazetteer feature would be a binary feature indicating whether the word was in my gazetteer. (In this case, lexicon of a kind)
  - For example, if a word is a "person", and it is found in the lexicon "first-name" lexicon, then "is_first_name" will be added as a feature vector of the given word.
  - How is it helpful?
    Having words belonging to a category, would be its feature. This would happen for many words, and since there will be many of these words, that would categorize them in to that category.

    For example, for the above case, if different sports teams are found in the respective gazetteers, they both would have "is_first_name" as a feature. This along with features helped getting better results.

- **Vowel Count:**

  - Vowel count for every word is also added as a feature.
  - This is more of the average vowel count, since I divided the number of vowels in a word with total characters in a line.
  - This, added as a feature gave better accuracy.
  - How is it helpful?
    Normalization gives us approximate features per word, and give good results. This is so because it is another way of finding he number of syllables in a word.

    This somehow adds on to be a good feature, and helps further in more accurate results.

- **Shape of the Word:**

  - Also called orthographic feature.
  - Shape of the word is added as a feature in the feature vector of the word.
    A word shape encodes the first and last two characters of the word, and a set of the encodings of the characters in between.
    Here, word shape of a given word is encoded as "X" for an uppercase letter, "x" for a lowercase letter, "d" for a digit and punctuation for the punctuation.

  - How is it helpful?

    This shape is added as a feature and will let us know which words have similar shapes, and will match the features of named entities. For example, all names of a person would have word shape as "Xxxx", and would help us find that these are indeed proper nuns, or to be specific even person names.

Along with other features, it is beneficial for us to get better results.