CO3251-NATURAL LANGUAGE PROCESSING

ASSIGNMENT-02

NAME - U.D.P.D UDUWELA

INDEX NO - 18/ENG/112

REGISTRATION NO - EN91361

DATE OF SUBMISSION-02/07/2021

How to use the Alexa

- We can ask any item from the Alexa. Alexa gives most suitable findings.

  As example-

  - ✓ If need a pen. Ask ‘what is the shelf number of pens’, ‘where are pens. likewise including the name of the particular good, you can ask questions. Then it will show the relevant information for that good on the console.
  - ✓ If Alexa cannot find such good display error message or ask again and again to insert new one or exit from the system.
- You can use greeting such as ‘hello’, ‘hi ‘. Then chat bot will reply in similar manner.
- To print the data, use “print” keyword. Then chat bot automatically invoke the printer process from OS. In the video I haven’t connected a printer. Therefore, can’t print it as a hard copy. But program automatically detects an available printer from your OS and try to print the available data write on sample.txt file.
- At the end, all chat history and generated google voice clips and text files are removed from the system. No need to care about the pervious garbage’s. Here the voice clips can be delete if the program terminate by using correct keyword. otherwise, the randomly named mp3 files will be keep in the main directory until they are removed manually.
- Please use the application in clam and quiet place. because google voice might not able to identify some words correctly. But Alexa tries show most similar findings.
- To exit from the chatbot use command like ‘Thanks’, ‘Thank you’, ‘exit’ and ‘bye’.

Technology use for Alexa

✓ Text Pre- Processing with NLTK

The main issue with text data is that it is all in text format (strings). However, the Machine learning algorithms need some sort of numerical feature vector in order to perform the task. So, before we start with any NLP project, we need to pre-process it to make it ideal for working. Basic text pre-processing includes:

1. **Converting the entire text into uppercase or lowercase**, so that the algorithm does not treat the same words in different cases as different

2. **Tokenization**: Tokenization is just the term used to describe the process of converting the normal text strings into a list of tokens i.e., words that we actually want. Sentence tokenizer can be used to find the list of sentences and Word tokenizer can be used to find the list of words in strings.

3. **Removing Noise** i.e. everything that isn't in a standard number or letter.

4. **Removing Stop words**. Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words

5. **Stemming**: Stemming is the process of reducing inflected (or sometimes derived) words to their stem, base or root form — generally a written word form. Example if we were to stem the following words: "Stems", "Stemming", "Stemmed", "and Stemtization", the result would be a single word "stem".

6. **Lemmatization**: A slight variant of stemming is lemmatization. The major difference between these is, that, stemming can often create non-existent words, whereas lemmas are actual words. So, your root stem, meaning the word you end up with, is not something you can just look up in a dictionary, but you can look up a lemma. Examples of Lemmatization are that "run" is a base form for words like "running" or "ran" or that the word "better" and "good" are in the same lemma so they are considered the same.

7.
8. **Bag of Words**: After the initial preprocessing phase, we need to transform text into a meaningful vector (or array) of numbers. The bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:
   - ➢ A vocabulary of known words.
   - ➢ A measure of the presence of known words.

   Why is it is called a "bag" of words? That is because any information about the order or structure of words in the document is discarded and the model is only concerned with whether the known words occur in the document, not where they occur in the document. The intuition behind the Bag of Words is that documents are similar if they have similar content. Also, we can learn something about the meaning of the document from its content alone.

- There are some other packages and libraries for voice recognitions and play sound in the program. Before using the program, the python packages should be download to your project.

Example - speech recognition, play sound

On the other hand, there are some libraries to use for printing. All the libraries are clearly comment in the python code