

NLP Techniques

Knowledge base:

1. Text preprocessing:
 - a. To get the corpus text in a workable format, some preprocessing was done. First the text had stopwords removed and was converted to lower case. Then the text was sentence tokenized to break up the book into sentences and fed into the keyword extractor.
2. Yake Keyword extraction:
 - a. Because TF-IDF relies of the documents within the corpus to be unique and NLPete's corpus was a single textbook, Yake keyword extractor was used to determine important words within the corpus. Yake is a statistical keyword extractor that uses features within a single text to find relevant words.
3. POS tagging:
 - a. POS tagging, or part of speech tagging, is a natural language processing technique where each word in a sentence is converted into a tuple in the form of a word and a tag. The knowledge base uses spacy's POS tagger to determine if a sentence is valid or not.

Chat flow:

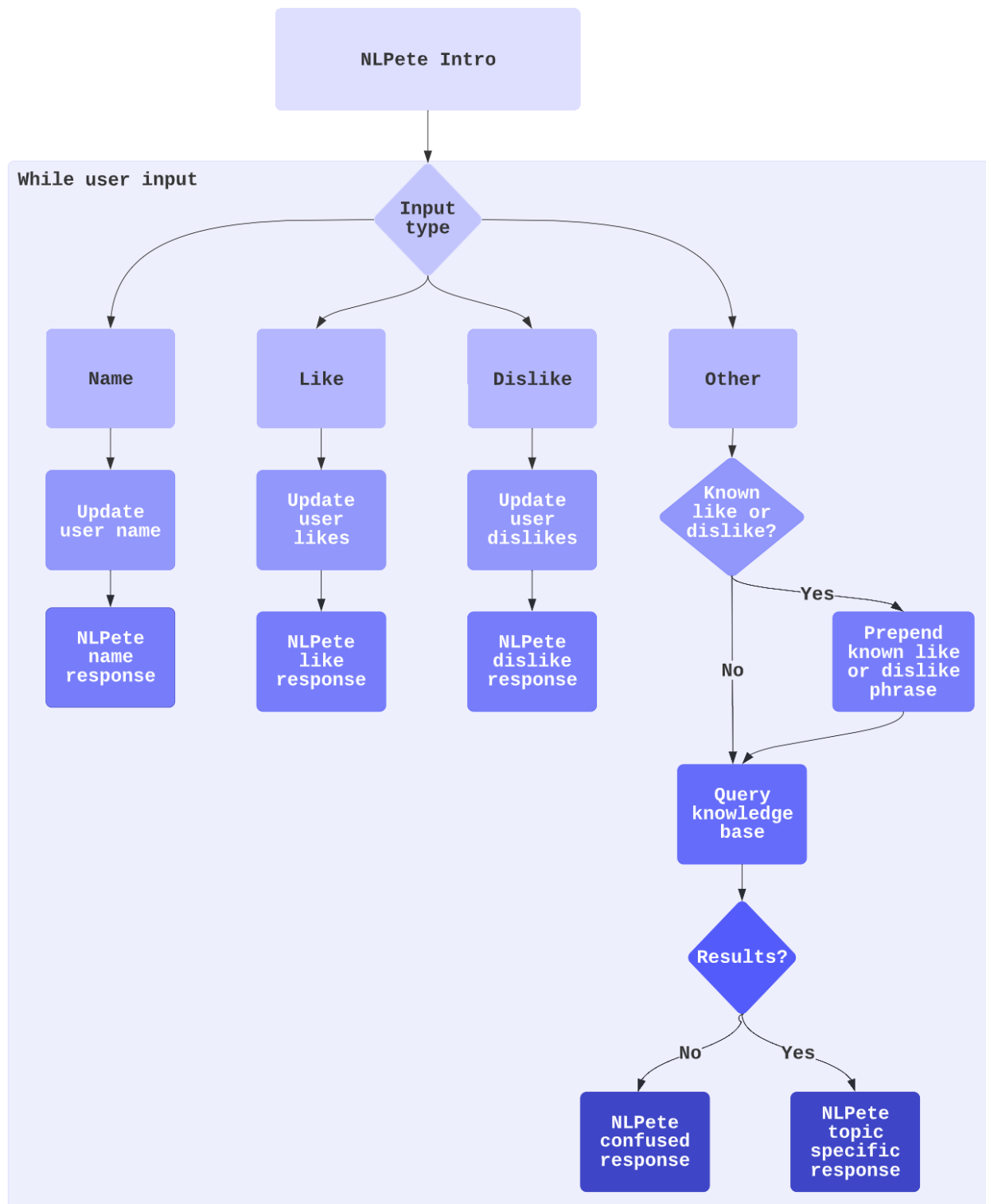
1. Dependency parsing:
 - a. Dependency parsing is a natural language processing technique that establishes dependencies between parts of a sentence to determine syntactical structure. NLPete used this technique to build the user's profile and learn their likes or dislikes. For instance, NLPete would take: "i like avocados" and find the DOBJ of "like" to determine the user likes something.
2. POS tagging:
 - a. Also used in the knowledge base, Spacy's POS tagger was used to help NLPete find a user's name. For instance, if the user typed "my name is katie", NLPete would see:

my|PRON name|NOUN is|AUX Katie|PROPN

And look for PROPN indicating a name.

II. Dialog

Logic Tree



Interactions

```
Me
my name is katie

NLPete
Nice to meet you, katie! How can I help you today?

katie
what is natural language processing

NLPete
This reveals a few harsh truths about NLP work: NLP results are not perfect because language is messy.

katie
language is messy

NLPete
The choice of a corpus is important because each body of text will be unique in the focus and the types of language.

katie
i like language

NLPete
Me too, language is sweet!
```

```
katie
i like pickles

NLPete
Me too, pickles is sweet!

katie
i hate pickles

NLPete
I thought you liked pickles. Oh well, your loss.
```

```
katie
bye

NLPete
Leaving so soon? Okay :(
```

```
katie
logistic regression

NLPete
Logistic Regression Despite its name, logistic regression, performs classification, not regression.

katie
can computers understand human speech

NLPete
Semantics Before looking at these approaches, the question arises: How does the human brain extract meaning from sentences?
```

```
NLPete
That went straight over my head. Can we move on?

katie
what is dependency parsing

NLPete
Syntax and Parsing Figure 10.6: Clause Structure 10.2 Dependency Parse A dependency parse shows relationships between the words of a sentence in an acyclic graph.
```

III. Evaluation

Strengths

NLPete has a lot of strengths! His responses are very quick, and he has a very human sounding dialogue when having casual discussion. He can handle words similar to other words such as “natural language processing” and “nlp”. He also deals with contractions and doesn’t require punctuation or end tokens. One of the coolest things about NLPete is his user interface.

Weaknesses

NLPete isn’t completely without flaw. His responses in regards to natural language processing topics are sometimes a little too specific when the user wants a general description. Sometimes the characters in the text cause problems for the textual interface as well.

IV. Appendices

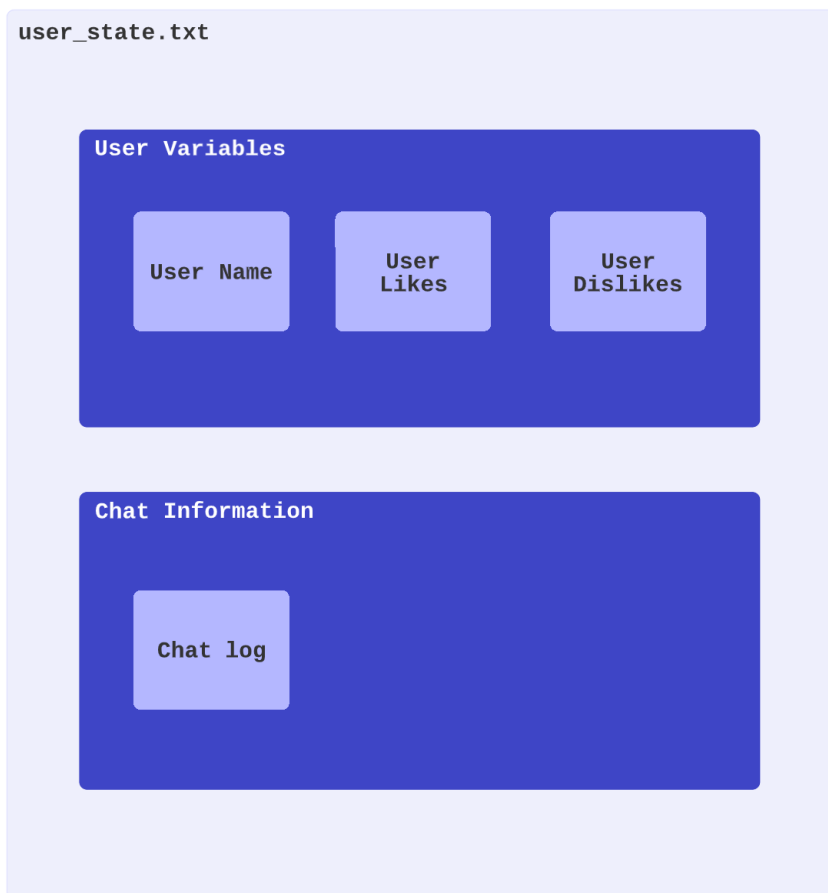
Knowledge Base

Once text preprocessing was completed, building the knowledge base itself was relatively easy. A Python dictionary structure was used where extracted keywords were the key and the values were lists of sentences with containing those key words. This object was then pickled. Because the corpus was so large, NLPete’s knowledge base has over 1000 key word entries. For example, the keyword ‘lemmas’ has a list of 7 sentences relevant to it.

'lemmas': ['The entries in a lexicon are lemmas , the base form of the word.', 'The next part of the code below uses a dictionary comprehension to make a dictionary of lemma:count entries for unique lemmas.', 'Dictionary Counts of Lemmas.', 'Recall that lemmas are the dictionary form of a word, so lemmas are just entries in the WordNet lexicon.', 'Annotating lemmas and named-entities may be useful.', 'The online notebook shows how to extract information token by token, including lemmas, POS, and more.', 'Additionally, word lemmas and parts of speech are gathered, along with named entity information.']

User Models

The user model represents a single user interaction with NLPete. The output file `user_state.txt` stores the user's name, their likes, their dislikes, and the entire chat log for analysis later.



```
user_state.txt - Notepad
File Edit Format View Help
User Name: eric
User likes: you, kinglsey, attitude
User dislikes: language
Chat log:
NLPete
Hi! I'm NLPete, your comprehensive guide on natural language processing! What's your name?
```