TOKENIZER :

Input : An English paragraph consisting of Words, Emoticons and other punctuation( comma, full stop ) ?Usernames

Output : An array of arrays which in turn consist of 3-tuple(word , Sentiment , Priority)

Notes :

1. Sentiment can be -1 (Negative) , 0 (Undefined, No sentiment , Word not recognized ) or 1. It is initialized to 0 in this step .Only after lexicon look-up, correct sentiment is attached.)
2. Priority is either 0 (normal ) or 1 (high). High for All Capitalized word and Emoticons.

Steps :

Sentence tokenize

Word tokenize (White space )

Then separate full stand-alone punctuations, shorten words using Regular Expressions