Tweet Tokenizer

B VARSHIT
201402029

Challenges Faced

1. Made a good regex for the urls so that htt... wouldn't get identified as a URL.
2. Included most of the symbols for tokenizing the smileys.
3. Smileys had to be decoded so that they are visible in files.
4. Took care of the non-ascii characters like …
5. Could not handle characters like é. They get seperated from words like Pokémon etc i.e it becomes Pok, émon