

# Tokenization

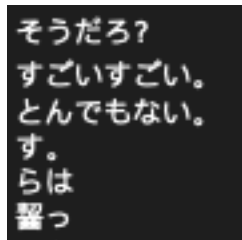
## QUESTIONS from 01a file:

To execute the code:

```
echo 'sentence to tokenise.' | python3 tokenization.py dictionary-file  
or  
cat 'file to be tokenized' | python3 tokenization.py dictionary-file
```

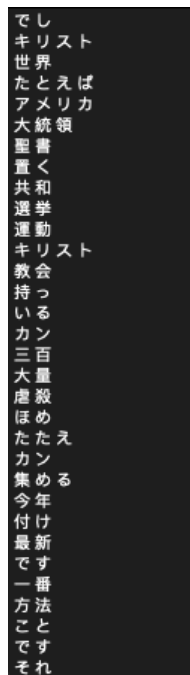
### Dictionary File

Dictionary file contains list of words( tokens ) of the language and each word should be in new line



そうだろ?  
すごいすごい。  
とんでもない。  
す。  
らは  
髷っ

After executing it output should look like



でし  
キリスト  
世界  
たえ  
ば  
アメリカ  
大統領  
聖書  
置く  
共和  
選挙  
運動  
キリスト  
教会  
持つ  
いる  
カン  
三  
百  
大  
量  
殺  
め  
た  
え  
た  
カン  
集  
め  
る  
今  
年  
付  
最  
新  
で  
一  
番  
方  
法  
と  
す  
れ  
そ

- brief description of its performance, with examples to support your findings  
 Comparing the performance, WER gives very low accuracy.  
 Some words were separated out into small new tokens. For example:  
 “ るそうです。” word is considered as single token as it was present in the dictionary as a whole but for  
 This work “ という感じでした。” which is tokenized into “ という感じ” , “ でし” and “ た。” as this word was not there in dictionary.

### QUESTIONS from 01b file:

- Why should we split punctuation from the token it goes with ?

Punctuations are distinct tokens in and of itself. These are typically used to give clarity to the writing by pausing, stopping, showing emotion, etc. For example, ‘Data’ and ‘Data!’ ‘Can mean different. So, when punctuation are removed all words are treated equally.

- Should abbreviations with space in them be written as a single token or two tokens?
  - How about numerals like 134 000 ?

Yes, abbreviations with space between them should be treated as same token  
 example U S A and USA are same. Same with the numerals.

- If you have a case suffix following punctuation, how should it be tokenised ?

Case suffix should be treated as same token after removing punctuation.

- Should contractions and clitics be a single token or two (or more) tokens ?

These should be treated as single tokens.