Tokenisation

March 20, 2024

[ ]:
```
#python stemmimg example
```

[10]:
```
import nltk
from nltk.stem import PorterStemmer
nltk.download("punkt")
```

```
[nltk_data] Downloading package punkt to /home/nmit/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
```

[10]:
```
True
```

[11]:
```
# Initialize Python porter stemmer
ps = PorterStemmer()
```

[13]:
```
# Example inflections to reduce
example_words = ["program","programming","programer","programs","programmed"]
```

[14]:
```
# Perform stemming
print("{0:20}{1:20}".format("--Word--","--Stem--"))
for word in example_words:
   print ("{0:20}{1:20}".format(word, ps.stem(word)))
```

```
--Word--            --Stem--
program             program
programming         program
programer           program
programs            program
programmed          program
```

[15]:
```
#python tokenise example
```

[16]:
```
import string
from nltk.tokenize import word_tokenize
```

[17]:
```
example_sentence = "Python programmers often tend like programming in python␣
 ↪because it's like english. We call people who program in python pythonistas."
```

[18]:
```
# Remove punctuation
```

```
example_sentence_no_punct = example_sentence.translate(str.maketrans("", "",␣
 ↪string.punctuation))
```

[19]:
```
# Create tokens
word_tokens = word_tokenize(example_sentence_no_punct)
```

[20]:
```
# Perform stemming
print("{0:20}{1:20}".format("--Word--","--Stem--"))
for word in word_tokens:
    print ("{0:20}{1:20}".format(word, ps.stem(word)))
```

```
--Word--            --Stem--
Python              python
programmers         programm
often               often
tend                tend
like                like
programming         program
in                  in
python              python
because             becaus
its                 it
like                like
english             english
We                  we
call                call
people              peopl
who                 who
program             program
in                  in
python              python
pythonistas         pythonista
```

[21]:
```
#Python Lemmatization example
```

[22]:
```
from nltk.stem import WordNetLemmatizer
nltk.download("wordnet")
nltk.download("omw-1.4")
```

```
[nltk_data] Downloading package wordnet to /home/nmit/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /home/nmit/nltk_data…
```

[22]:
```
True
```

[23]:
```
# Initialize wordnet lemmatizer
wnl = WordNetLemmatizer()
```

[24]:
```
# Example inflections to reduce
example_words = ["program","programming","programer","programs","programmed"]
```

[25]:
```
# Perform lemmatization
print("{0:20}{1:20}".format("--Word--","--Lemma--"))
for word in example_words:
   print ("{0:20}{1:20}".format(word, wnl.lemmatize(word, pos="v")))
```

```
--Word--            --Lemma--
program             program
programming         program
programer           programer
programs            program
programmed          program
```

[ ]: