Stemming using NLTK

SECTION 2

Eng. Fatma

What is Stemming ?

- After tokenization has sentences divided into words, stemming is a procedure to unify words and extract the root, base form of each word.

- Stemming usually removes prefixes or suffixes such as -er, -ion, -ization from words to extract the base or root form of a word.

What is Stemming ?

- **Example**:

- computers, computation, and computerization.

- compute is the stem of these words.

How to perform Stemming ?

- NLTK provides practical solution to implement stemming without sophisticated programming.
- Let us try two commonly used methods:
- (1) Porter Stemmer and
- (2) Snowball Stemmer in NLP.

Porter stemmer

Porter Stemmer is the earliest stemming technique used in 1980s.

Its key procedure is to remove words common endings and parse into generic forms.

This method is simple and used in many NLP applications effectively.

```
# Import PorterStemmer as p_stem
from nltk.stem.porter import PorterStemmer as p_stem
p_stem().stem("computer")
```

```
'comput'
```

Porter stemmer

```
# Import PorterStemmer as p_stem
from nltk.stem.porter import PorterStemmer as p_stem
p_stem().stem("dogs")
```

```
'dog'
```

Porter stemmer

Note: Stemmer may output an invalid word when dealing with special words e.g.

tradit is acquired if suffix -ional is removed. tradit is not a word in English,

it is a root form.

```
# Import PorterStemmer as p_stem
from nltk.stem.porter import PorterStemmer as p_stem
p_stem().stem("traditional")
```

Porter stemmer

Note: Stemmer may output an invalid word when dealing with special words e.g.

tradit is acquired if suffix -ional is removed. tradit is not a word in English,

it is a root form.

```
# Import PorterStemmer as p_stem
from nltk.stem.porter import PorterStemmer as p_stem
p_stem().stem("traditional")
```

Porter stemmer

Note: Porter Stemming will remove suffixes -s or -es to extract root form, that may result in single form such as apes, bags, dogs, etc. But in some cases, it will generate non-English words such as gener, jungl and queri.

```
# Define some plural words
w_plu = ['apes','bags','computers','dogs','egos','frescoes','generous','hats','igloos','jungles', 'kites','learners','mice',
         'natives','openings','photos','queries','rats','scenes', 'trees','utensils','veins','wells','xylophones','yoyos',
         'zens']
from nltk.stem.porter import PorterStemmer as p_stem
w_sgl = [p_stem().stem(wplu) for wplu in w_plu]
print(' '.join(w_sgl))
```

```
ape bag comput dog ego fresco gener hat igloo jungl kite learner mice nativ open photo queri rat scene tree utensil vein well
xylophon yoyo zen
```

Snowball stemmer

Snowball Stemmer provides improvement in stemming results as compared with

Porter Stemmer and provides multi-language stemming solution.

One can check languages using languages() method. Import from NLTK package to invoke

Snowball Stemmer:

```
# Import Snowball Stemmer as s_stem
from nltk.stem.snowball import SnowballStemmer as s_stem
# Display the s_stem language set
print(s_stem.languages)
```

Snowball stemmer

Snowball Stemmer provides improvement in stemming results as compared with

Porter Stemmer and provides multi-language stemming solution.

One can check languages using languages() method. Import from NLTK package to invoke

Snowball Stemmer:

```
# Import Snowball Stemmer as s_stem
from nltk.stem.snowball import SnowballStemmer as s_stem
# Display the s_stem language set
print(s_stem.languages)
```

Snowball stemmer

```
# Import Snowball Stemmer as s_stem
from nltk.stem.snowball import SnowballStemmer as s_stem
s_stem_ENG = s_stem(language="english")
w_plu = ['apes','bags','computers','dogs','egos','frescoes','generous','hats','igloos','jungles', 'kites','learners','mice',
         'natives','openings','photos','queries','rats','scenes', 'trees','utensils','veins','wells','xylophones','yoyos',
         'zens']

# Apply Snowball Stemmer onto the plural words
sgls = [s_stem_ENG.stem(wplu) for wplu in w_plu]
print(' '.join(sgls))
```

```
ape bag comput dog ego fresco generous hat igloo jungl kite learner mice nativ open photo queri rat scene tree utensil vein w
ell xylophon yoyo zen
```

Snowball stemmer

Try to compare with previous stemmer. What are the differences?

1. Snowball Stemmer achieved similar results as Porter Stemmer in most cases

except in generously where Snowball Stemmer came up with a meaningful

root form generous instead of gener in Porter Stemmer

2. Try some plural words to compare performance between Porter Stemmer vs

Snowball Stemmer