Normalization

Text Normalization – Wiki Definition

Text normalization

Text normalization is the process of transforming **text** into a single canonical form that it might not have had before. **Normalizing text** before storing or processing it allows for separation of concerns, since input is guaranteed to be consistent before operations are performed on it.

Text normalization - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Text_normalization** ▾

Text Normalization

- Before almost any natural language processing of a text, the text has to be normalized.
- At least three tasks are commonly applied as part of any normalization process:
  1. Segmenting/tokenizing words from running text.
  2. Normalizing word formats.
  3. Segmenting sentences in running text.

- Finland's capital → Finland Finlands Finland's ?
- what're, I'm, isn't → What are, I am, is not
- Hewlett-Packard → Hewlett Packard ?
- state-of-the-art → state of the art ?
- Lowercase → lower-case lowercase lower case ?
- San Francisco → one token or two?
- m.p.h., PhD. → ??

Punctuation

구두점

```
import re
from string import punctuation
from nltk.tokenize import word_tokenize

sentence = 'i'd like to learn more somthing. i'd like to learn more somthing.'
removePattern = re.compile('[{0}]'.format(re.escape(punctuation)))

tokens = word_tokenize(sentence)
print(tokens)

cleanTokens = list()

for term in tokens:
    newTerm = removePattern.sub('', term)

    if newTerm:
        cleanTokens.append(newTerm)

print(cleanTokens)
```

Stopwords

Stopwords

```
from nltk.corpus import stopwords

stopwords.fileids()
stopwords.words('english')
stop = stopwords.open('english').read()
```

```
cleanWords = list()

for term in 'i like you'.split():
    if term in stop:
        print(term, '[Skipped]')
    else:
        print(term, '[Passed]')
        cleanWords.append(term)

cleanWords
```

Remove stopwords

In [10]:
```
import nltk

stopword = nltk.corpus.stopwords.words('english')# All English Stopwords
```

In [11]:
```
# Function to remove Stopwords
def remove_stopwords(tokenized_list):
    text = [word for word in tokenized_list if word not in stopword]# To remove all stopwords
    return text

data['body_text_nostop'] = data['body_text_tokenized'].apply(lambda x: remove_stopwords(x))

data.head()
```

Out[11]:

	label	body_text	body_text_clean	body_text_tokenized	body_text_nostop
0	ham	I've been searching for the right words to tha...	Ive been searching for the right words to than...	[ive, been, searching, for, the, right, words,...	[ive, searching, right, words, thank, breather...
1	spam	Free entry in 2 a wkly comp to win FA Cup fina...	Free entry in 2 a wkly comp to win FA Cup fina...	[free, entry, in, 2, a, wkly, comp, to, win, f...	[free, entry, 2, wkly, comp, win, fa, cup, fin...
2	ham	Nah I don't think he goes to usf, he lives aro...	Nah I dont think he goes to usf he lives aroun...	[nah, i, dont, think, he, goes, to, usf, he, l...	[nah, dont, think, goes, usf, lives, around, t...
3	ham	Even my brother is not like to speak with me. ...	Even my brother is not like to speak with me T...	[even, my, brother, is, not, like, to, speak, ...	[even, brother, like, speak, treat, like, aids...
4	ham	I HAVE A DATE ON SUNDAY WITH WILL!!	I HAVE A DATE ON SUNDAY WITH WILL	[i, have, a, date, on, sunday, with, will]	[date, sunday]

불용어

```
korStop = '은, 는, 이, 가, 께, 을, 를, 고, 께서, 게, 에게'

cleanWords = list()

for term in '어머님 은 자장면 이 싫다 고 하셨어'.split():
    if term in korStop:
        print(term, '[Skipped]')
    else:
        print(term, '[Passed]')
        cleanWords.append(term)

cleanWords
```

그 외

대소문자	Case-folding by reducing all letters to lower case
짧은 어휘	Removing words with very a short length
저빈도 어휘	Removing rare words
정규표현식	

한글 불용어와 비속어 처리

형태	품사	비율
이	VCP	0.01828
있	VA	0.011699
하	VV	0.009774
것	NNB	0.009733
들	XSN	0.006898
그	MM	0.005327
되	VV	0.003613
수	NNB	0.003474
이	NP	0.003361
보	VX	0.00331
않	VX	0.002976
없	VA	0.00292
나	NP	0.00269
사람	NNG	0.002074
주	VV	0.001885
아니	VCN	0.001871
등	NNB	0.001822
같	VA	0.001725
우리	NP	0.001715
때	NNG	0.001686
년	NNB	0.001648
가	VV	0.001619
한	MM	0.001584
지	VX	0.001538