딥러닝을 활용한 자연어 분석

문장분석

김용범

무영인터내쇼날

자연어 분석과정

1. **형태론(Morphology) :** 단어와 형태소를 연구
2. **통사론(syntax) :** 문법적 구조분석(Parsing)
3. **의미론(Senmantics) :** 단어 의미차이 **ex) 뉘앙스, 톤, 의도(긍/부정)**

Phrase
토큰 활용

불용어 처리

단어의 활용

Stop Words

1. **연관성이 낮은 단어**들을 제외하고 분석
2. **내용**과 **목적**에 따라서, **불용어 처리여부** 및 해당 **목적에 맞는 불용어 말뭉치 DB**등을 판단해야 한다

Stop Words

In [10]:
```
texts = 'I like such a Wonderful Snow Ice Cream'
texts = texts.lower()
texts
```

Out[10]:
```
'i like such a wonderful snow ice cream'
```

In [11]:
```
from nltk import word_tokenize
tokens = word_tokenize(texts)
tokens
```

Out[11]:
```
['i', 'like', 'such', 'a', 'wonderful', 'snow', 'ice', 'cream']
```

Stop Words

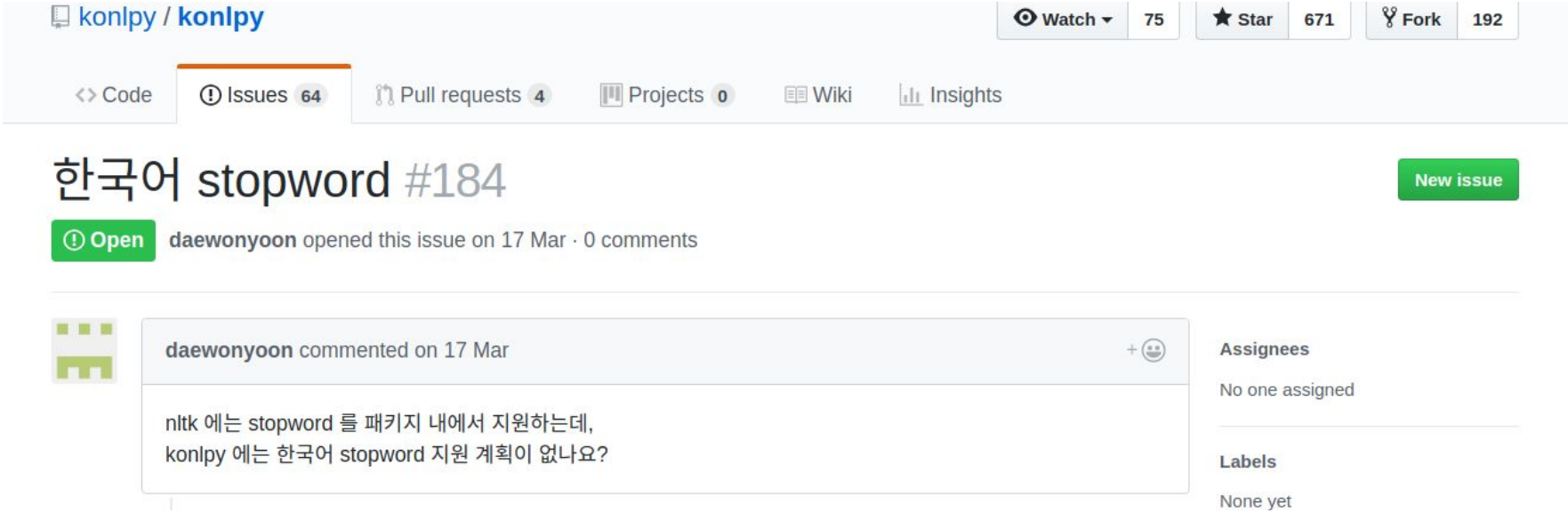
In [12]:
```
from nltk.corpus import stopwords
stopwords.words('english')[::18]
```

Out[12]:
```
['i', 'her', 'those', 'an', 'into', 'further', 'such', 'now', 'mightn']
```

In [13]:
```
tokens = [word   for word in tokens
                 if word not in stopwords.words('english')]
print(tokens)
```

```
['like', 'wonderful', 'snow', 'ice', 'cream']
```

한글의 경우


konlpy / konlpy
Watch 75
Star 671
Fork 192
Code
Issues 64
Pull requests 4
Projects 0
Wiki
Insights
한국어 stopword #184
New issue
Open
daewonyoon opened this issue on 17 Mar · 0 comments
daewonyoon commented on 17 Mar
nltk 에는 stopword 를 패키지 내에서 지원하는데,
konlpy 에는 한국어 stopword 지원 계획이 없나요?
Assignees
No one assigned
Labels
None yet

한글의 경우

In [14]:
```
f = open('./data/한국어불용어100.txt', 'r')
s = f.read(); f.close()

stop_words = [ txt.split('\t')[:2]  for txt in s.split('\n') ]
stopword   = {}
for txt in stop_words:
    try:    stopword[txt[0]] = txt[1]
    except: pass
stopword
```

Out[14]:
```
{'이': 'NP',
 '있': 'VA',
 '하': 'VV',
 '것': 'NNB',
 '들': 'VV',
 '그': 'MM',
 '되': 'VV',
 '수': 'NNB',
 '보': 'VX',
 '않': 'VX',
 '없': 'VA',
```

Stop Words 사용자 정의 방법

1. **분야별(주식,리포트,연설문) 정제된 자료**를 **수집**한 뒤
2. Token을 나눈 뒤 **빈도대비 중요도**를 측정
3. **빈도대비 중요도** 측정결과를 저장 및 활용
4. 측정은 후술할 **tf-idf** 등을 활용

Phrase N-Gram

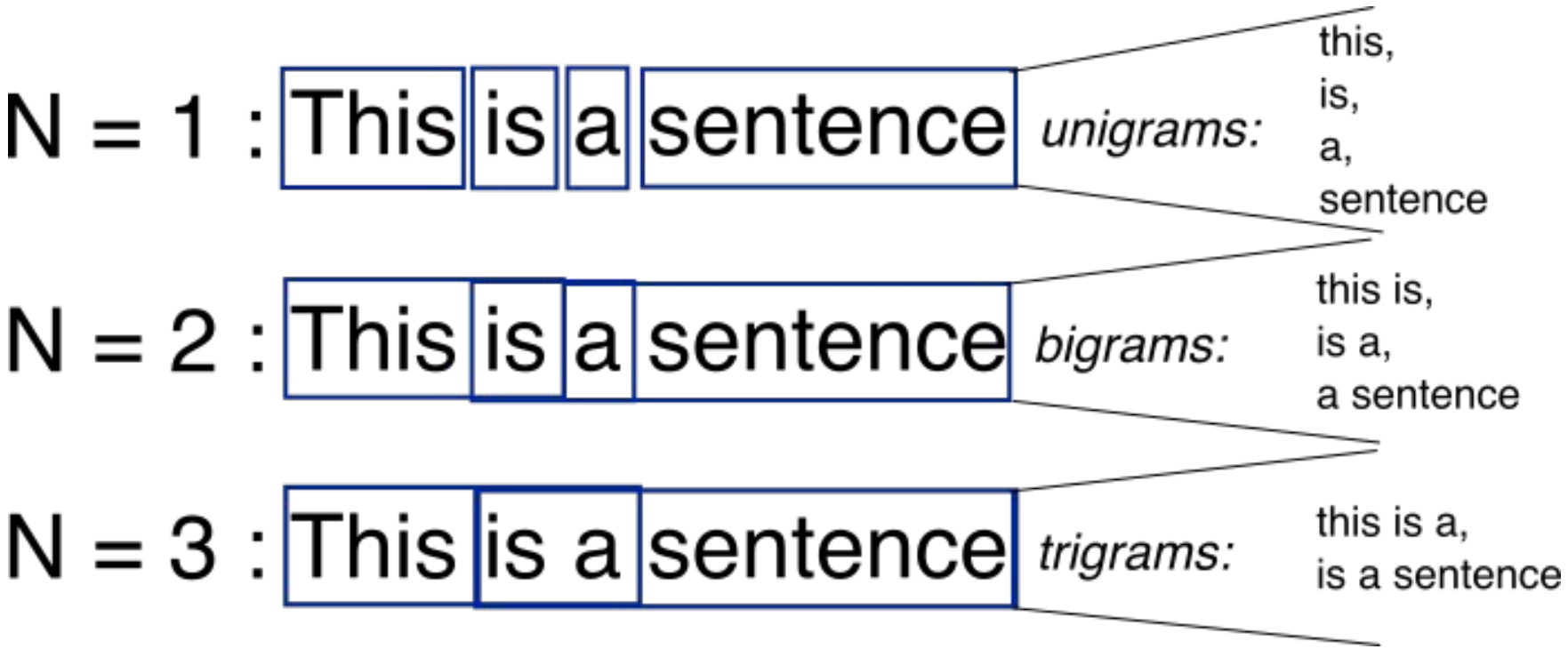
N-Gram

Token 집합

N Gram

1. **문장, 절**의 비교 분석시 **token 갯수 일치**가 필요
2. **독립적 / 균일한 관리** 가능한, **분류기준**이 필요

N Gram


N = 1 : This is a sentence
unigrams:
this,
is,
a,
sentence
N = 2 : This is a sentence
bigrams:
this is,
is a,
a sentence
N = 3 : This is a sentence
trigrams:
this is a,
is a sentence

N Gram

In [1]:
```
# 독일 퀘르버 재단 연설문 : 베를린 선언
f      = open('./data/베를린선언.txt', 'r')
texts_org = f.read()
f.close()
```

In [2]:
```
from nltk.tokenize import word_tokenize
word_token = word_tokenize(texts_org)
```

In [3]:
```
from nltk.util import ngrams

texts_sample = [txt for txt in ngrams(word_token, 3)]
texts_sample[:5]
```

Out[3]:
```
[('존경하는', '독일', '국민'),
 ('독일', '국민', '여러분'),
 ('국민', '여러분', ','),
 ('여러분', ',', '고국에'),
 (',', '고국에', '계신')]
```

PMI

N-gram 상관분석

Point wise Mutual Information

점 위치 상호관계를 활용한 정보

1. **단어간의 거리**를 비교측정하여 **객체의 상관성**을 분석한다
2. 연어 (**근접어:collocation**) 관계 활용하여 분석가능한 객체를 생성한다 (**Bi-gram, Tri-gram**)
3. PMI 는 **단어간의 상관관계 확률론**을 근거로, **단어간의 독립을 가정할 때 발생확률** 과 **문서에서 측정된 동시발생확률** 을 비교하여 상관성을 분석한다

PMI - 한글만 추출

In [27]:
```
from nltk.tokenize import RegexpTokenizer
re_capt = RegexpTokenizer('[가-힣]\w+')
raw_texts = re_capt.tokenize(texts_Berlin_raw)
raw_texts[:10]
```

Out[27]:
```
['존경하는', '독일', '국민', '여러분', '고국에', '계신', '국민', '여러분', '하울젠', '쾨르버재단']
```

In [28]:
```
texts = ''
for txt in raw_texts:
    texts += txt + " "
texts[:200]
```

Out[28]:
```
'존경하는 독일 국민 여러분 고국에 계신 국민 여러분 하울젠 쾨르버재단 이사님과 모드로 동독 총리님을 비롯한 내외 귀빈 여러분 먼저 냉전과 분단을 넘어 통일을 이루고 힘으로 유럽통합과 국제평화를 선도하고 있는 독일과 독일 국민에게 무한한 경의를 표합니다 오늘 자리를 마련해 주신 독일 정부와 쾨르버 재단에도 감사드립니다 아울러 얼마 별세하신 헬무트 총리의 가족'
```

PMI - 한글에 태그를 추가

In [6]:
```
%%time
# 베를린 선언문에 Tag 속성 추가하기
from konlpy.tag import Okt
twitter = Okt()
tagged_words = twitter.pos(texts, stem=True)
```

```
CPU times: user 8.57 s, sys: 230 ms, total: 8.8 s
Wall time: 3.6 s
```

PMI - 연어관계 객체를 생성한 뒤, 객체 상위 PMI Bi-gram 출력

In [30]:
```
from nltk import collocations

finder = collocations.BigramCollocationFinder.from_words(tagged_words)
finder
```

Out[30]:
```
<nltk.collocations.BigramCollocationFinder at 0x7fc643f2ddd8>
```

In [31]:
```
# top 10 n-grams with highest PMI
measures = collocations.BigramAssocMeasures()
finder.nbest(measures.pmi, 10)
```

Out[31]:
```
[(('가능하며', 'Adjective'), ('불가', 'Noun')),
 (('가스', 'Noun'), ('관', 'Noun')),
 (('가운데', 'Noun'), ('현재', 'Noun')),
 (('감사', 'Noun'), ('드립니', 'Verb')),
 (('갖춰', 'Verb'), ('지', 'PreEomi')),
 (('같은', 'Adjective'), ('공감', 'Noun')),
 (('거나', 'Eomi'), ('깨져', 'Verb')),
 (('건너지', 'Verb'), ('않기', 'Verb')),
 (('걷어', 'Verb'), ('차는', 'Verb')),
 (('검증', 'Noun'), ('가능하며', 'Adjective'))]
```

PMI - 연어관계 객체를 생성한 뒤, 객체 상위 PMI Tri-gram 출력

In [32]:
```
finder = collocations.TrigramCollocationFinder.from_words(tagged_words)
finder
```

Out[32]:
```
<nltk.collocations.TrigramCollocationFinder at 0x7fc643f2db70>
```

In [33]:
```
measures = collocations.TrigramAssocMeasures()
finder.nbest(measures.pmi, 10)
```

Out[33]:
```
[(('가능하며', 'Adjective'), ('불가', 'Noun'), ('역적', 'Noun')),
 (('가스', 'Noun'), ('관', 'Noun'), ('연결', 'Noun')),
 (('가운데', 'Noun'), ('현재', 'Noun'), ('생존', 'Noun')),
 (('같은', 'Adjective'), ('공감', 'Noun'), ('대', 'Suffix')),
 (('거나', 'Eomi'), ('깨져', 'Verb'), ('서도', 'Noun')),
 (('검증', 'Noun'), ('가능하며', 'Adjective'), ('불가', 'Noun')),
 (('견', 'Noun'), ('지하', 'Noun'), ('면서', 'Noun')),
 (('과도', 'Josa'), ('같은', 'Adjective'), ('공감', 'Noun')),
 (('들어서는', 'Verb'), ('대전', 'Noun'), ('환', 'Noun')),
 (('록', 'Eomi'), ('앞장서서', 'Verb'), ('돕겠', 'Verb'))]
```

Tf-idf

상대빈도분석

Term Frequency-Inverse Document Frequency

1. 문서의 내용을 쉽게 벡터로 표현하는 고전적 방식
2. **Term Frequency** : 특정 용어의 발생빈도
3. (문서 Token 출현빈도) / (문서 전체 Token 갯수)
4. **Inverse Document Frequency** (문서 빈도의 역)
5. 문서를 이해하는데 Token의 중요도
6. **일반 문서대비 출현빈도**

Term Frequency-Inverse Document Frequency

1. **TF**는 해당 문서만 있으면 바로 연산이 가능하지만
2. **IDF**는 모집단의 Corpus 별 통계값 (일반문서의 Token 출현빈도)이 있어야 연산가능

Term Frequency-Inverse Document Frequency

	Commonly Used Terms				
Tag Name	food	cakes	cooking	song	music
itunes	0.00019	0.00062	0.00080	0.02106	0.02319
food	0.04970	0.02544	0.01890	0.00024	0.01136
podcast	0.00842	0.00120	0.01538	0.01568	0.02053
government	0.01080	0.00108	0.00592	0.00022	0.00340
comics	0.00364	0.00282	0.00055	0.00229	0.01021
baking	0.04971	0.02544	0.02147	0.00018	0.00125
cooking	0.04450	0.02249	0.02347	0.00018	0.00108
dessert	0.02290	0.02739	0.02819	0.01488	0.01337
recipe	0.02162	0.00956	0.01830	0.00019	0.00019
foodblog	0.04971	0.01034	0.01538	0.00019	0.00013

Term Frequency-Inverse Document Frequency

	name character	token_stem character	count_per_doc integer	count_of_docs integer	tfidf numeric
111	Akebono Brake Corporation	act_euro	1	1	0.0809
112	Akebono Brake Corporation	act_euro_flyer	1	1	0.0809
113	Akebono Brake Corporation	aftermarket	1	2	0.0687
114	Akebono Brake Corporation	aftermarket_brake	1	1	0.0809
115	Akebono Brake Corporation	aftermarket_brake_usa	1	1	0.0809
116	Akebono Brake Corporation	akebono	3	1	0.2426
117	Akebono Brake Corporation	akebono_brake	2	1	0.1617
118	Akebono Brake Corporation	akebono_brake_trust	1	1	0.0809
119	Akebono Brake Corporation	akebono_proud	1	1	0.0809
120	Akebono Brake Corporation	akebono_proud_manufactur	1	1	0.0809

TF - IDF (트럼프 취임연설문 불러오기 / 전처리)

In [1]:
```
f = open('./data/trump.txt', 'r')
texts = f.read()
f.close()
```

In [2]:
```
texts = texts.lower()

from nltk.tokenize import RegexpTokenizer
re_capt = RegexpTokenizer(r'[a-z]\w+')
token   = re_capt.tokenize(texts)
token[:10]
```

Out[2]:
```
['chief',
 'justice',
 'roberts',
 'president',
 'carter',
 'president',
 'clinton',
 'president',
 'bush',
 'president']
```

TF - IDF (Stop Word)

In [3]:
```
from nltk.corpus import stopwords
stopwords_eng = stopwords.words('english')
stopwords_eng[:10]
```

Out[3]:
```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your']
```

In [4]:
```
texts = [txt    for txt in token
                if txt not in stopwords_eng]
document = ''
for txt in texts:
    document += txt + ' '
document[:500]
```

Out[4]:
```
'chief justice roberts president carter president clinton president bush president obama fel
low americans people world thank citizens america joined great national effort rebuild count
ry restore promise people together determine course america world many many years come face
challenges confront hardships get job done every four years gather steps carry orderly peace
ful transfer power grateful president obama first lady michelle obama gracious aid throughou
t transition magnificent thank today cere'
```

TF - IDF (Tf-idf 학습객체 생성)

In [6]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_vec   = TfidfVectorizer()
transformed = tfidf_vec.fit_transform(raw_documents = [document])
transformed
```

Out[6]:
```
<1x455 sparse matrix of type '<class 'numpy.float64'>'
        with 455 stored elements in Compressed Sparse Row format>
```

In [7]:
```
import numpy as np
transformed = np.array(transformed.todense())

index_value   = {i[1]:i[0] for i in tfidf_vec.vocabulary_.items()}
fully_indexed = {index_value[column]:value  for row in transformed
                                            for (column,value) in enumerate(row)}
```

TF - IDF (Numpy 결과를 Pandas Series 객체로 변환/ 조회하기)

In [8]:
```
import pandas as pd
tfidf = pd.Series(fully_indexed).sort_values(ascending=False)
tfidf[:10]
```

```
/usr/lib/python3.6/importlib/_bootstrap.py:219: RuntimeWarning: numpy.dtype size changed, may indicate binary incom
patibility. Expected 96, got 88
  return f(*args, **kwds)
/usr/lib/python3.6/importlib/_bootstrap.py:219: RuntimeWarning: numpy.dtype size changed, may indicate binary incom
patibility. Expected 96, got 88
  return f(*args, **kwds)
```

Out[8]:
```
america     0.423714
american    0.233042
people      0.211857
country     0.190671
nation      0.190671
one         0.169485
every       0.148300
great       0.127114
never       0.127114
new         0.127114
dtype: float64
```

In [9]:
```
tfidf['obama']
```

Out[9]:
```
0.06355703979105537
```