In [9]:

```
from os import listdir

def fileids(path, ext='txt'):
    return [path+file for file in listdir(path) if file.split(".")[-1] == ext]

len(fileids(""))
```

Out[9]:

102

In [20]:

```
def filecontent(file):
    with open(file, encoding="utf-8") as fp:
        content = fp.read()
    return content
```

In [6]:

```
def ngram(term, n=2):
    return [term[i:i+n] for i in range(len(term) - n + 1)]
```

In [18]:

```
import re
from string import punctuation

pattern = dict()

#구두점
pattern1 = re.compile(r"[{0}]".format(re.escape(punctuation)))
#corpus = pattern1.sub(" ", corpus)
pattern["punc"] = pattern1
# print(len(corpus))
# corpus

#불용어
pattern2 = re.compile(r"[A-Za-z0-9]{7,}")
#corpus = pattern2.sub(" ", corpus)
pattern["stop"] = pattern2
#print(len(corpus))

#이메일
pattern3 = re.compile(r"Ww{2,}@Ww{3,}.(.Ww{2,})+")
pattern3 = re.compile(r"Ww{2,}@(.?Ww{2,})+")
#corpus = pattern3.sub("", corpus)
pattern["email"] = pattern3

#도메인
pattern4 = re.compile(r"(.?Ww{2,}){2,}")
#corpus = pattern4.sub("", corpus)
pattern["domain"] = pattern4

#한글 이외
pattern5 = re.compile(r"[^가-힣0-9]+")
#corpus = pattern5.sub(" ", corpus)
pattern["nonkorean"] = pattern5

#반복되는 공백문자
pattern6 = re.compile(r"Ws{2,}")
#corpus = pattern6.sub(" ", corpus)
pattern["whitespace"] = pattern6
```

In [122]:

```
newslist = list()

for i in range(0,102):
    content = filecontent(fileids("")[i])
    #print(content)
    for _ in ["email", "punc", "stop", "whitespace"]:
        content = pattern[_].sub(" ", content)
        newslist.append(content)

newslist
```

Out[122]:

```
['스타벅스의 디지털 혁신 뒤에는 MS 있었다스타벅스의 디지털 혁신 뒤에는 마이크로소프트(MS)가 있었다.WnWn김영욱 한국마이크로소프트 부장이 16일 한국마이크로소프트 서울 광화문 본사에서 열린 '빌드 2019' 미디어 디브리핑을 하고 있다. /이정민 기자Wn Wn16일 한국MS 서울 광화문 본사에서 열린 '빌드 2019' 미디어 디브리핑(보고·설명)에서 김영욱 한국MS 부장은 "스타벅스는 제품부터 고객, 파트너, 공급자에 이르는 모든 디지털 혁신을 애저와 애저 IoT, 애저 스피어, 애저 AI, 애저 블록체인을 통해 이뤘다"고 말했다.WnWn빌드 2019는 지난주 미국 시애틀에서 열린 MS의 연례 개발자 콘퍼런스다. 빌드 2019에서 사티아 나델라 MS 최고경영자(CEO)가 스타벅스 사례를 직접 소개한 바 있다.WnWn우선 세계 각지의 커피 농장에서 생산되는 원두가 전세계 3만개 이상의 스타벅스 매장에 도착하기까지의 모든 과정은 블록체인으로 관리된다. 원두 산지에서부터 최종 포장까지의 모든 과정을 추적한다. 고객은 이를 통해 자신이 마시는 커피 원두의 출처부터 포장시기까지 알 수 있다.WnWn전세계 스타벅스 매장의 커피 머신에는 사물인터넷(IoT) 기술이 적용돼 있다. 수온이나 압력 등이 최적의 상태를 유지할 수 있도록 하며 클라우드를 통해 레시피가 자동으로 업데이트되기도 한다.WnWn스타벅스는 MS의 애저와 애저 AI를 활용해 지능형 메뉴추천시스템 '딥 브루(Deep Brew)'도 개발했다. AI로 사용자의 성향을 파악해 메뉴를 추천하거나 날씨와 매장별, 시간대별 인기 메뉴를 추천해 주기도 한다.WnWn김영욱 부장은 올해 빌드의 키워드로 △프라이버시와 보안 △AI와 클라우드 △개방성 △디지털 트랜스포메이션 등을 꼽았다.WnWn김 부장은 "이번 빌드를 통해
```

In [44]:

```
from nltk.tokenize import word_tokenize
from konlpy.tag import Komoran
from string import punctuation
import re
from collections import defaultdict

ma = Komoran()

#content = filecontent(fileids("news")[-2])

indexTerm = defaultdict(int)
indexTerm1 = defaultdict(int)
indexTerm2 = defaultdict(int)
indexTerm3 = defaultdict(int)
indexTerm4 = defaultdict(int)
indexTerm5 = defaultdict(int)

for i in range(0,102):
    for term in newslist[i].split():
        indexTerm1[term] += 1

for _ in indexTerm1:
    for t in ma.pos(_):
        indexTerm2[t] += 1
        indexTerm[t] += 1 #원시형태소 = 품사
        if len(t[0]) > 1: #음절 길이로 정규화
            indexTerm3[t[0]] += 1 # 원시형태소
            indexTerm[t[0]] += 1
        if t[1].startswith("N"):
            indexTerm4[t[0]] += 1 #명사
            indexTerm[t[0]] += 1
        for n in ngram(t[0]): #바이그램
            indexTerm5[n] += 1
            indexTerm[n] += 1

print(len(indexTerm1),len(indexTerm2), len(indexTerm3),len(indexTerm4), len(indexTerm5), len(indexTe
# min(indexTerm5.values()), max(indexTerm5.values()),
min(indexTerm.values()), max(indexTerm.values()) )

indexTerm
```

```
4143 2576 1998 1750 2448 5847 1 343
```

Out[44]:

In [50]:

```
for i in range(0,102):
    tokens1 = newslist[i].split() #원시어절
    tokens2 = word_tokenize(newslist[i]) #[word_tokenize(content) for token in tokens1] #구두점 분리
    tokens3 = [_ for token in tokens2 for _ in ma.pos(token)] #형태소-품사
#     tokens 1 + tokens2 + tokens3 => Lexicon(controlled vocab)
    tokens4 = [token[0] for token in tokens3]
    tokens5 = [token[0] for token in tokens3 if token[1].startswith("N")]
    tokens6 = [_ for token in tokens4 for _ in ngram(token)]
    #print(content, "₩n")
    #print(len(content.split()))
print(len(tokens1))
print(len(tokens2))
print(len(tokens3))
print(len(tokens4))
print(len(tokens5))
print(len(tokens6))
print(len(tokens1 + tokens2 + tokens3))
print(len(tokens1 + tokens2 + tokens3 + tokens4 + tokens5 + tokens6))
print(len(set(tokens1 + tokens2 + tokens3 + tokens4 + tokens5 + tokens6)))
```

```
83
83
164
164
70
109
330
673
311
```

In [103]:

```
from konlpy.corpus import kobill
from konlpy.tag import Komoran
from math import sqrt

# ma = Komoran()

DTM = defaultdict(lambda:defaultdict(int))
for idx in fileids(""):
    for term in filecontent(idx).split():
        for token in ma.pos(term):
            DTM[idx]["/".join(token)] += 1
```

In [104]:

```
TDM = defaultdict(lambda:defaultdict(int))
for idx, termList in DTM.items():
    for term, freq in termList.items():
        TDM[term][idx] = freq

TWM = defaultdict(lambda:defaultdict(float))
DVL = defaultdict(float)
N = len(DTM)
for idx, termList in DTM.items():
    maxTF = max(termList.values()) #단어 빈도
    for term, freq in termList.items():
        TF = freq / maxTF
        IDF = log2(N/len(TDM[term]))
        TWM[term][idx] = TF*IDF
        DVL[idx] += TWM[term][idx]**2

for idx, length in DVL.items():
    DVL[idx] = sqrt(length)

DVL #벡터의 길이
```

Out[104]:

```
defaultdict(float,
            {'F:/news/IT201905160.txt': 9.346882316101825,
             'F:/news/IT201905161.txt': 5.538483419147847,
             'F:/news/IT201905610.txt': 6.323355140201782,
             'F:/news/IT2019051611.txt': 9.948683254314824,
             'F:/news/IT2019051612.txt': 10.180963297911923,
             'F:/news/IT2019051613.txt': 10.568651526046676,
             'F:/news/IT2019051614.txt': 10.306342890495447,
             'F:/news/IT2019051615.txt': 3.538750307285323,
             'F:/news/IT2019051616.txt': 5.080412316158881,
             'F:/news/IT201905162.txt': 11.476904452559806,
             'F:/news/IT201905163.txt': 12.962504243877193,
             'F:/news/IT201905164.txt': 8.154237392019796,
             'F:/news/IT201905165.txt': 8.157925332237566,
             'F:/news/IT201905166.txt': 6.420376924269586,
             'F:/news/IT201905167.txt': 5.27438807983166,
             'F:/news/IT201905168.txt': 11.332953659876795,
             'F:/news/IT201905169.txt': 9.590892942477419,
```

In [114]:

```
query = "서울시에 거래되는 아파트의 전세값은?"
#의 와 에 때문에 문제의 소지가 있기는 함 그래도 그 셋이 W->0이 되기는 함

TQM = defaultdict(int)
QWM = defaultdict(float)

for term in query.split():
    for token in ma.pos(term):
        TQM["/".join(token)] += 1

alpha = 0.5
maxTF = max(TQM.values())
for term, freq in TQM.items():
    TF = alpha + (1-alpha)*(freq / maxTF)
    DF = len(TWM[term]) if len(TWM[term]) > 0 else 1
    IDF = log2(N/DF)
    QWM[term] = TF*IDF #IDF = N / 1

#IDF 관건
TQM
#QWM
maxTF
```

Out[114]:

```
1
```

In [115]:

```
candidateList = defaultdict(float)

for term, weight1 in QWM.items():
    for doc, weight2 in TWM[term].items():
        innerProduct = weight1 * weight2
        candidateList[doc] += innerProduct

for doc, sim in candidateList.items():
    candidateList[doc] = sim / DVL[doc]
```

In [118]:

```
from nltk.tokenize import sent_tokenize

k = 5
for doc, sim in sorted(candidateList.items(), key=lambda x:x[1], reverse=True)[:k]:
    print("문서이름:{0} / 유사도:{1:.4f}".format(doc, sim))
    print(sent_tokenize(filecontent(doc))[:3])
    print()
```

```
문서이름:F:/news/경제201905165.txt / 유사도:1.0317
['[환율마감] 원/달러 2.9원↑원/달러 환율이 상승(원화가치 하락) 마감했다.', '16일 서
울 외환시장에서 오후 3시30분 기준 원/달러 환율은 전 거래일 대비 2.9원 오른 1191.5원
에 거래를 마쳤다.', '이날 원/달러 환율은 전 거래일보다 1.6원 내린 1187.0원으로 출발
해 1191.5원에 마감했다.']

문서이름:F:/news/국제201905161.txt / 유사도:1.0200
['화웨이, 美 W'거래금지W'에 발끈.."소송도 불사"도널드 트럼프 미 행정부가 중국 화웨이
를 '거래 제한 명단' 에 올리겠다고 발표하자 화웨이가 강하게 반발했다고 16일 중국 환
구시보가 전했다.', '미국 동부 시간으로 15일 미 상무부는 화웨이와 70개 계열사를 거래
제한 기업 명단에 올린다고 발표했다.', '이날 화웨이는 "우리는 미 정부와 소통해 제품
안전을 보장하기를 희망한다.']

문서이름:F:/news/정치2019051611.txt / 유사도:0.7531
['[뉴스현장] 트럼프, 6월 하순 방한..문 대통령과 정상회담 外<출연 : 정대진 아주대 통
일연구소 교수>WnWn트럼프 미국 대통령이 내달 하순 일본에서 개최되는 주요 20개국 정상
회의를 계기로 방한해 문 대통령과 정상회담을 하기로 했습니다.', '어떤 논의들이 이뤄질
지 관심이 집중됩니다.', '자세한 내용 정대진 아주대 통일연구소 교수와 짚어보겠습니
다.']

문서이름:F:/news/경제201905168.txt / 유사도:0.4054
['[마감시황] 외국인 매도에..코스피 하락 마감[서울경제] 코스피가 16일 외국인의 매도에
하락 마감했다.', '이날 코스피는 전 거래일보다 25.09포인트(1.20%) 내린 2,067.69로 거
래를 마쳤다.', '지수는 전 거래일보다 2.10포인트(0.10%) 오른 2,094.88로 출발해 등락을
거듭하다 하락 마감했다.']

문서이름:F:/news/경제2019051612.txt / 유사도:0.3980
['4대은행, 1분기 정규직 1000여명 짐쌌다..비대면거래 확대, 점포폐쇄 등 영향비대면 거
래확대, 영업점 통폐합, 희방퇴직 등으로 올해 1·4분기 4대 시중은행의 정규직 약 1000명
이 짐을 싼 것으로 나타났다.', '16일 신한·KB국민·우리·KEB하나 등 4대 시중은행들의
분기보고서를 분석한 결과 1·4분기 기준 정규직 직원수는 총 5만6120명으로 집계됐다.',
'이는 지난해 말(5만7082명)과 비교해 962명 줄어든 것이다.']
```

In [119]:

```
candidateList = defaultdict(float)

for term, docList in TWM.items():
    for doc, weight1 in TWM[term].items():
        weight2 = QWM[term]
        candidateList[doc] += (weight1 - weight2)**2
        print(term, doc, (weight1 - weight2)**2)

for doc, sim in candidateList.items():
    candidateList[doc] = sqrt(sim)
```

```
스타벅스/NNP F:/news/IT201905160.txt 9.695741054511057
의/JKG F:/news/IT201905160.txt 0.0007313014982271662
의/JKG F:/news/IT201905161.txt 0.09028413558360071
의/JKG F:/news/IT2019051610.txt 0.03834767666090132
의/JKG F:/news/IT2019051612.txt 0.036438552160799956
의/JKG F:/news/IT2019051613.txt 0.08933627720725212
의/JKG F:/news/IT2019051614.txt 0.12597810965553913
의/JKG F:/news/IT2019051615.txt 0.04298427695135231
의/JKG F:/news/IT2019051616.txt 0.064274545742622
의/JKG F:/news/IT201905162.txt 0.0925553458693757
의/JKG F:/news/IT201905164.txt 0.10255439985526392
의/JKG F:/news/IT201905165.txt 0.035833773413131126
의/JKG F:/news/IT201905166.txt 0.0925553458693757
의/JKG F:/news/IT201905167.txt 0.11520054525904613
의/JKG F:/news/IT201905168.txt 0.06799984594484745
의/JKG F:/news/경제201905160.txt 0.0731301498227166
의/JKG F:/news/경제201905161.txt 0.08529900675321661
의/JKG F:/news/경제2019051610.txt 0.09551693038069105
의/JKG F:/news/경제2019051611.txt 0.024340656375904185
```

In [111]:

```
for doc in DTM:
    print(len(DTM[doc]), sum(DTM[doc].values()))
candidateList
```

```
212 483
224 583
342 806
85 136
225 471
222 527
127 341
470 1214
119 176
166 466
86 161
235 530
391 1058
267 672
456 1287
187 435
4 4
94 153
283 672
```

In [120]:

```
k = 5
for doc, sim in sorted(candidateList.items(), key=lambda x:x[1], reverse=True)[:k]:
    print("문서이름:{0} / 거리:{1:.4f}".format(doc, sim))
    print(sent_tokenize(filecontent(doc))[:3])
    print()
```

```
문서이름:F:/news/국제201905166.txt / 거리:22.1727
["Korea's auto output, exports up but domestic sales down in April"]

문서이름:F:/news/국제2019051612.txt / 거리:19.7065
['British rock band Queen coming to Korea for concert']

문서이름:F:/news/국제2019051610.txt / 거리:19.2278
['KITA sets up advisory panel on trade in Washington']

문서이름:F:/news/국제201905168.txt / 거리:17.1175
['Samsung SDI showcases optimized ESS in Europe']

문서이름:F:/news/국제201905169.txt / 거리:15.3807
["'교육예산 삭감'에 항의하는 브라질 시위대(상파울루 AP=연합뉴스) 15일(현지시간) 브라
질 상파울루에서 진행된 정부의 교육예산 삭감에 반대하는 시위.", '자이르 보우소나루 대
통령 정부가 재정적자 완화 방안의 하나로 교육예산을 삭감하겠다고 밝힌 가운데 이날 오
전 수도 브라질리아와 최대 도시 상파울루 등에서 시작된 시위는 오후 들어 전국의 대도시
로 번졌다.', 'leekm@yna.co.kr']
```

정제한 데이터에서 코사인 / 유사도 검색하는 거 생각해보기