

유튜브 댓글을 확인해 보자

01 사전 작업

- 한글 폰트 적용, konlpy 설치

목차

01 라이브러리 불러오기

02 텍스트 데이터 분석

03 텍스트 데이터 시각화-워드 클라우드

01 라이브러리 불러오기 및 한글 폰트 설치

목차로 이동하기

```
In [1]: ### 나눔 고딕 설치
!apt-get update -qq # 설치를 업데이트
!apt-get install fonts-nanum* -qq # 설치한다. fonts-nanum*
```

```
Selecting previously unselected package fonts-nanum.
(Reading database ... 121913 files and directories currently installed.)
Preparing to unpack .../fonts-nanum_20200506-1_all.deb ...
Unpacking fonts-nanum (20200506-1) ...
Selecting previously unselected package fonts-nanum-coding.
Preparing to unpack .../fonts-nanum-coding_2.5-3_all.deb ...
Unpacking fonts-nanum-coding (2.5-3) ...
Selecting previously unselected package fonts-nanum-eco.
Preparing to unpack .../fonts-nanum-eco_1.000-7_all.deb ...
Unpacking fonts-nanum-eco (1.000-7) ...
Selecting previously unselected package fonts-nanum-extra.
Preparing to unpack .../fonts-nanum-extra_20200506-1_all.deb ...
Unpacking fonts-nanum-extra (20200506-1) ...
Setting up fonts-nanum-extra (20200506-1) ...
Setting up fonts-nanum (20200506-1) ...
Setting up fonts-nanum-coding (2.5-3) ...
Setting up fonts-nanum-eco (1.000-7) ...
Processing triggers for fontconfig (2.13.1-4.2ubuntu5) ...
```

```
In [4]: import matplotlib.pyplot as plt
import matplotlib.font_manager as fm
import matplotlib as mpl
import numpy as np

# 폰트 파일 경로
path = '/usr/share/fonts/truetype/nanum/NanumGothicEco.ttf'

# 폰트 프로퍼티 생성
font_prop = fm.FontProperties(fname=path, size=10)
font_name = font_prop.get_name()
print(font_name) # NanumGothic Eco

# 폰트 매니저에 폰트 추가
```

```

fm.fontManager.addfont(path)

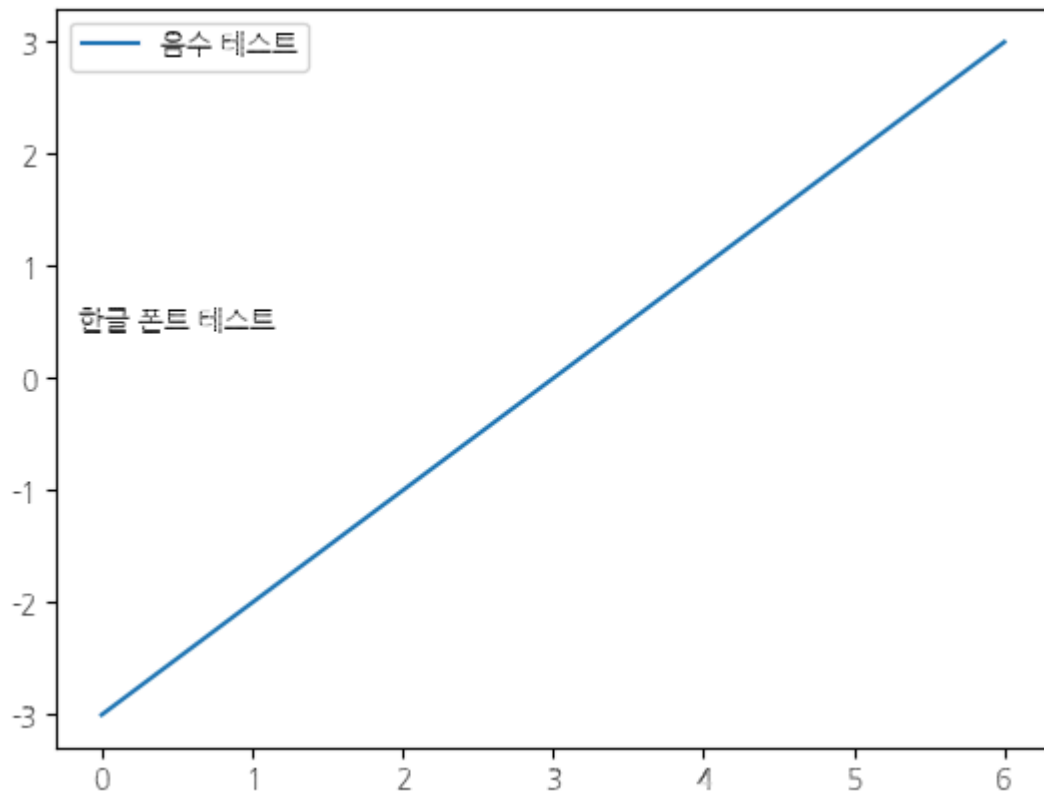
# matplotlib에 폰트 설정
plt.rc('font', family=font_name)

# 음수 표시되도록 설정
mpl.rcParams['axes.unicode_minus'] = False

# 예제 플롯
plt.figure()
plt.plot([-3, -2, -1, 0, 1, 2, 3], label='음수 테스트')
plt.text(0.5, 0.5, '한글 폰트 테스트', ha='center', va='center', fontproperties=
plt.legend()
plt.show()

```

NanumGothic Eco



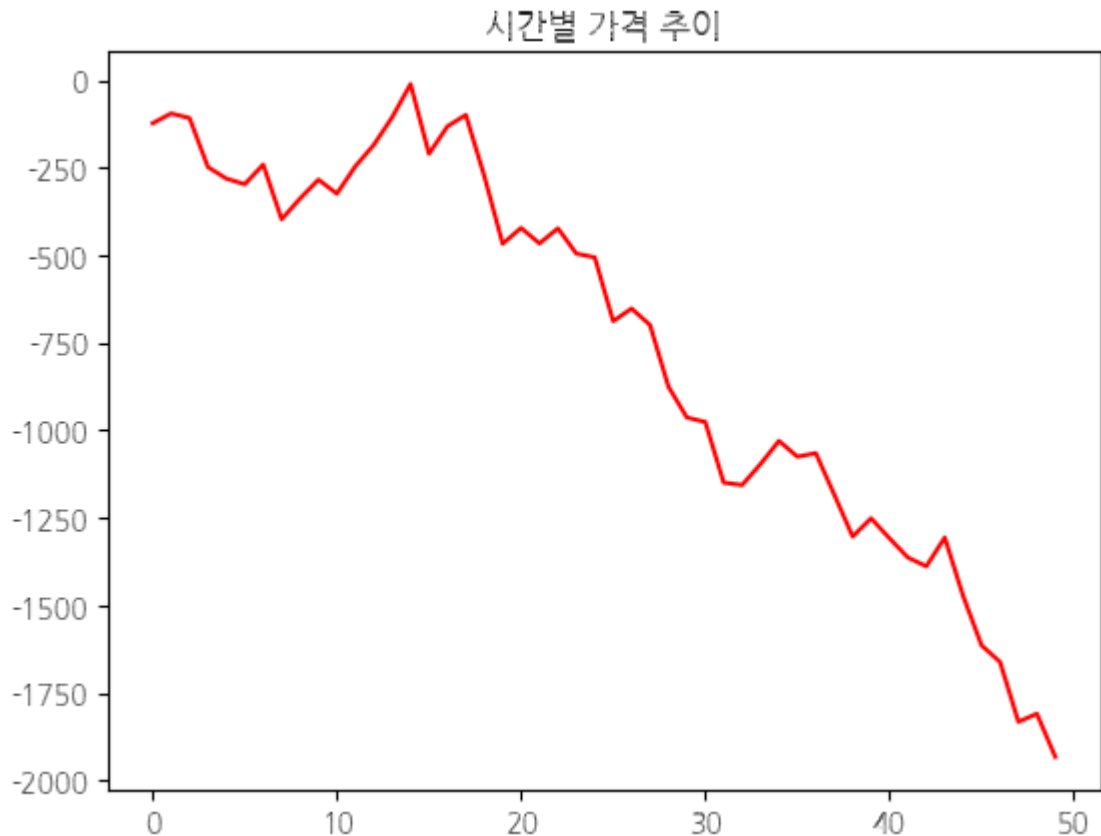
```

In [5]: # 데이터 준비
data = np.random.randint(-200, 100, 50).cumsum()

# 그래프를 그려 한글 확인
plt.plot(range(50), data, 'r')
plt.title('시간별 가격 추이')

```

Out[5]: Text(0.5, 1.0, '시간별 가격 추이')



In [6]: `pip install konlpy`

```
Collecting konlpy
  Downloading konlpy-0.6.0-py2.py3-none-any.whl (19.4 MB)
    ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 19.4/19.4 MB 36.4 MB/s eta 0:00:00
Collecting JPYpe1>=0.7.0 (from konlpy)
  Downloading JPYpe1-1.5.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (488 kB)
    ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 488.6/488.6 kB 35.8 MB/s eta 0:00:00
0
Requirement already satisfied: lxml>=4.1.0 in /usr/local/lib/python3.10/dist-packages (from konlpy) (4.9.4)
Requirement already satisfied: numpy>=1.6 in /usr/local/lib/python3.10/dist-packages (from konlpy) (1.25.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from JPYpe1>=0.7.0->konlpy) (24.0)
Installing collected packages: JPYpe1, konlpy
Successfully installed JPYpe1-1.5.0 konlpy-0.6.0
```

02 텍스트 데이터 분석

목차로 이동하기

In [8]: `doc_ko = open("youtubu_comment.csv").read()
doc_ko[1:1000]`

Out[8]: '글\너릴 땀슈퍼맨이 인생 히어로였는데 언제부터가 아이언맨이 되버린..ㅋㅋ\너"형, 3000만
 큼 사랑해!"\너스탠리복장정리할 거라고 믿습니다. 그게 약속이니 까음!!\너{내가 아이언맨을 제
 일 좋아하는 히어로인 이유}1.원작코믹스를 재밌게 봄2.내 인생 첫 히어로였음3.아이언맨이 나오
 는 영화마다 다른 슈트임4.개그를 재밌게 많이 함5.로다주여서6.로다주여서7.로다주여서8.
 역시로다주여서\n7:13에서 토르의 번개를 맞고 충전이 200프로까지 된 이유 아이언맨2에서 위플
 레쉬의 전기공격에 계속 과부하되는 문제점을 겪어 이를 보완하고 자업그레이트한 것\n이번에는
 어벤져스 무기와 장비를 알아봐요\n근데 이거 진짜 모르는 사람 많은데 마크원 날다가 부해돼서 떨어
 어진 건 제작미흡이 아니라 나사를 시간이 없어서 대충 조여서 그런 영화 대사 중에 잉센이 시간이 없
 다고 하니 나사를 대충 쪼이라 했음\nI....am....Ironman.\n예전에 로다주가 촬영 끝나고 아
 이언맨 슈트 중 하나를 그대로 입고도 망쳤다 스탭들한테 다시 잡혀왔다는데 로다주가 인터뷰 중
 가장 갖고 싶은 슈트란 질문에 마크42라고 답한 걸 보면 그 슈트가 마크42일 가능성이 높겠네요 ㅋㅋ
 \n토니 안 죽었으면 갓킬러도 리뷰했을 거 같다\n중간에 마크41 설명하는데서 이름에 마크39로
 잘못 나왔네요 ㅎㅎ\n인피니티워 예고편에 헬크버스터 탁!!하고 날라왔을 때 진짜 마크50은 눈길
 조차도 안하고 에이조브울트론에서 워낙 헬크버스터가 비중이 많았기에 인피니티워에도 등장
 해서 헬크버스터랑 아이언 스파이더 전투 씬 만기대했는데...어라??마크50번 신되는 장면 보고 진
 짜기 절했다 🤔🤔\n마크50에서 나노 입자 나올 때 호로호를 ㅇ허덜 ㄹ러 ㅈ ㅎ덜 ㄹ러 ㅈ ㅎ허덜 이러
 면서 나온답니다.\n다음은 타노스 복장 총정리 ㅋㅋㅋㅋㅋㅋㅋㅋ\n다음은!?~닥터 스트
 레인지가즈아~~~ㅋㅋㅋㅋㅋㅋ\n울트론이랑 싸울 때 헬크버스터를 입었다면?\n15:32 스타
 크에겐 여러 인공지능이 있었군요. 다음 영상 주제.\n마블 뽕이 빠져서 구독자랑 조회수가 안느네
 형..ㅈㅈ 블랙위도우 개봉이 빨리 되면 좋겠당\n4:17 대표적인 오류입니다 이때 아이언맨이 맞은
 건 탱크의 달려있는 기관총입니다\n제일 잘 정리하신 거 같네요 잘 보고 갑니다\n다음엔 온두화 살
 정리해주세요 ㄱㄱㄱ\n메일 목록도리의 복장정리를 해주세요\n마블에 관련된 총정리는 메일 목
 도리님 영상이 제일 재밌어용~ㅎㅎ\n사실 마크85의 하이라이트는 인피니티 스톤들의 힘을 견딜 수
 있는 내구력이라고 생각함 우주에서 최고라고 할 수 있는 니다벨리르'

```
In [9]: import nltk
from konlpy.tag import Kkma      ### 꼬꼬마
from konlpy.tag import Hannanum  ### 한나눔

# OKT 클래스를 이용한 명사확인
from konlpy.tag import Okt       ### Okt
```

```
In [10]: t = Kkma()
```

```
In [11]: %time

doc_nouns = t.nouns(doc_ko)
doc_nouns
```

CPU times: user 1min 57s, sys: 1.47 s, total: 1min 59s
 Wall time: 1min 45s

```
Out[11]: ['땃',
          '땃글',
          '글',
          '슈퍼맨',
          '아이언',
          '턴',
          '인생',
          '인생히어로',
          '히',
          '어로',
          '맨',
          '형',
          '3000',
          '3000만큼 사랑',
          '만큼',
          '사랑',
          '스탠리',
          '스탠리복장정리',
          '복장',
          '정리',
          '거',
          '약속',
          '음',
          '내가',
          '내가아이언',
          '히어로',
          '이유',
          '1',
          '원작',
          '원작코믹스',
          '코믹스',
          '봄',
          '봄2',
          '2',
          '내',
          '3',
          '영화',
          '슈트',
          '슈트임4',
          '임',
          '4',
          '개그',
          '함',
          '함5',
          '5',
          '다주',
          '다주여',
          '여',
          '6',
          '7',
          '8',
          '역',
          '역시',
          '시',
          '13',
          '토',
          '토르',
          '르',
          '번개',
          '충전',
```

'200',
'200프로',
'프로',
'이유아이언맨',
'아이',
'언',
'계속',
'부화',
'보완',
'문제점',
'위',
'위플레쉬',
'플레쉬',
'전기',
'전기공격',
'공격',
'업그레이드',
'이번',
'기와',
'기와장비',
'장비',
'날',
'부',
'건',
'중',
'마크',
'미흡',
'나사',
'시간',
'마크원',
'날다가',
'건제작',
'영화대사',
'다가',
'제작',
'원',
'이거',
'대사',
'사람',
'예전',
'촬영',
'데로',
'스태프',
'인터뷰',
'중하나',
'예전으로',
'아이언맨',
'하나',
'에로',
'질문',
'마크42',
'42',
'답',
'마크42알가능성',
'일',
'가능성',
'갓',
'갓킬러',
'니',
'킬러',

'리뷰',
'중간',
'마크41설명',
'41',
'데',
'설명',
'이름',
'마크39',
'39',
'인',
'예고편',
'헐크',
'헐크버스터',
'버스터',
'때',
'때진짜마크50',
'진짜',
'50',
'안',
'티',
'오브',
'비중',
'전투',
'에이지',
'스파이',
'티워에',
'눈',
'워에',
'조',
'등장',
'트론에',
'조차도',
'차도',
'씬',
'기대',
'마크50변신',
'변신',
'장면',
'장면보고',
'보고',
'기절',
'마크50',
'홀',
'호로',
'홀ㅇ허럴ㄹ',
'ㅇ',
'허',
'나노',
'나노입자',
'랄ㄹ',
'입자',
'러',
'러ㅏㅎ랄ㄹ',
'ㅏ',
'ㅎ',
'러ㅏ',
'허랄',
'다음',
'노',
'노스복장총정리',

'스복',
'장총',
'닥터',
'닥터스트레인지가즈',
'스트레인',
'지가',
'즈',
'트론',
'크버스',
'크버스터',
'터',
'15',
'32',
'32스타크',
'인공',
'스타크',
'인공지능',
'지능',
'겐',
'다음영상주제',
'영상',
'주제',
'뽕',
'마블',
'조회',
'조회수가',
'수가',
'구독',
'구독자량',
'자량',
'블랙',
'위도',
'개봉',
'위도우',
'우',
'줄겟',
'줄겟닷',
'닷',
'17',
'17대표적',
'오류',
'이때',
'탱크',
'대표적',
'이때아이언',
'기관총',
'화살',
'화살정리',
'윤',
'주세',
'메일목',
'복장정리',
'메일목도리',
'도리',
'총',
'관련',
'도리님',
'총정리',
'용',
'사실',

'사실마크85',
'85',
'틀',
'수',
'스톤',
'우주',
'최고',
'벨리',
'재현',
'엔드',
'틀렛',
'내구력',
'기술력',
'엔드게임',
'렛',
'워프',
'하이라이트',
'게임',
'힘',
'생각',
'17형',
'그거',
'전차포',
'내가보기',
'보기',
'115',
'정도',
'20',
'흙',
'피터',
'자신',
'그램',
'작동',
'처음',
'공백',
'때흙',
'비행기',
'처음부분',
'장치',
'부분',
'이것',
'파',
'파프롬',
'프롬',
'대량',
'뒷',
'저녁',
'돌리',
'누구',
'포함',
'시간여행',
'시뮬레이션',
'업',
'스티브',
'스티브가토',
'여행',
'거절',
'그날',
'가토',
'방법',

'모색',
'저스',
'저스3',
'경우',
'수',
'가슴',
'시작',
'기자',
'기자회견',
'회견',
'엠',
'엠아이언',
'맨아이언맨',
'끝',
'핑거',
'핑거스냅하면서',
'스냅',
'하면서',
'묵도',
'묵도리',
'리',
'지금',
'마지막',
'에',
'스텐',
'스텐리',
'의',
'부담감',
'부담감주기',
'주기',
'췌',
'자체',
'22',
'19',
'19명령',
'명',
'령',
'안그레도',
'그레도',
'나',
'번',
'번부탁',
'부탁',
'농부',
'노스',
'즈아이미',
'미',
'정',
'그것',
'장난',
'렉',
'렉스부',
'스부',
'로켓',
'피해',
'어벤',
'스가',
'측정',
'빛',
'노스머리',

'머리',
'작',
'ㄱ',
'다음슈',
'슈',
'스탠리ㄱ',
'마블영화',
'분석',
'히등장',
'스파이더맨',
'뒤',
'더',
'닝',
'어우',
'어우전',
'전',
'25',
'25스트라이커',
'스트라이커',
'제타',
'제타노스피부색이랑',
'피부색',
'이랑',
'노스옷총정리',
'스옷',
'카',
'유물',
'카마르타지',
'마르',
'타지',
'16',
'44',
'비브라',
'비브라늄',
'늄',
'아이언맨안경정리',
'안경',
'가인',
'가인크레더블혈크',
'크레',
'더블',
'건',
'형님',
'피터파',
'친구',
'시리즈',
'모든',
'건어',
'건어때',
'다나',
'넘',
'링',
'넘비',
'비',
'넘버링',
'ㅈ',
'ㅈㅈㅈㅈ',
'ㅈ',
'복잡',
'감사',

'감사감사',
'0',
'발음',
'기',
'욕',
'연도',
'연도제압',
'제압',
'스가디',
'디',
'인류',
'베',
'베티로스',
'로스',
'26',
'26스팸',
'스팸',
'출',
'5252',
'5252목도',
'공',
'공계가안',
'계',
'가안',
'저',
'추측',
'마크85',
'테스트',
'테스트모델',
'모델',
'시발점',
'아',
'ㅋ',
'02',
'02오',
'오',
'39스타',
'스타',
'부스트',
'맨역',
'로',
'마블망',
'망',
'아웃',
'라이더',
'외형',
'외형총정리',
'잔',
'생',
'참고',
'생기진',
'기진',
'종료',
'있음',
'뭐',
'구미',
'론',
'43',
'10',
'29',

'29싱크로',
'싱크로',
'전부',
'엑스트라',
'22분',
'분',
'영',
'영성만',
'성만',
'수고',
'근',
'걸',
'명작',
'오늘',
'걸형',
'근데이',
'마블정주행',
'데이',
'정주',
'행',
'고',
'고ㅏㅓ',
'ㅏㅓ',
'20분',
'분',
'닛',
'메밀',
'팔',
'팔콘',
'콘',
'원',
'솔',
'페',
'폐기',
'카터',
'리',
'리키',
'키',
'너',
'크으',
'존재',
'뵙',
'난',
'난293개언어',
'293',
'개',
'언어',
'ㅅㅅ',
'ㅅㅅㅏ',
'뷔',
'뷔쥘',
'쥘',
'하드웨어',
'거지',
'찰',
'찰뷔스',
'뷔스',
'소프트',
'소프트웨어',
'웨어',

'구',
'06',
'06오후',
'오',
'호',
'흑',
'개꿀',
'미니',
'미니아크리액터',
'아크리',
'액',
'6개',
'동굴',
'말리',
'말리부',
'아이언맨2초반',
'초반',
'후반',
'아이언맨3',
'저',
'저에너지',
'에너지',
'가타',
'때',
'쉽',
'오른손잡이',
'이',
'이당',
'자',
'이제',
'이제드렉스',
'드렉스',
'티스',
'토르복장정리',
'야쵸',
'자동차',
'스탠스',
'리리리리',
'넘버',
'저스1',
'때34개',
'34',
'양산',
'생산',
'양산형',
'마찬가지',
'동일',
'이건',
'무도',
'무도리',
'영어',
'영어발음',
'02초',
'초',
'즈오',
'마크세',
'세',
'스탠리복장가즈',
'복',
'장가',

'하워드',
'조수',
'자비',
'자비스토니',
'스토니',
'해피',
'모건',
'영상인덱스',
'인덱스',
'개약',
'약',
'오왓',
'왓',
'묵',
'묵돌이',
'돌이',
'담',
'주문',
'어스',
'레이',
'엑스',
'어스파',
'레이미',
'맨유니',
'유니',
'버스',
'이정',
'이정도면',
'도면',
'형인',
'때아웃',
'점',
'저거',
'머리카락',
'편집',
'편집실력배리굿',
'실력',
'배리',
'굿',
'빵',
'짐',
'군항',
'단하',
'겁',
'겁나커',
'나커',
'무엇',
'데폭발',
'겁나크',
'전미사일',
'폭발',
'나크',
'미사일',
'겁나작',
'나작',
'팔라똥',
'스중에',
'포',
'포츠',
'츠',

'가능',
'튜브',
'취소',
'유',
'반복',
'계획',
'묵살',
'조회수8000',
'회수',
'8000',
'600',
'600대',
'대',
'감동',
'동안',
'동안감사',
'죄',
'죄송',
'송',
'2번',
'농담',
'우리',
'로드',
'로드행',
'우리스타',
'스냅',
'구독자',
'구독자반',
'반',
'3000천만큼안사랑',
'천',
'만큼',
'안사랑',
'영',
'재',
'재밌으',
'밌으',
'3000만큼힘',
'갯당',
'흑우',
'쥬',
'비링',
'해피복장',
'노스복장정리',
'21',
'18',
'라',
'아무',
'라나노',
'틀렛이아',
'렛이아',
'얘기',
'얘기안',
'자연',
'애',
'저번',
'대본',
'이래',
'슈트정리',
'저번워머신',

'워',
'머신',
'완전',
'완전사탄인정',
'사탄',
'인정',
'폐기카',
'캡틴',
'캡틴여친',
'친',
'행복',
'건강',
'슈트더많',
'더많',
'저스총정리',
'17이때',
'법칙',
'물리학',
'무시',
'전설',
'슈트전설',
'ㅏ',
'ㅏㅇ',
'명령',
'알람',
'컷',
'우리',
'노스형님가야조',
'가야',
'조',
'7.83',
'분3000만큼사랑',
'17대공포',
'대공포',
'탱크주포',
'주포',
'거의',
'캣',
'캣단',
'단',
'ㅋㅈㄱ',
'ㄱ',
'배',
'조종',
'마크1아크리액터',
'아크',
'리액터',
'가동력',
'가동력원인',
'원인',
'내용',
'컨텐츠',
'12',
'12부추요',
'추요',
'이건ㄹㅈㄷ',
'ㄹ',
'ㄷ',
'뒤통수',
'워머신',

'인정임ㅅㄱ',
'0.3000',
'0.3000만큼사랑',
'ㄹㅇ아이언',
'손실',
'구',
'의상',
'10해주시',
'해',
'주시',
'처음시',
'발',
'필',
'필발',
'와우',
'총정리한번',
'한번',
'명령수행',
'수행',
'7.77',
'구독잭팟',
'잭',
'팟',
'ㄹㅇ마크1',
'파이어',
'파이어뱃임',
'뱃',
'50번',
'번',
'힐',
'왕',
'츠감사함',
'합',
'당',
'메밀형고마웁',
'고마',
'웁',
'응',
'마크51',
'51',
'84',
'3000년지',
'년',
'나도',
'미스',
'지',
'테리',
'유투',
'튜',
'혈버',
'혈버2.0',
'2.0',
'마크49',
'49',
'레',
'레스큐',
'스큐',
'정확',
'영상제작',
'갈게',

'갈게여',
'57',
'이이',
'|',
'|ㅇ|',
'메밀묵도리님',
'학원',
'2시간',
'규',
'규독',
'독',
'로다주여',
'배우',
'유튜버들',
'버들',
'눈물',
'이',
'한로',
'한로다주여',
'제일',
'존경',
'팬',
'날마블',
'이젠',
'오진',
'아빠',
'아빠인로다주',
'남편',
'레고아이언',
'고아',
'이언',
'맨역할',
'역할',
'그',
'매력',
'당신',
'거부',
'영원',
'니스',
'업글',
'설정',
'달래',
'독알',
'영상시작',
'알',
'요구',
'원점',
'형18만큼사랑',
'전마블영화광',
'영화광',
'삭제',
'삭제장면',
'이번영상',
'아이언맨1',
'사막',
'서로',
'서로디가토',
'훗날',
'썬',
'유튜버중한명',

'튜버',
'한명',
'고생',
'고',
'고르',
'우즈베키스',
'키스',
'로난',
'웰케슬플까',
'착용',
'완벽',
'형영상',
'있어근데퀴실버',
'11',
'면',
'본적',
'드래건',
'면라이더',
'ㄷㄷ',
'개인적',
'비나노',
'마크45',
'45',
'디자인',
'46',
'46아크테크슈트',
'테크',
'남자',
'50시발점',
'일',
'일회성',
'회성',
'스킬',
'캠',
'3000만 큼',
'하심',
'님',
'변',
'치즈',
'변화해',
'치즈버거',
'화해',
'버거',
'결국',
'결국아이언맨',
'끓엇',
'분',
'닥터마법정리',
'마법',
'화려',
'에릭',
'몽',
'호',
'후',
'레게',
'헤이',
'양',
'ㅏ양',
'ㅓ',
'ㅓ우',

```
'투우우',
'예',
'슈트이름',
'해머',
'해머헤드',
'헤드',
'원편맨',
'편',
'현상',
'현상범',
'범',
'13000',
'만큼사랑',
'왈',
'48',
'48무게라',
'무게',
'기보',
'헬기',
'기보단',
'마블영화페이지즈',
'이즈',
'지형',
'52',
'자살',
'직원',
'창조주',
'메밀묵도리님고생',
'조사',
'노스웁고',
'스웁',
'팬서',
'서',
'페퍼',
'거십간지남',
'십간',
'지남',
'영어공부',
'공부',
'건팩트',
...]
```

```
In [12]: # nltk.Text()를 이용하여 nltk가 가지는 많은 기능을 사용 가능해짐.
ko = nltk.Text(doc_nouns, name="아이언맨")
print(type(ko))
print(len(ko.tokens))
```

```
<class 'nltk.text.Text'>
2310
```

```
In [13]: ### 단어들의 사용 횟수 확인 - 빈도 분석
ko.vocab()
```

```
Out[13]: FreqDist({'리': 3, '분': 3, '대': 3, '만큼': 2, '여': 2, '시': 2, '건': 2, '하
나': 2, '일': 2, '데': 2, ...})
```

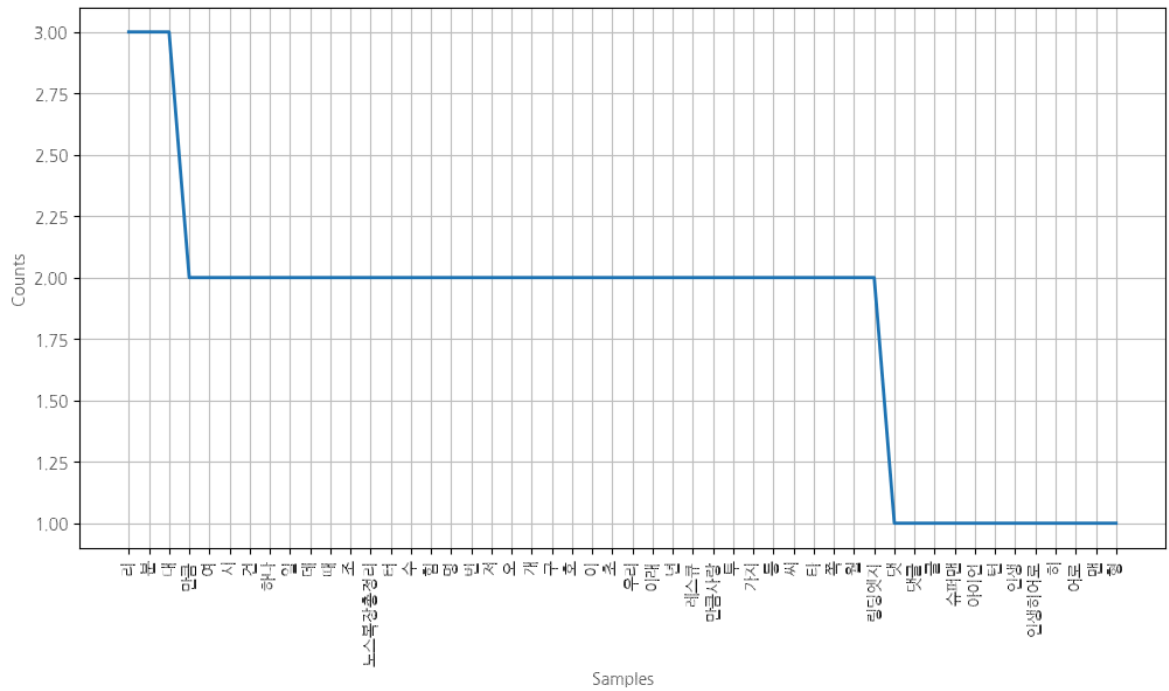
```
In [14]: most_fre = ko.vocab().most_common(50)
most_fre
```

```
Out[14]: [('리', 3),
          ('분', 3),
          ('대', 3),
          ('만큼', 2),
          ('여', 2),
          ('시', 2),
          ('건', 2),
          ('하나', 2),
          ('일', 2),
          ('데', 2),
          ('때', 2),
          ('조', 2),
          ('노스북장총정리', 2),
          ('터', 2),
          ('수', 2),
          ('힘', 2),
          ('명', 2),
          ('번', 2),
          ('자', 2),
          ('오', 2),
          ('개', 2),
          ('구', 2),
          ('호', 2),
          ('이', 2),
          ('초', 2),
          ('우리', 2),
          ('이래', 2),
          ('년', 2),
          ('레스큐', 2),
          ('만큼사랑', 2),
          ('투', 2),
          ('가지', 2),
          ('등', 2),
          ('씨', 2),
          ('타', 2),
          ('쪽', 2),
          ('월', 2),
          ('링딩엣지', 2),
          ('땃', 1),
          ('땃글', 1),
          ('글', 1),
          ('슈퍼맨', 1),
          ('아이언', 1),
          ('턴', 1),
          ('인생', 1),
          ('인생히어로', 1),
          ('히', 1),
          ('어로', 1),
          ('맨', 1),
          ('형', 1)]
```

```
In [15]: ### 중복된 단어를 제거한 개수를 확인
         print(len(set(ko.tokens)))
```

2269

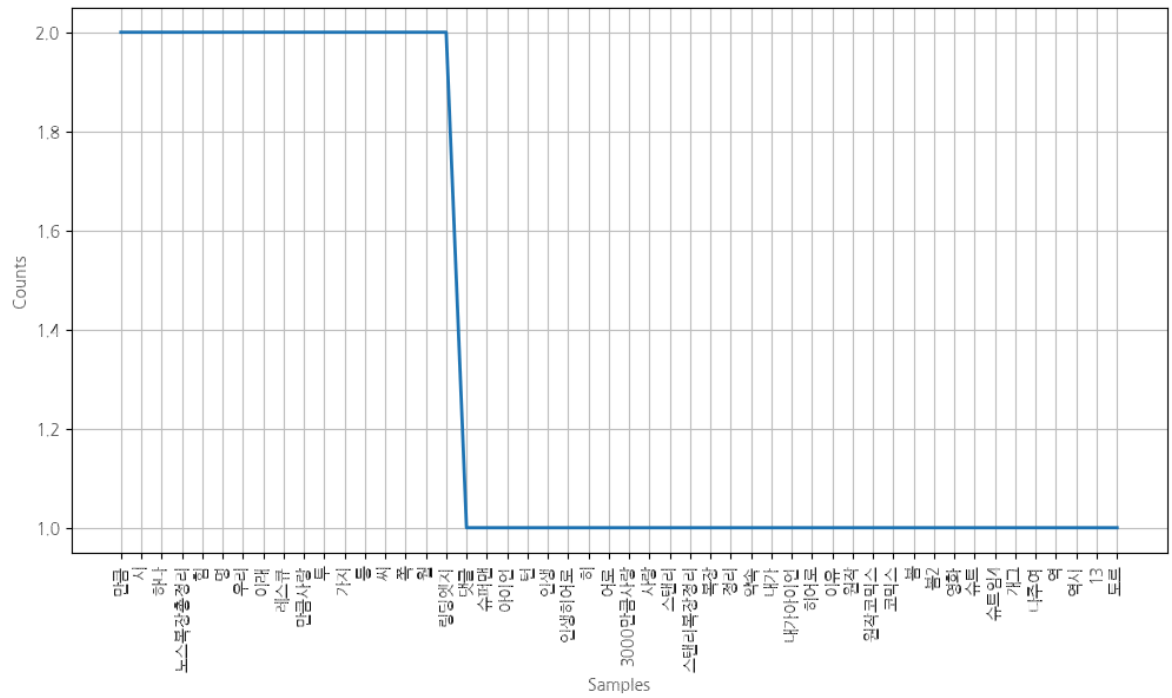
```
In [16]: plt.figure(figsize=(12, 6))
         ko.plot(50)
         plt.show()
```



```
In [17]: ### 한글에서는 따로 불용어 사전이 없어, 따로 만들거나 또는 파일로 부터 불러올 수
stop_words = ['리', '일', '이', '분', '내',
              '여', '건', '나', '개', '번',
              '때', '오', '수', '구', '터',
              '초', '저', '조', '대', '년', '데', '호', '타', '차', '외', '댓',
              '임', '함', '원', '날', '글', '중', '부', '함5', '토', '뒷',
              '거', '음', '형', '맨', '3000', '메', '도', '군', '님', '1', '2', '3',
              '6', '7', '8', '9', '다가', '다주']

new_ko = [ ]
for one_word in ko:
    if one_word not in stop_words:
        new_ko.append(one_word)
```

```
In [18]: ### nltk Text 객체 만들기
new_ko = nltk.Text(new_ko, name="아이언맨2")
plt.figure(figsize=(12,6))
new_ko.plot(50)
```



```
Out[18]: <Axes: xlabel='Samples', ylabel='Counts'>
```

```
In [19]: most_fre = new_ko.vocab().most_common(50)
          most_fre
```



```
Out[19]: [('만큼', 2),
          ('시', 2),
          ('하나', 2),
          ('노스복장총정리', 2),
          ('힘', 2),
          ('명', 2),
          ('우리', 2),
          ('이래', 2),
          ('레스큐', 2),
          ('만큼사랑', 2),
          ('투', 2),
          ('가지', 2),
          ('등', 2),
          ('씨', 2),
          ('쪽', 2),
          ('월', 2),
          ('링딩엣지', 2),
          ('뎃글', 1),
          ('슈퍼맨', 1),
          ('아이언', 1),
          ('턴', 1),
          ('인생', 1),
          ('인생히어로', 1),
          ('히', 1),
          ('어로', 1),
          ('3000만큼사랑', 1),
          ('사랑', 1),
          ('스탠리', 1),
          ('스탠리복장정리', 1),
          ('복장', 1),
          ('정리', 1),
          ('약속', 1),
          ('내가', 1),
          ('내가아이언', 1),
          ('히어로', 1),
          ('이유', 1),
          ('원작', 1),
          ('원작코믹스', 1),
          ('코믹스', 1),
          ('봄', 1),
          ('봄2', 1),
          ('영화', 1),
          ('슈트', 1),
          ('슈트임4', 1),
          ('개그', 1),
          ('다주여', 1),
          ('역', 1),
          ('역시', 1),
          ('13', 1),
          ('토르', 1)]
```

텍스트의 단어어 분포 확인 (dispersion_plot)

```
In [20]: from wordcloud import WordCloud, STOPWORDS

import numpy as np
from PIL import Image
```

```
In [21]: data = new_ko.vocab().most_common(1000)
```

03 텍스트 데이터 시각화-워드 클라우드

목차로 이동하기

```
In [22]: ### 워드 클라우드 표현을 위한 데이터 생성
### 약간의 시간이 필요.
wc = WordCloud(background_color='white',
                max_words=200,
                # mask=Car_mask,
                contour_width=3,
                contour_color="steelblue",
                font_path=path).generate_from_frequencies(dict(data))
```

```
In [23]: plt.figure(figsize=(12,8))
plt.imshow(wc)
plt.axis("off")
plt.show()
```

