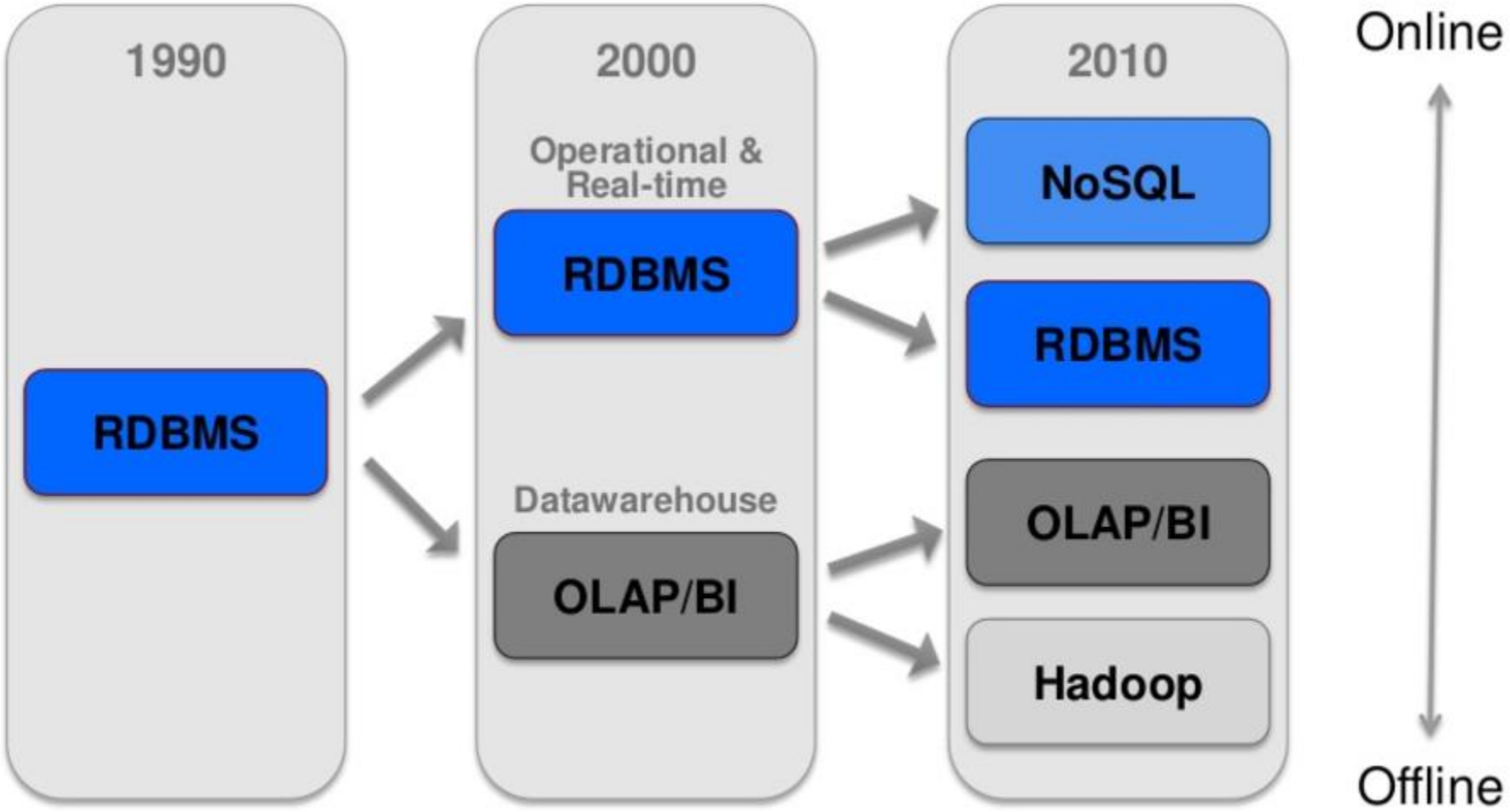
2018 [Database System] 실습 3

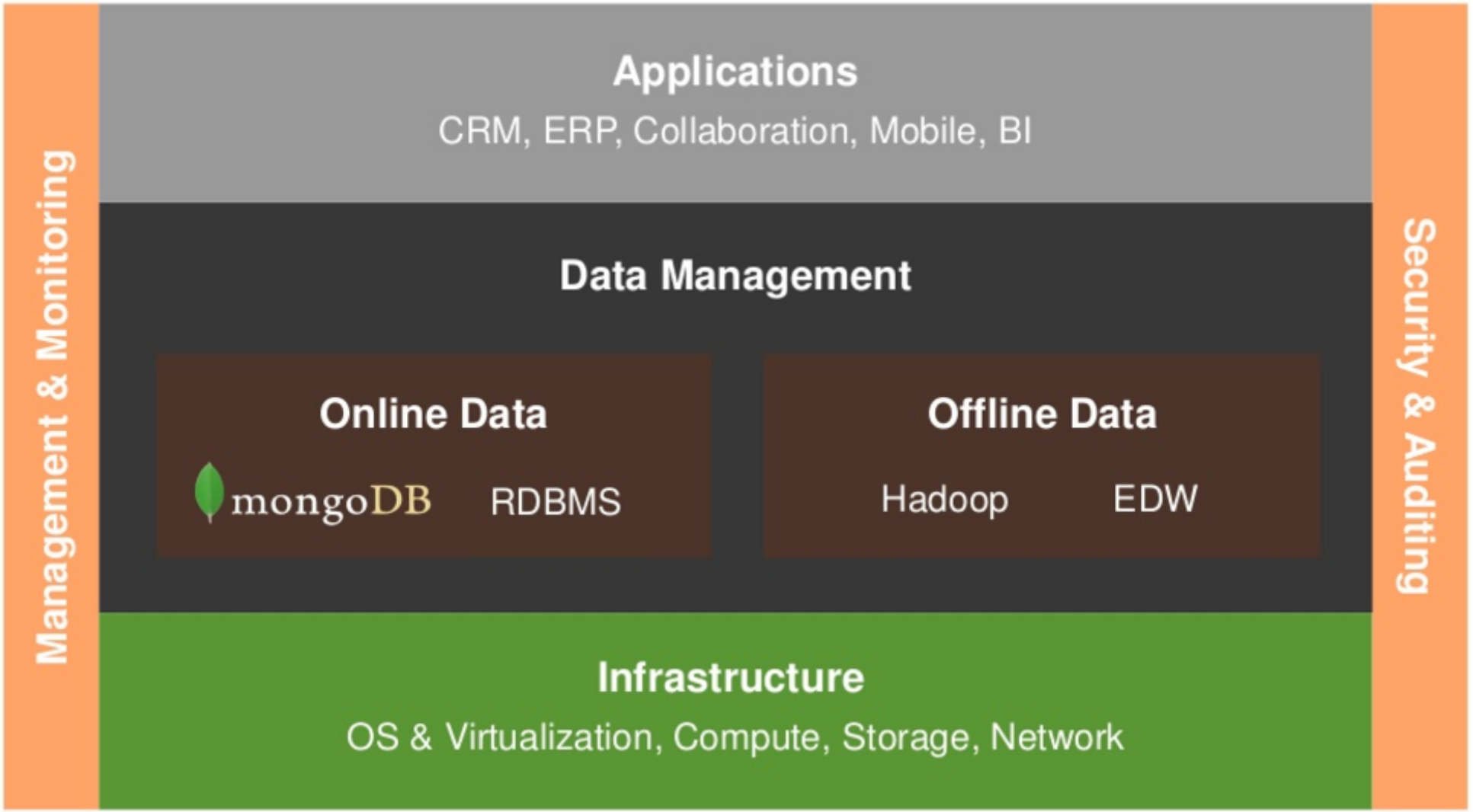
MongoDB 접속 및 형태소 분석

Sogang Univ. Database Lab.

The Evolution of Databases



MongoDB and Enterprise ID Stack



MongoDB vs. SQL

MongoDB	SQL
Database	Database
Collection	Table/View
Document	Record/Tuple
PK: _id Field	PK: Any Attribute(s)
Dynamic Schema	Fixed Schema
Index	Index
Embedded Document	Joins
Shard	Partition

Document Data Model

Relational

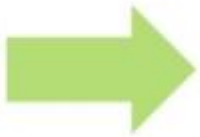
Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

— no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2



MongoDB

```
{
  first_name: 'Paul',
  surname: 'Miller'
  city: 'London',
  location: [45.123,47.232],
  cars: [
    { model: 'Bentley',
      year: 1973,
      value: 100000, … },
    { model: 'Rolls Royce',
      year: 1965,
      value: 330000, … }
  ]
}
```

실습 서버 접속

- **서버 정보**
  - Host address: dbpurple.sogang.ac.kr
  - Port: 22

- **계정 정보** *자신의 학번으로 입력하세요.
  - ID: DB학번 (e.g. DB20181234)
  - 초기 PW: DB학번 (e.g. DB20181234)

MongoDB 접속

- DB 접속
  - 서버에 원격 터미널 접속 후 다음을 입력
    - mongo -u “db학번” -p “db학번” --authenticationDatabase “db학번”
    - “”를 포함할 것.
- Database 변경
  - use db학번

```
sehwa@dbpurple:~/DBProject$
sehwa@dbpurple:~/DBProject$ mongo -u "dbta" -p "dbta" --authenticationDatabase "dbta"
MongoDB shell version: 3.0.14
connecting to: test
> use dbta
switched to db dbta
>
```

- Collection 목록 보기
  - show collections

```
> show collections
news
system.indexes
>
```

MongoDB 접속

- Collection 내용 보기
  - db.news.find()
  - db.news.findOne()

```
> db.news.findOne()
{
        "_id" : ObjectId("59211fe36c9d342a2080c08e"),
        "title" : "박 대통령, 사드 '목청'…우병우엔 '침묵'",
        "url" : "http://news.naver.com/main/read.nhn?mode=LS2D&
4&oid=032&aid=0002718576",
        "datetime" : ISODate("2016-08-02T22:56:00Z"),
        "content" : "· 국무회의 주재…"저의 남은 소명은 나라 지키
혜 대통령은 2일 "북한이 핵능력을 고도화시키면서 핵 탄도미사일 성
는 상황인데도 고고도미사일방어체계(THAAD·사드) 배치를 둘러싼 감
이 타들어가는 심정"이라고 말했다. 하지만 도덕성 논란으로 사퇴 요
수석 문제에 대해선 침묵했다.박 대통령은 이날 국무회의에서 "사드
일 위협으로부터 우리 국민을 보호하기 위해서 오랜 고심과 철저한
이라며 이같이 말했다. 박 대통령은 또 "저도 가슴 시릴 만큼 아프게
에게 남은 유일한 소명은 대통령으로서 나라와 국민을 각종 위협으로
것"이라고 감성적으로 호소했다.그러면서 "사드 배치 문제를 비롯한
을 청취하고 문제를 적극적으로 해결해나가기 위해 지역 대표인 국회
만날 것"이라고 밝혔다. 경북 성주 군민 등의 철회 요구를 거부하면
하겠다는 뜻을 드러낸 것이다. 박 대통령은 이르면 4일 대구·경북 의
할 것으로 보인다.그러나 박 대통령은 우 수석 문제는 일절 거론하지
모들과 국무위원들에게 "힘들게 하루하루를 보내고 있는 우리 국민들
전 노력해주기를 바란다"고 했다. 우 수석을 안고 가겠다는 뜻을 내
NS [트위터] [페이스북]▶ [인기 무료만화 보기]©경향신문 (www.khan.
포 금지",
        "press" : "경향신문"
}
>
```

```
> db.news.find()
{ "_id" : ObjectId("59211fe36c9d342a2080c08e"), "title" : "
우엔 '침묵'", "url" : "http://news.naver.com/main/read.nhn?
id2=264&oid=032&aid=0002718576", "datetime" : ISODate("2016
" : "· 국무회의 주재…"저의 남은 소명은 나라 지키는 것" 감성
북한이 핵능력을 고도화시키면서 핵 탄도미사일 성능을 끊임없이
도미사일방어체계(THAAD·사드) 배치를 둘러싼 감등이 멈추지 않
"이라고 말했다. 하지만 도덕성 논란으로 사퇴 요구에 직면한
침묵했다.박 대통령은 이날 국무회의에서 "사드 배치는 북한의
리 국민을 보호하기 위해서 오랜 고심과 철저한 검토를 거쳐서
다. 박 대통령은 또 "저도 가슴 시릴 만큼 아프게 부모님을 잃
소명은 대통령으로서 나라와 국민을 각종 위협으로부터 안전하
으로 호소했다.그러면서 "사드 배치 문제를 비롯한 지역 현안들
을 적극적으로 해결해나가기 위해 지역 대표인 국회의원과 단체
밝혔다. 경북 성주 군민 등의 철회 요구를 거부하면서 직접 나
드러낸 것이다. 박 대통령은 이르면 4일 대구·경북 의원들을 만
다.그러나 박 대통령은 우 수석 문제는 일절 거론하지 않았다.
들에게 "힘들게 하루하루를 보내고 있는 우리 국민들을 생각해
바란다"고 했다. 우 수석을 안고 가겠다는 뜻을 내비친 것이다
이스북]▶ [인기 무료만화 보기]©경향신문 (www.khan.co.kr), 무
s" : "경향신문" }
{ "_id" : ObjectId("59211fe36c9d342a2080c096"), "title" : "
에 최선 다할 것\"", "url" : "http://news.naver.com/main/rea
=100&sid2=264&oid=003&aid=0007385669", "datetime" : ISODate
ontent" : "【서울=뉴시스】전진환 기자 = 박근혜 대통령이 2일
국무회의에 입장하고 있다. 2016.08.02 amin2@newsis.com【서
근혜 대통령은 2일 \"우리 정부는 입양인 여러분을 위한 지원에
다. 박 대통령은 이날 오후 서울 소공동 롯데호텔에서 열린 '20
낸 영상메시지에서 \"국외 입양인 여러분께 모국방문과 모국어
기회를 제공하고, 모국에 돌아와 생활하는 입양인 지원도 충실
밝혔다. 박 대통령은 입양인들에게 \"여러분은 어린 나이에 모
지만 입양가정의 따뜻한 보살핌과 자신의 노력으로 밝고 희망찬
여러분의 모국 대한민국 역시 과거 세계에서 가장 가난한 나라
```

MongoDB 접속

- Collection index 구성
  - datetime(시간) 순으로 인덱스 구성
  - db.news.ensureIndex({“datetime”:1})

```
> db.news.ensureIndex({"datetime":1})
{
        "createdCollectionAutomatically" : false,
        "numIndexesBefore" : 1,
        "numIndexesAfter" : 2,
        "ok" : 1
}
>
>
> db.system.indexes.find()
{ "v" : 1, "key" : { "_id" : 1 }, "name" : "_id_", "ns" : "dbta.news" }
{ "v" : 1, "key" : { "datetime" : 1 }, "name" : "datetime_1", "ns" : "dbta.news" }
```

- 그 외 명령어는 아래 링크를 참조
  - https://docs.mongodb.com/manual/tutorial/query-documents/

Pymongo

- Python 코드를 통한 mongoDB 접속
  - 파일 생성
    - shell에서 vi editor를 통해 python code 파일 생성
    - vi DBex#3_학번.py
  - 라이브러리 import

```
#-*- coding: utf-8 -*-

import datetime
import time
import sys
import MeCab
import operator
from pymongo import MongoClient
from bson import ObjectId
from itertools import combinations
```

  - 메뉴 함수 정의
    - 본 실습에서는 0,1번과 관련된 함수 구현을 목적으로 함.

```
def printMenu():
    print "0. CopyData"
    print "1. Morph"
    print "2. print morphs"
    print "3. print wordset"
    print "4. frequent item set"
    print "5. association rule"
```

```
def p0():
    """
    TODO:
    CopyData news to news_freq
    """
def p1():
    """
    TODO:
    Morph news and update news db
    """
def p2():
    """
    TODO:
    input  : news url
    output : news morphs
    """
def p3():
    """
    TODO:
    copy news morph to  new db named news_wordset
    """
def p4():
    """
    TODO:
    input  : news url
    output : news wordset
    """
def p5(length):
    """
    TODO:
    make frequent item_set
    and inset new dbs (dbname = candidate_L+"length")
    ex) 1-th frequent item set dbbname = candidate_L1
    """
def p6(length):
    """
    TODO:
    make strong association rule
    and print all of strong rules
    by length-th frequent item set
    """
```

Pymongo

- Python 코드를 통한 mongoDB 접속
  - Data Copy 함수 정의

```
def p0():
    col1 = db['news']
    col2 = db['news_freq']

    col2.drop()

    for doc in col1.find():
        contentDic = {}
        for key in doc.keys():
            if key != "_id":
                contentDic[key] = doc[key]
        col2.insert(contentDic)
```

  - MongoDB 접속
    - dbta1 대신에 db학번

```
stop_word = {}
DBname = "dbta2"
conn = MongoClient('dbpurple.sogang.ac.kr')
db = conn[DBname]
db.authenticate(DBname,DBname)
```

```
if __name__ == "__main__":
    make_stop_word()
    printMenu()
    selector = input()
    if selector == 0:
        p0()
    elif selector == 1:
        p1()
        p3()
    elif selector == 2:
        url = str(raw_input("input news url:"))
        p2(url)
    elif selector == 3:
        url = str(raw_input("input news url:"))
        p4(url)
    elif selector == 4:
        length = int(raw_input("input length of the frequent item:"))
        p5(length)
    elif selector == 5:
        length = int(raw_input("input length of the frequent item:"))
        p6(length)
```

Pymongo

- Python 코드를 통한 mongoDB 접속
  - Data Copy 실행
    - Shell에서 다음 명령어를 수행
    - chmod 750 DBex#3_학번.py
    - Python DBex#3_학번.py

```
python DBprj#3_dbta.py
0. CopyData
1. Morph
2. print morphs
3. print wordset
4. frequent item set
5. association rule
0
```

  - Data Copy 여부 확인
    - mongodb 접속 후 collection 확인

```
> show collections
news
news_freq
system.indexes
>
```

```
> db.news_freq.findOne()
{
        "_id" : ObjectId("5b067d8e6c9d340fe739dalc"),
        "title" : "남양주경찰서, 진접파출소에 24시간 개방 도서관 마련",
        "url" : "http://news.naver.com/main/read.nhn?mode=LS2D&mid=shm&sidl=103&sid2=
        "datetime" : ISODate("2016-06-17T10:52:00Z"),
        "content" : "(남양주=연합뉴스) 권숙희 기자 = 경기 남양주경찰서는 진접파출소
다고 17일 밝혔다. 어린이용 도서 200권, 일반 도서 400권 등 장서 600권을 보유한 도서
지난 16일 개관했다. 365일 24시간 문을 여는 파출소 업무 특성에 따라 도서관도 연중
열람대장에 간단한 인적사항을 기재한 뒤 빌려볼 수 있다. 경찰 관계자는 \"주민들이 책
공유하고 불안 요소를 쉽게 건의하는 기회를 제공할 것으로 기대된다\"고 밝혔다. suki
놀리비아 \"우리를 거지로 본다\"▶ [핫클릭] 가장 오래된 산소찾아…\"131억광년전 초기은하
지갑 '슬쩍'<저작권자 (c) 연합뉴스, 무단 전재-재배포 금지>",
        "press" : "연합뉴스"
}
>
```

Pymongo

- Python 코드를 통한 mongoDB 접속
  - MorpAnalysis 함수 구성

```
## Make stop word
def make_stop_word():
    f = open("wordList.txt",'r')
    while True:
        line = f.readline()
        if not line: break
        stop_word[line.strip('\n')] = line.strip('\n')
    f.close()
```

형태소 분석 및 불용어 제거

```
def morphing(content):
    t = MeCab.Tagger('-d/usr/local/lib/mecab/dic/mecab-ko-dic')
    nodes = t.parseToNode(content.encode('utf-8'))
    MorpList = []
    while nodes:
        if nodes.feature[0] == 'N' and nodes.feature[1] == 'N':
            w = nodes.surface
            if not w in stop_word:
                try:
                    w = w.encode('utf-8')
                    MorpList.append(w)
                except:
                    pass
        nodes = nodes.next
    return MorpList
```

형태소 분석기 불러오기

```
def p0():
    col1 = db['news']
    col2 = db['news_freq']

    col2.drop()

    for doc in col1.find():
        contentDic = {}
        for key in doc.keys():
            if key != "_id":
                contentDic[key] = doc[key]
        col2.insert(contentDic)
```

형태소 분석 및 불용어 제거

MongoDB에 저장

Pymongo

- Python 코드를 통한 mongoDB 접속
  - Make new dbs

```
def p3():
    coll = db['news_freq']
    col2 = db['news_wordset']
    col2.drop()
    for doc in coll.find():
        new_doc = {}
        new_set = set()
        for w in doc['morph']:
            new_set.add(w.encode('utf-8'))
        new_doc['word_set'] = list(new_set)
        new_doc['url'] = doc['url']
        col2.insert(new_doc)
```

news_wordset db생성

Pymongo

- Python 코드를 통한 mongoDB 접속
  - MorpAnalysis 수행
    - Python DBex#3_학번.py

```
june244@dbpurple:~$ python DBprj#3_dbta.py
0. CopyData
1. Morph
2. print morphs
3. print wordset
4. frequent item set
5. association rule
1
```

  - MorpAnalysis 확인
    - mongodb 접속 후 collection 확인

```
> db.news_wordset.findOne()
{
        "_id" : ObjectId("5b065ca66c9d3466a8489bd0"),
        "url" : "http://news.naver.com/main/read.nhn?mode=LS2D&mid=shm&sid1=102&sid2=256&oid=421&aid=0002099866",
        "word_set" : [
                "김 부 겸 ",
                "민 주 당 ",
                "오 후 ",
                "대 학 교 ",
                "전 주 시 ",
                "대 학 ",
                "문 요 한 ",
                "전 주 ",
                "본 부 ",
                "대 구 ",
                "수 성 ",
                "최 신 ",
                "방 문 ",
                "갑 ",
                "국 회 의 원 ",
```

[참고] Python

- if 문

예시	수행 결과

A = 1
B = 2
if A == B:
  print "A == B"
else:
  print "A != B"
 	
if A >=1 and B>=1:
  print "True"
else:
  print "False"
	A !=B
	True

[참고] Python

- for 문
  - range()

예시	수행 결과
num = [] **for** i **in** range(1, 11): num.append(i) print num	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

  - List 이용

예시	수행 결과
templist =['one', 'two', 'three’] **for** i **in** templist: print i	one two three

  - Dictionary 이용

예시	수행 결과
dic = {’a’:1, ’b:2, ’c': 3} **for** k **in** dic: print k, dic[k]	a 1 b 2 c 3

[참고] Python

- List 자료형
  - list는 값의 시퀀스(sequence)
  - list에 들어 있는 값을 element 또는 item이라고 함
- List 자료형 예시
  - list 생성 및 element 추가하기

예시	수행 결과
characters = ['A', 'B', 'C'] numbers = [10, 15] empty = [] print characters, numbers, empty	['A', 'B', 'C'] [10, 15] []

예시	수행 결과
characters = [] characters.append('A') characters.append('B') characters.append('C') print characters	['A', 'B', 'C']

[참고] Python

- List의 element 삭제
  - pop()
    - List의 특정 인덱스의 element를 list에서 삭제하고 해당 element를 반환
    - 인덱스 없이 사용하면 마지막 element 삭제하고 해당 element를 반환

예시	수행 결과
numbers = [10, 15, 20, 25] print numbers numbers.pop(2) x = numbers.pop() print numbers print x	[10, 15, 20, 25] [15, 25] 10

  - remove()
    - 지우려는 element를 알고 있는 경우

예시	수행 결과
characters = ['A', 'B', 'C'] characters.remove('B') print characters	['A', 'C']

[참고] Python

- List의 element 변경

예시	수행 결과
numbers = [10, 15, 20, 25] print numbers numbers[2] = 7 print numbers	[10, 15, 20, 25] [10, 15, 7, 25]

- List slice

예시	수행 결과
numbers = [10, 15, 20, 25, 30] t = numbers[1:3] print t	[15, 20, 25]

[참고] Python

- Dictionary 자료형
  - Key와 Value의 쌍으로 표현하는 자료형

- Dictionary 쌍 추가

예시	수행 결과
dic1 = {'a':1, 'b':2, 'c': 3} print dic1 dic2 = {} dic2['d'] = 4 print dic2	{'a':1, 'b':2, 'c': 3} {'d': '4'}

- Dictionary 쌍 제거

예시	수행 결과
dic = {'a':1, 'b':2, 'c': 3} del dic['b'] print dic	{'a':1, 'c': 3}

[참고] Python

- Dictionary에서 key 로 value 얻기

예시	수행 결과
dic = {'a':1, 'b':2, 'c': 3} print dic['b']	2

- Key list 만들기
  - keys()

예시	수행 결과
dic = {'a':1, 'b':2, 'c': 3} print dic.keys()	['a', 'b', 'c']

- Value list 만들기
  - values()

예시	수행 결과
dic = {'a':1, 'b':2, 'c': 3} print dic.values()	[1, 2, 3]

[참고] Python

- set 자료형
  - 집합 자료형은 순서가 없고 중복을 허용하지 않는다.
- set 자료형 예시
  - set 생성 및 element 추가하기

예시	수행 결과
S1 = set([1,2,3]) S2 = set("Hello") print S1 print S2	Set([1,2,3]) Set(['H','e','l','o'])

  - https://wikidocs.net/1015

[참고] Python

- Frozenset
  - list나 set은 hashable 하지 않다.
  - ex) lis = [1,2,3]
    dic = {}
    dic[lis] = 1 -> error
    왜냐하면, lis가 가변하기 때문이다.
  - 따라서 frozenset을 통해 고정시켜야한다.
  - ex) lis = [1,2,3]
    dic = {}
    dic[frozenset(lis)] = 1 -> ok
- Permutation
  - https://www.geeksforgeeks.org/permutation-and-combination-in-python/

참고자료

- Pymongo API
  - https://api.mongodb.com/python/current/
- Python
  - https://wikidocs.net/book/1
  - https://docs.python.org/2/library/index.html
- MongodDB 명령어
  - https://docs.mongodb.com/manual/tutorial/query-documents/