빅데이터 처리 및 응용

지정주제

대표 포털 사이트의 실시간 검색어 비교

2014707073 김수환

목 차

2019 빅데이터 처리 및 응용

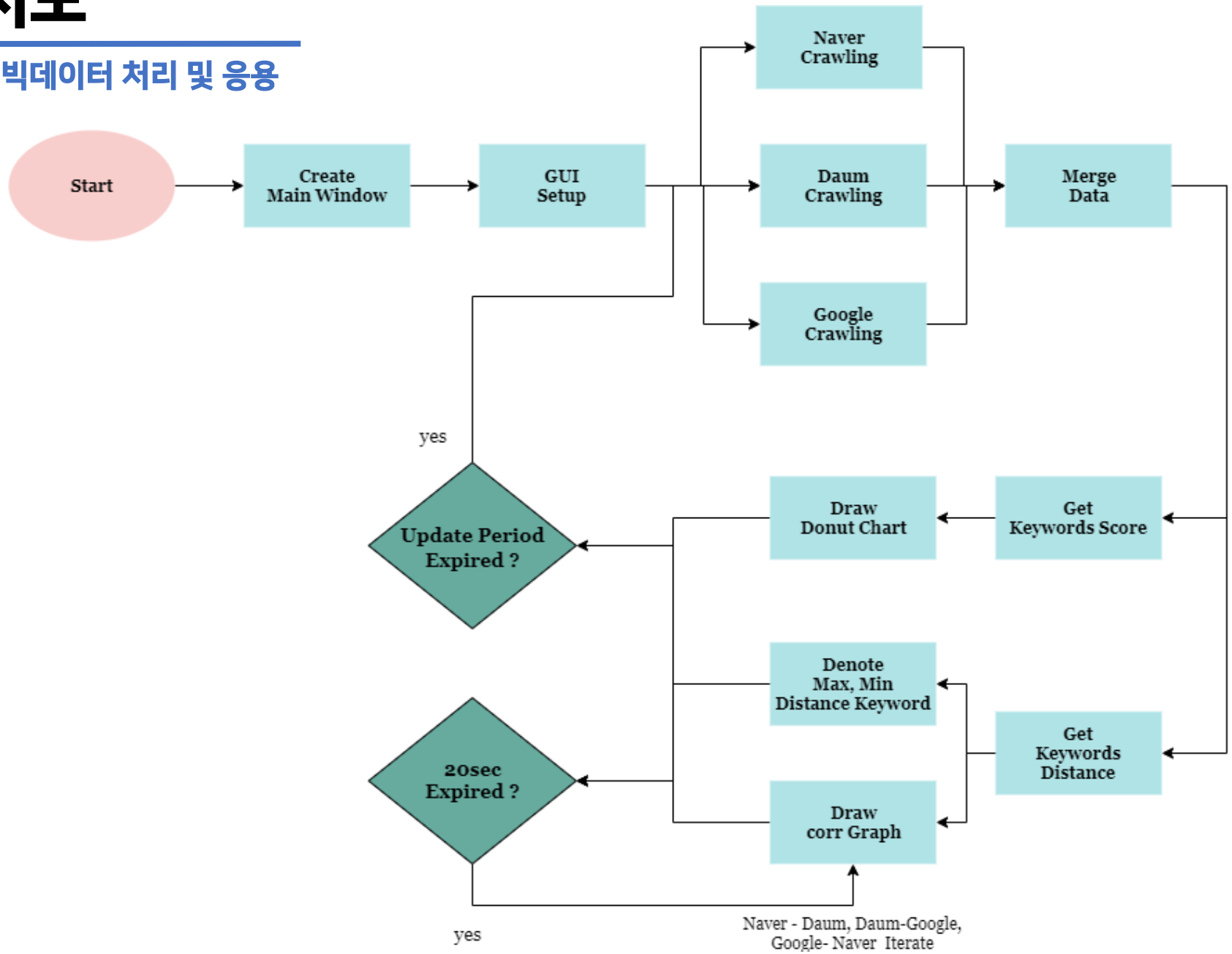
1 프로젝트 개요

2019 빅데이터 처리 및 응용

프로젝트명	**대표 포털 사이트의 실시간 검색어 비교**	
과목명	빅데이터 처리 및 응용	
개발기간	2019.11.07 ~ 2019.11.15	
개발 환경	Operating System	Windows 10 Edu.
	Language	Python 3.7.1
	Development Tools	PyCharm Community Edition 2019.2.4
	Library	- PyQt5 5.13.2 - NumPy 1.15.4 - Pandas 0.23.4 - Matplotlib 3.0.2 - bs4 0.0.1 - Selenium 3.141.0

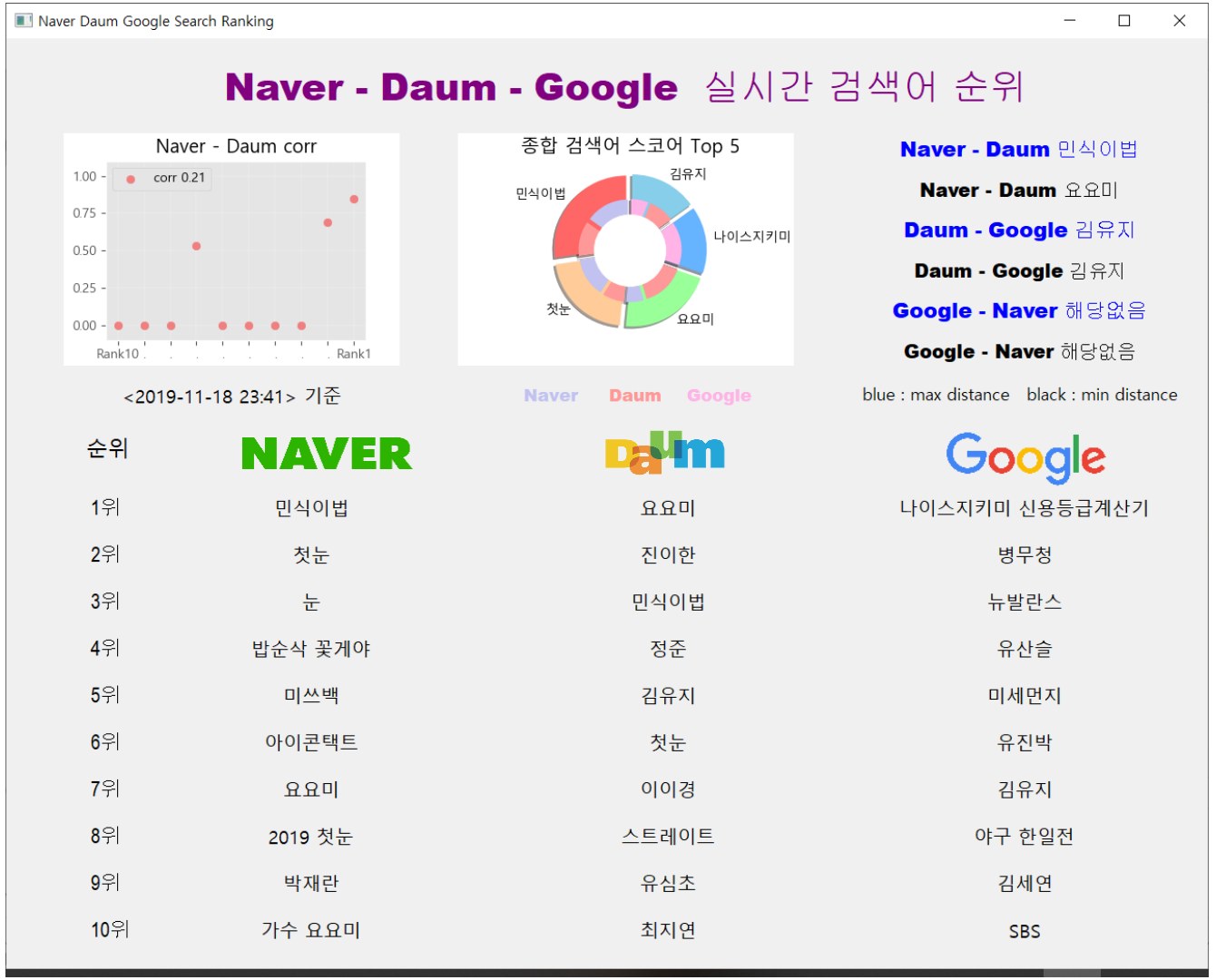
2 순서도

2019 빅데이터 처리 및 응용



3 구현 기능 – (1) GUI

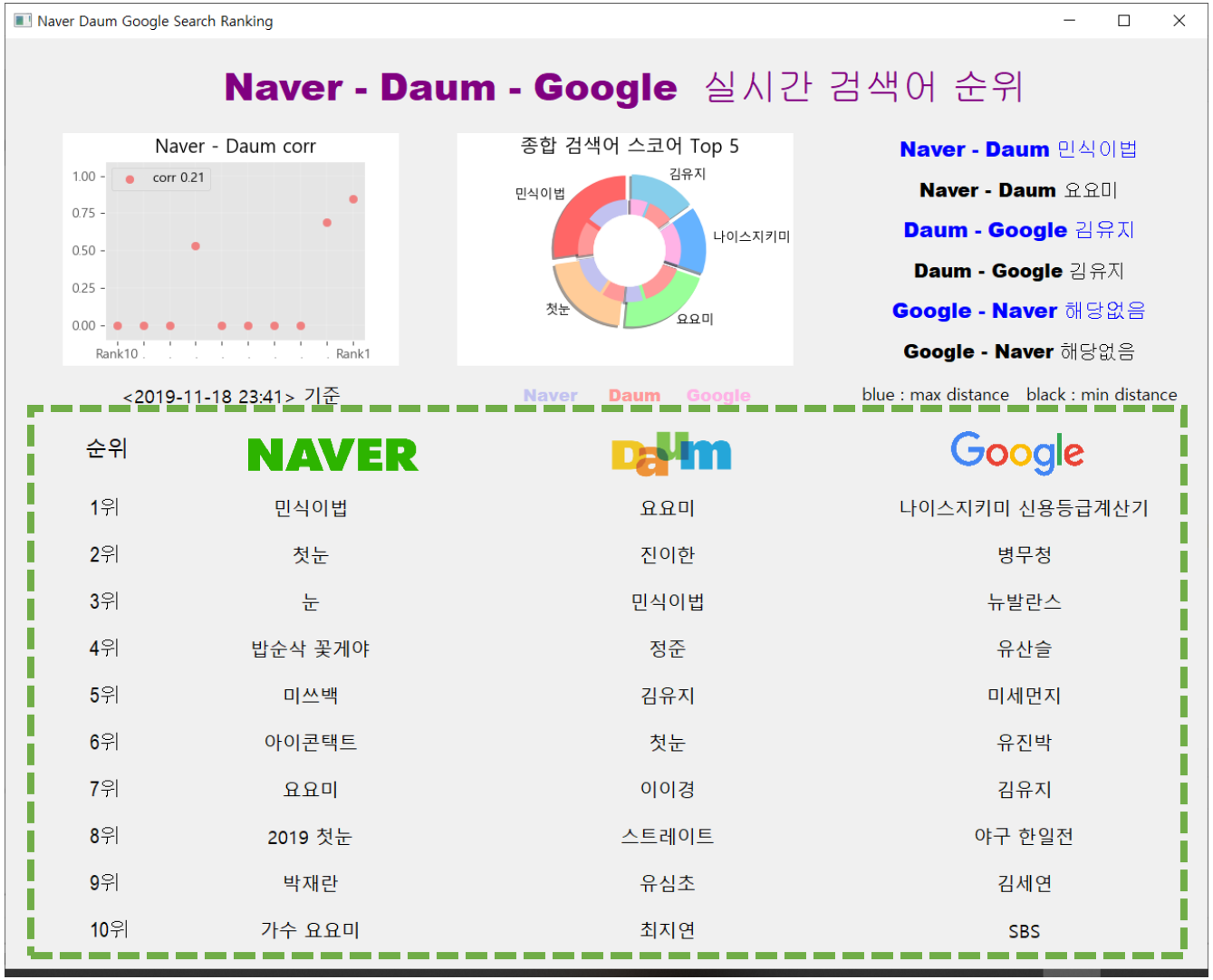
2019 빅데이터 처리 및 응용



대표 포털 사이트의 실시간 검색어 비교

3 구현 기능 – (2) 멀티스레딩 이용 대표 포털 검색어 크롤링

2019 빅데이터 처리 및 응용




```
class MultiCrawler():
    def __init__(self, rank_containers_list, queue):
        self.threads = list()  # Crawler thread를 저장할 리스트
        self.queue = queue     # thread 꼬임 방지용 큐
        search_engines = ['Naver', 'Daum', 'Google']  # 검색엔진 리스트

        for i, search_engine in enumerate(search_engines):
            self.threads.append(Crawler(search_engine,
                                        rank_containers_list[i],
                                        self.queue))

    # Thread Run
    def start(self):
        for thread in self.threads:
            thread.start()

    # Thread Join
    def join(self):
        for thread in self.threads:
            thread.join()
```

대표 포털 사이트의 실시간 검색어 비교

3 구현 기능 – (3) 크롤링 데이터 저장

2019 빅데이터 처리 및 응용

Time	Search Engine	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6	Rank7	Rank8	Rank9	Rank10
2019-11-18 17:20	Google	병무청	뉴발란스	임종석	나이스지키미 신용등급계산기	유산슬	야구 한일전	김세연	SBS	여우티 9900원	설리
	Naver	나이스지키미 신용등급계산기	정준	김유지	200억 발패치 힐링 패치	크리니크 프렌즈치크팝	커먼유니크5 억아울렛	나이스지키미	뉴발란스 블랙프라이데이	정준 김유지	나이스지키미 신용등급
	Daum	김유지	정준	유산슬	고유정	연애의 맛	이시아	한다감	1박 2일 시즌4	을지대학교	홍콩
2019-11-18 17:21	Google	병무청	뉴발란스	임종석	나이스지키미 신용등급계산기	유산슬	야구 한일전	김세연	SBS	여우티 9900원	설리
	Naver	나이스지키미 신용등급계산기	정준	김유지	200억 발패치 힐링 패치	크리니크 프렌즈치크팝	커먼유니크5 억아울렛	나이스지키미	정준 김유지	뉴발란스 블랙프라이데이	나이스지키미 신용등급
	Daum	김유지	정준	유산슬	고유정	연애의 맛	이시아	1박 2일 시즌4	을지대학교	한다감	홍콩

Column

[Time Search_Engine Rank1 Rank2 . . . Rank10]

포맷으로 저장

3 구현 기능 – (4) 상관관계

2019 빅데이터 처리 및 응용

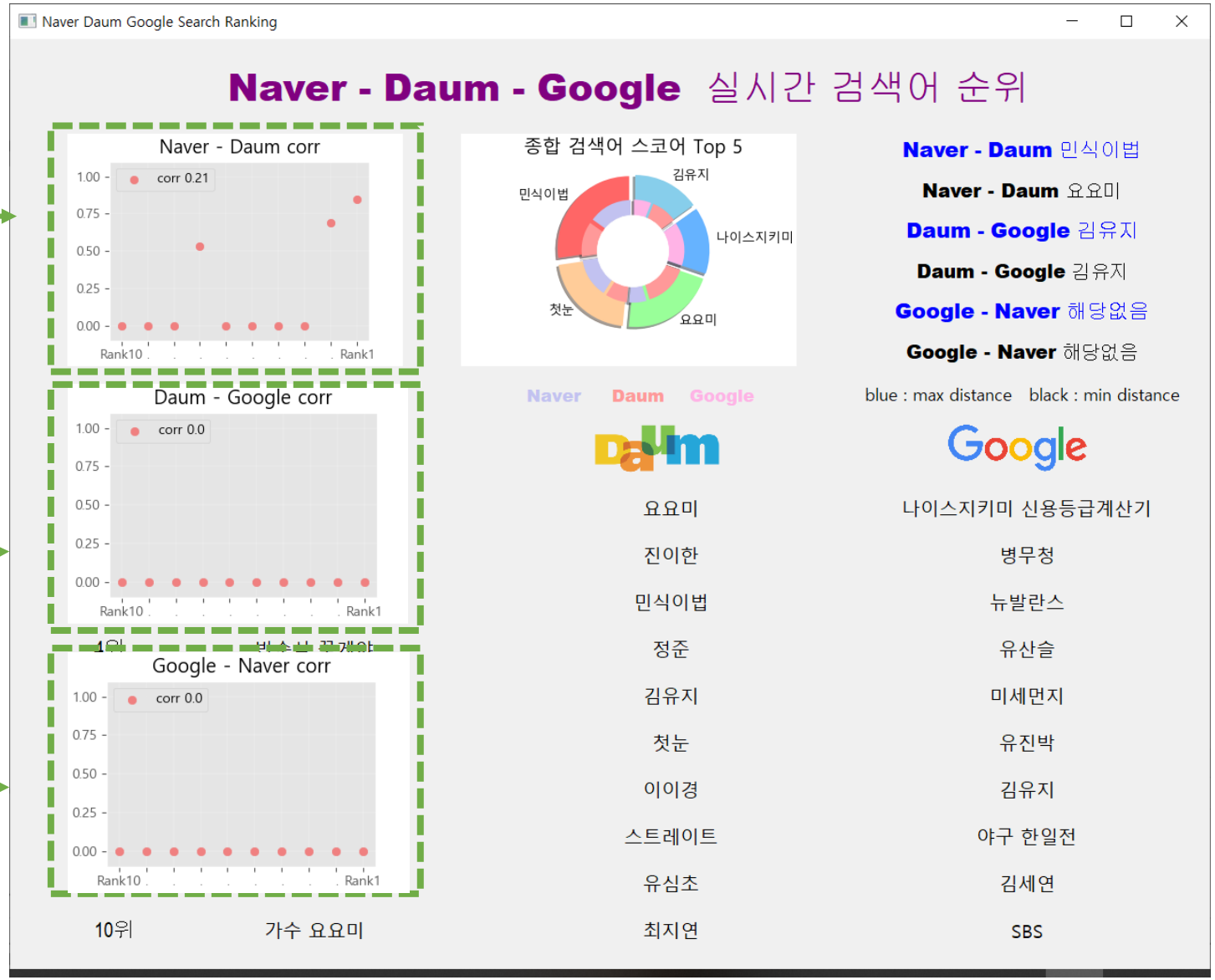
상관관계 표시

20초마다
주기적으로
아래 순서로
업데이트

1) Naver - Daum

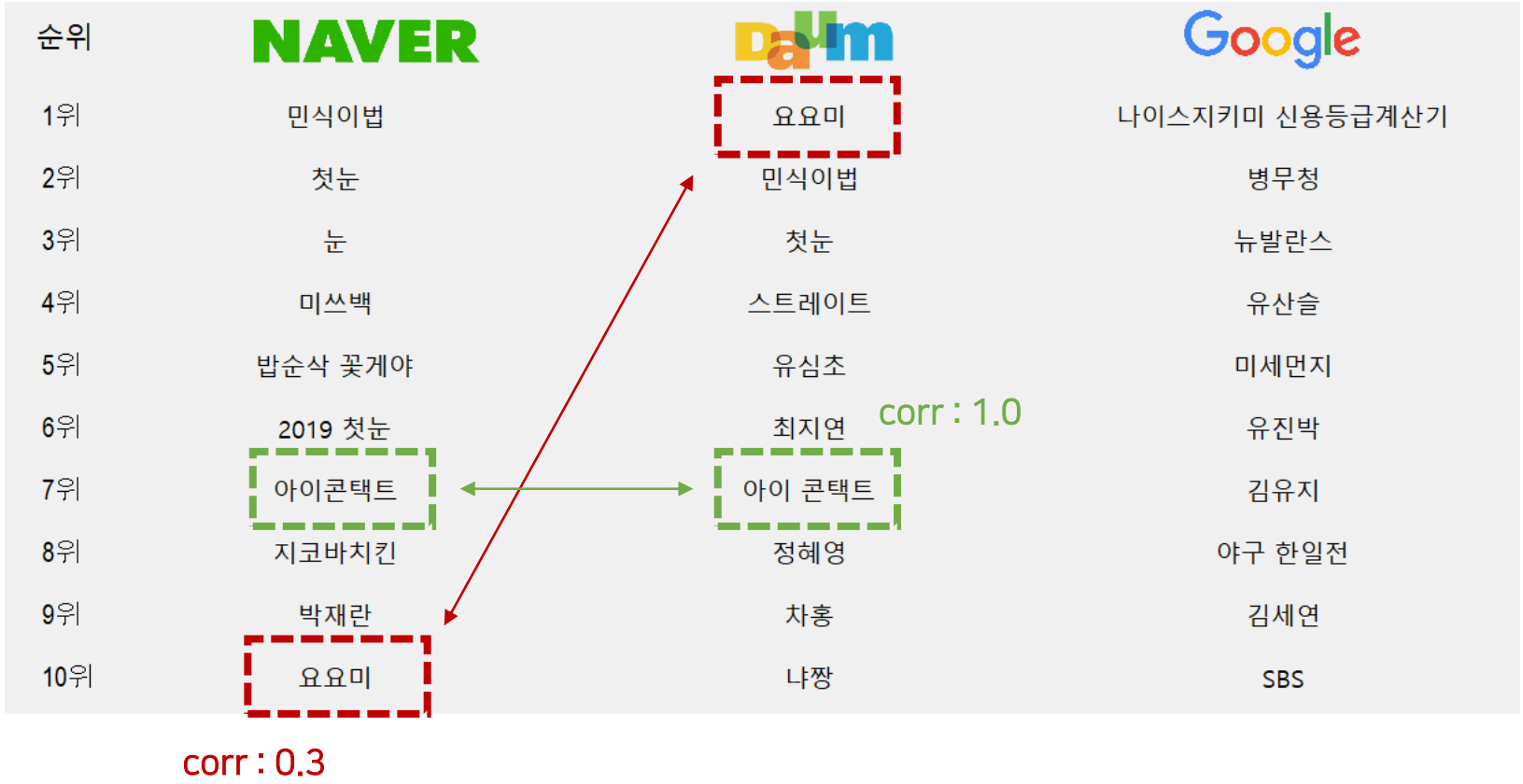
2) Daum – Google

3) Google - Naver



3 구현 기능 – (4) 상관관계

2019 빅데이터 처리 및 응용



< Example >

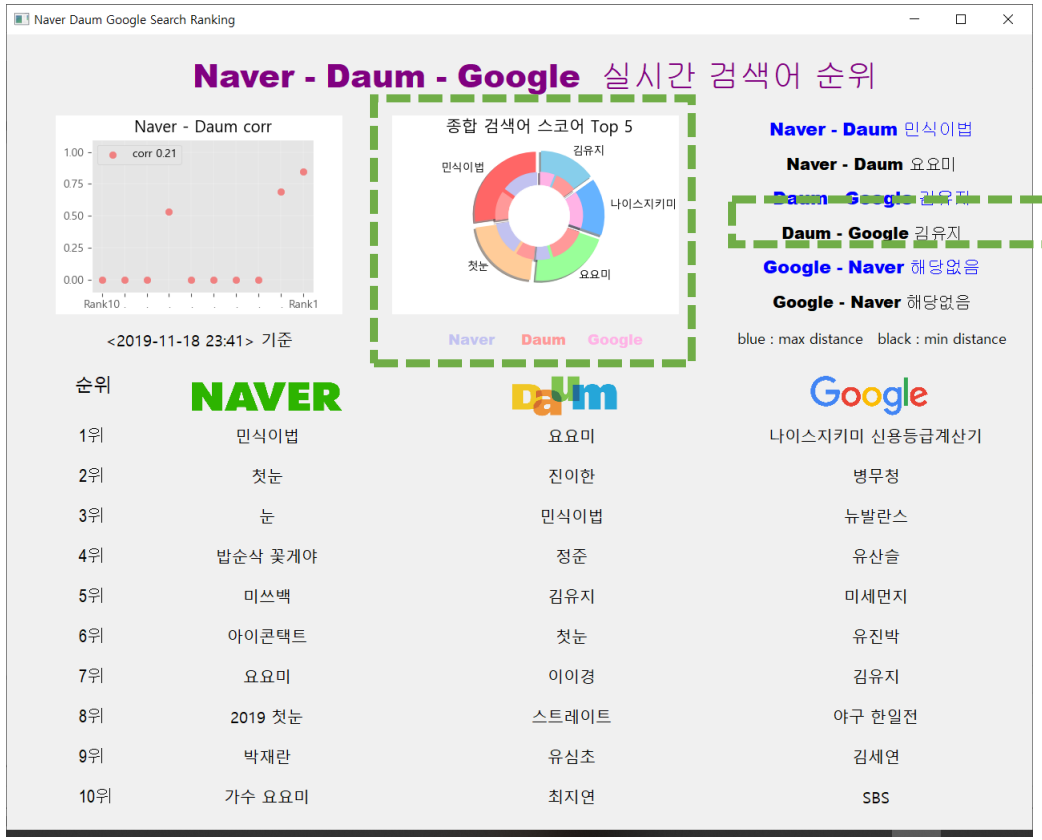
Distance	Corr
0	1.00
1	≒ 0.93
2	≒ 0.86
3	≒ 0.78
4	≒ 0.70
5	≒ 0.62
6	≒ 0.54
7	≒ 0.46
8	≒ 0.38
9	0.30
∞	0.00

< corr table >

검색어 순위 내에서 가장 차이가 많이 나는 **Distance**가 9인 경우도 상관관계가 있다고 판단하여
통상적으로 **뚜렷한 상관관계**가 있다고 하는 0.3을 최소값으로 Distance에 따라 차등적으로 corr 부여

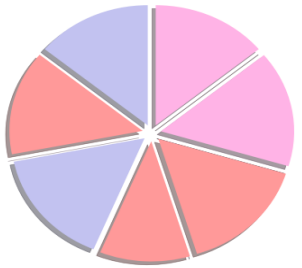
3 구현 기능 – (5) Donut + Sub chart

2019 빅데이터 처리 및 응용



< Top5 Keywords >

3 포털에서 Score Top5
키워드를 점수 순서로 Draw



< Portal Site Ratio >

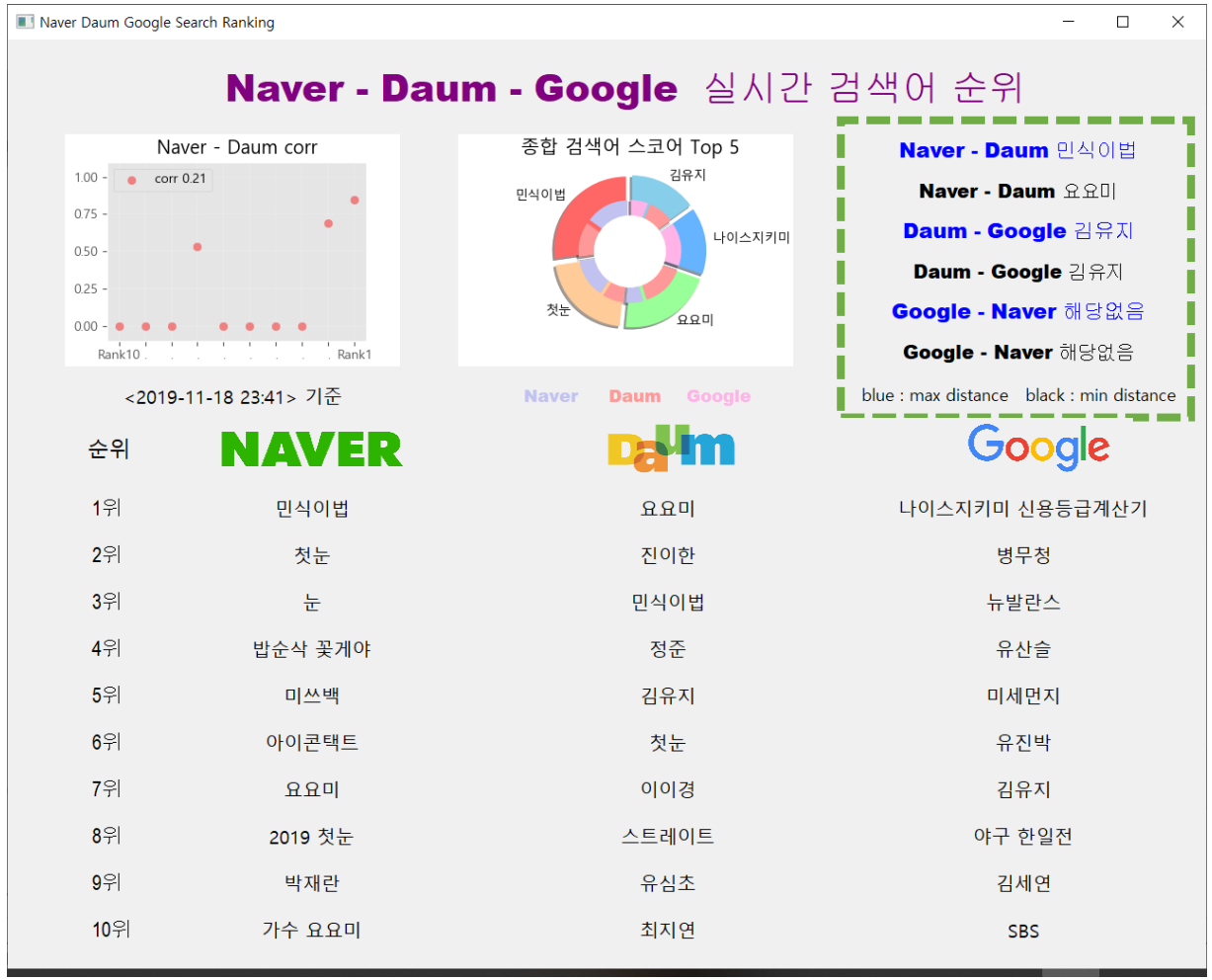
Top5 키워드에 대해서
각 포털 사이트의 점수 비율 표시

< 각 포털 사이트 순위별 점수 >

Rank	Score
1	10
2	9
3	8
4	7
5	6
6	5
7	4
8	3
9	2
10	1
else	0

3 구현 기능 – (6) 포털별 순위차이가 가장 큰 / 작은 검색어

2019 빅데이터 처리 및 응용



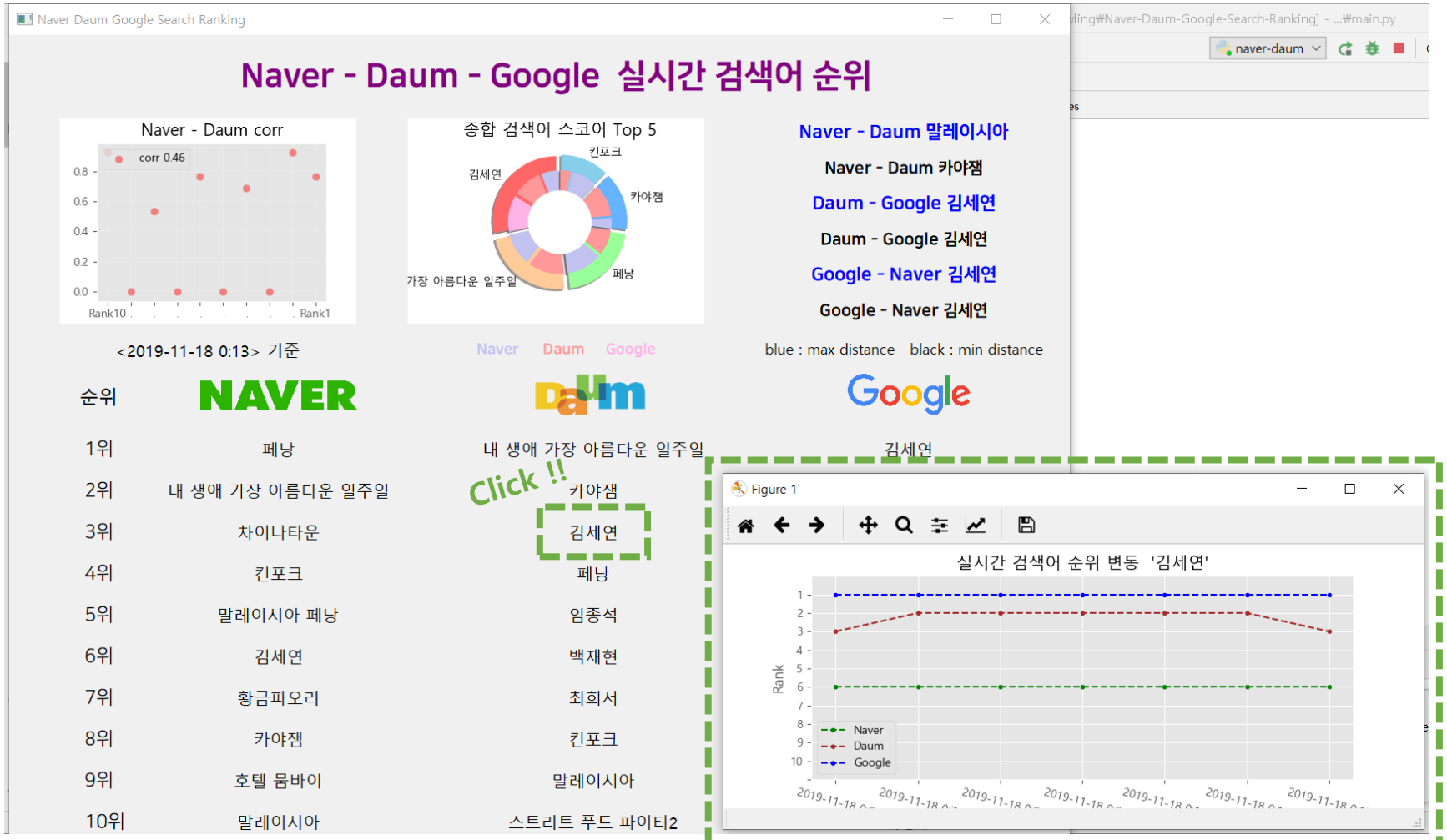
각 포털 사이트 간의
Max Distance, Min Distance를
계산해서 표시

Blue : Min Distance

Black : Max Distance

3 구현 기능 – (7) 각 검색어 클릭시 순위 변동 표시

2019 빅데이터 처리 및 응용



키워드 클릭시 해당 검색어의 순위변동을 주기별로 표시
(주기는 코드 상에서 설정 가능)

좌측 그래프는 테스트용으로 1분 단위의 순위변동 표시

3 구현 기능 – (8) 주기적 업데이트

2019 빅데이터 처리 및 응용

: 20초마다 업데이트

: 입력받은 주기마다 업데이트
sec 단위로 입력

4 최종 시연

2019 빅데이터 처리 및 응용

Q & A