1
분석배경

2
활용 데이터

3
데이터 처리 방안
& 활용 분석 기법

4
분석결과
(시각화)

5
서비스 활용 방안

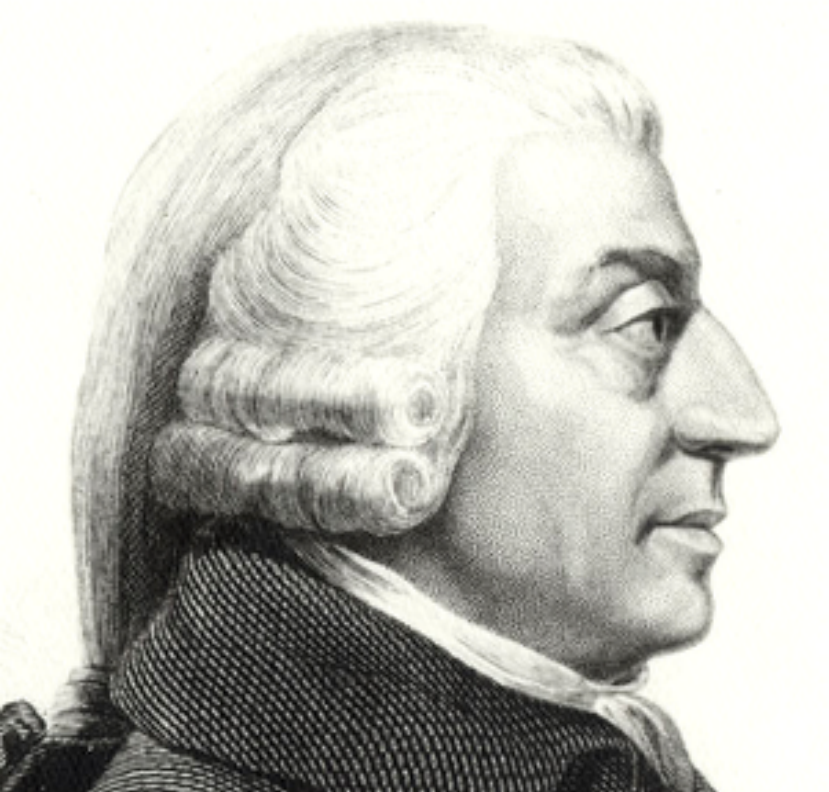
6
기대효과

1

분 석 배 경

인간은 합리적이다

애덤 스미스, ‘국부론’

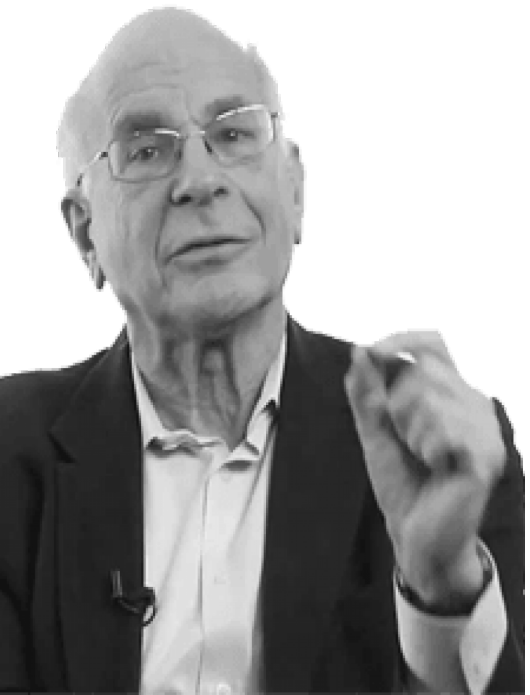


주가는 모든 정보를 반영한다

유진 파마, ‘효율적 시장 가설’ (강형 시장)

People place too much confidence in human judgement

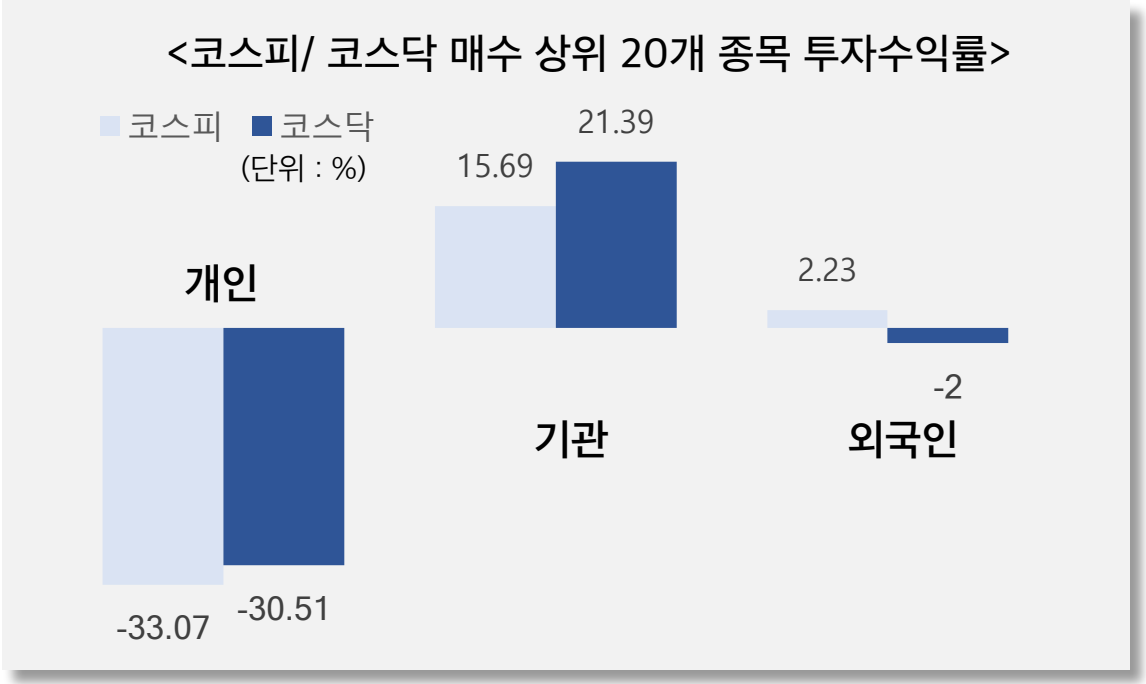
Thinking fast and slow (2011), Daniel Kahneman



System 1	System 2
Fast	Slow
Unconscious	Conscious
Automatic	Effortful
Everyday Decisions	Complex Decisions
Error prone	Reliable

국내 개인투자자의 경우 정보의 질이 상대적으로 떨어진다

'한국 주식 시장에서의 투자자 간 정보의 비대칭' (한국재무학회)



자료 : 한국거래소, 2016년 기준 출처 : 헤럴드 경제

개인투자자

개인투자자

百戰不勝

인간의 선택은 이성적이지 않다

금융시장 내 '정보의 비대칭성'

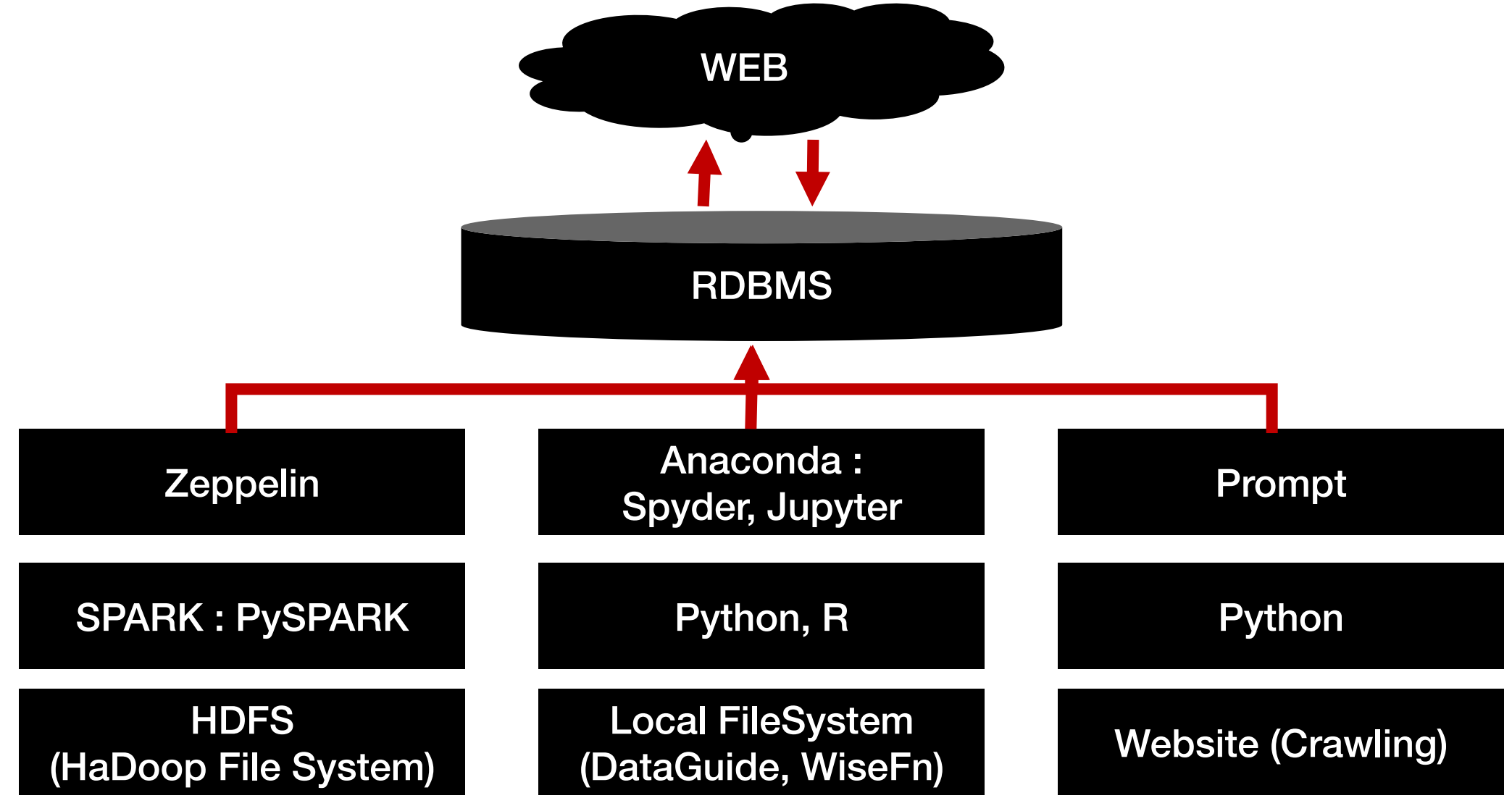
百戰安全

합리적인 투자 솔루션 제공




≠

진우
(眞友)
Portfolio
Strategy



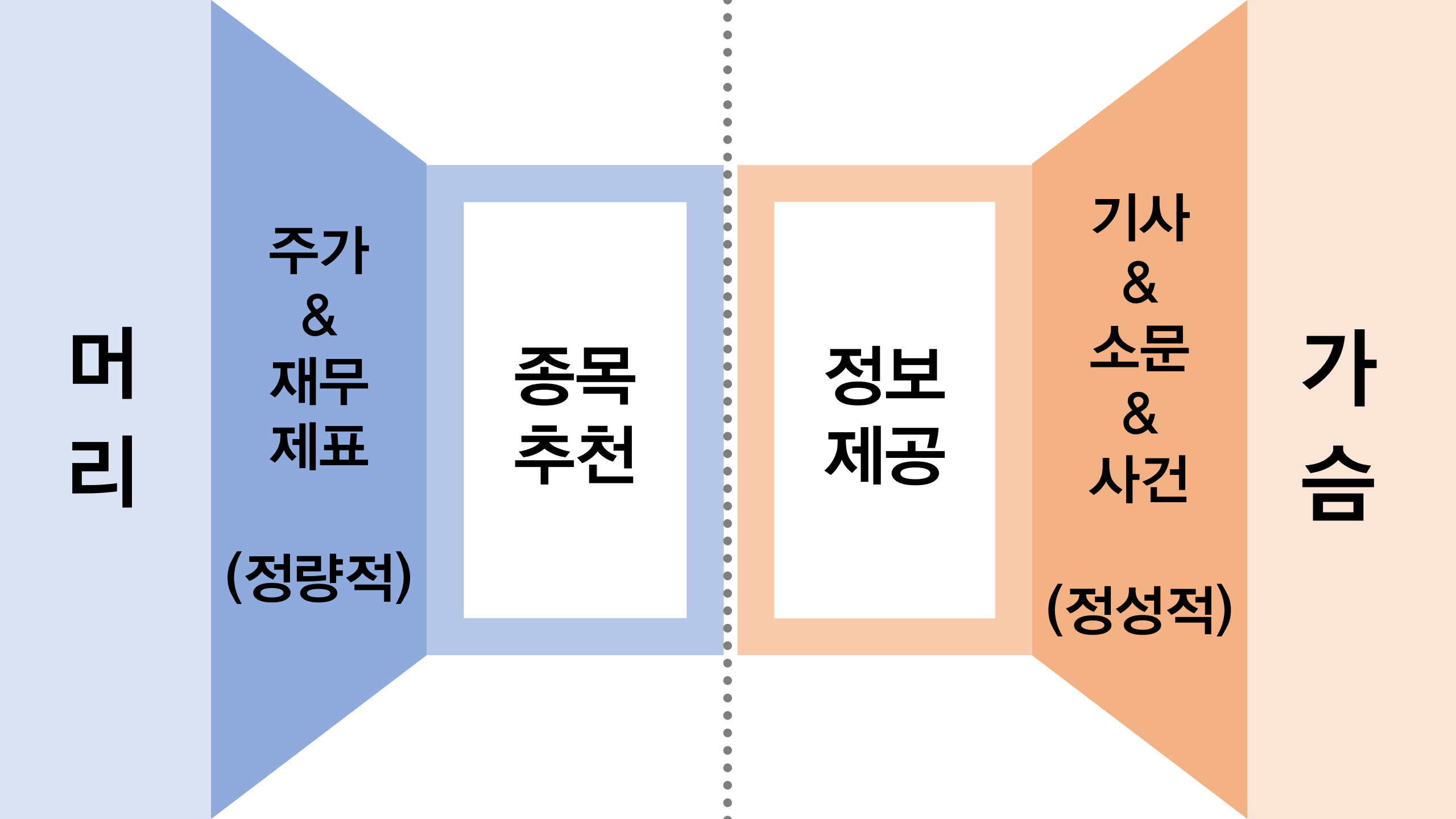
Model 1 : SVM with SGD

Model 2 : SVM + AdaBoost
Model 3 : Naïve Bayes
+ AdaBoost

Model 4 : BeautifulSOup,
Konlpy.tag, pytacloud

2

활 용 데 이 터


머리
주가
&
재무
제표
(정량적)
종목
추천
정보
제공
기사
&
소문
&
사건
(정성적)
가슴

2.1 정량 : 주가 관련 데이터 (수치)

총 139,735,532

약 1.4억 개

종목	시계열	지표(column)
KOSPI + KOSDAQ	2010.01.04 ~2017.08.28	주가, 거래량 등 (종목명, 날짜 제외)
1994개	1894일	37개

2.1 정량 : 주가 관련 데이터 (수치)

종가	시가	고가	저가	변동성(5일)	변동성(20일)
거래량	자기주식수	외국인보유비율	대차거래 체결 (5일)	대차거래상환 (5일)	공매도거래량 (5일)
베타(1일)	변동성(1일)	PER	PBR	PSR	PCR
					P.EBITDAPS
사모펀드 순매수수량	국가 순매수수량	기관계 순매수수량	금융투자 순매수수량	보험 순매수수량	투신 순매수수량
은행 순매수수량	기타금융 순매수수량	연기금 순매수수량	기타법인 순매수수량	개인 순매수수량	등록외국인 순매수수량
기타외국인 순매수수량	외국인계 순매수수량	기관 외국인계 순매수수량	전체 순매수수량	전체 매수수량	전체 매도수량

2.1 정량 : 주가 관련 데이터 (수치)

종가 시가 고가 저가	변동성(1일) 변동성(5일) 변동성(20일) 베타(1일)	PER PBR PSR PCR P.EBITDAPS	대차거래체결대 차거래상환 공매도거래량	사모펀드 국가 기관계 금융투자 보험 …
⌄	⌄	⌄	⌄	⌄
시장의 판단	민감도	재무적 요인 고려	하방 베팅	투자 주체 별 움직임

3

데이터 처리 방안 & 활용 분석 기법

3.1

주 가 데 이 터

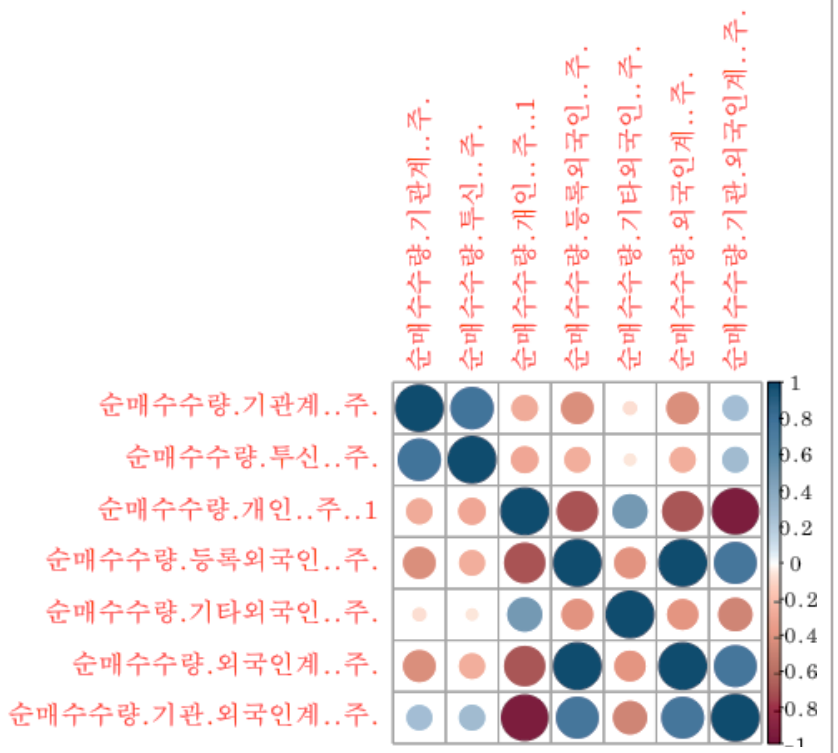
Data Handling (Domain & Cutting)

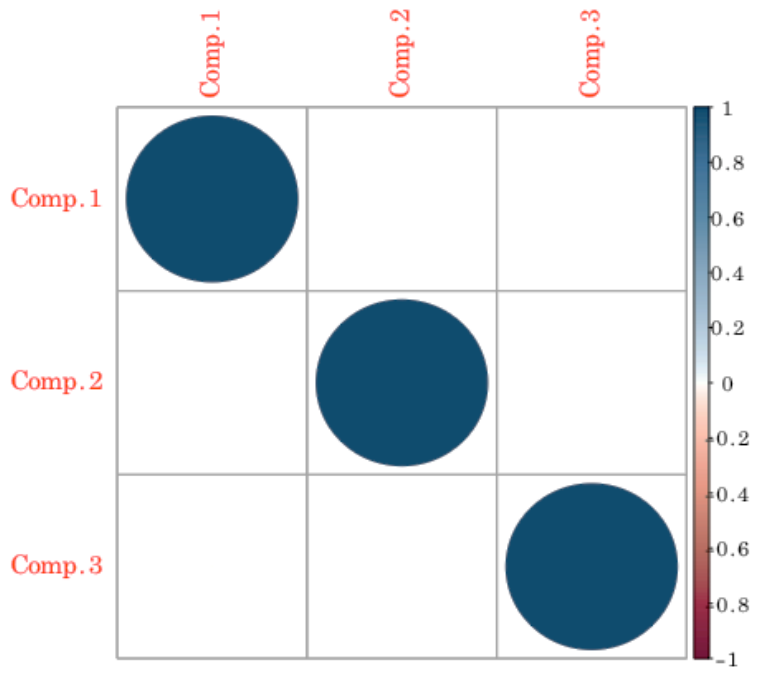
상호 연관성 높은 지표	거래 정지 기록 有 종목	데이터 10% 유실 지표	4분기 단위 지표 삭제
Ex) 거래량 & 거래대금	종가에서 null값 있는 종목 제거	거래가 없거나 유실된 날짜가 전체 일수 중 10% 넘어가는 경우	재무비율 中 리스트 level이 4개인 것 삭제
거래 관련 column 38개 –〉 19개	종목수 1994 –〉 1627	종목별 상이	ROA, ROE, ROIC, EPS 등 column삭제

Data Handling (Filtering & Scaling)

PCA (Principal component analysis)

Standard Scale





종목명	날짜	A1	A2	...	A36	A37	Up/Down
삼성전자	2010.01.04	XXXXX	XXXXXX		XXXXX	XXXXX	0
삼성전자	2010.01.05	XXXXX	XXXXXX		XXXXX	XXXXX	1
...					...		...
삼성전자	2017.08.28	XXXXX	XXXXX		XXXXX	XXXXX	1
한미약품	2010.01.04	XXXXX	XXXXX		XXXXX	XXXXX	1
한미약품	2010.01.05	XXXXX	XXXXX		XXXXX	XXXXX	0
...					...		

Machine Learning Algorithms

분류	SVM (Support Vector Machine)	강화	AdaBoost
	지도 학습. Hyper-plane을 이용해 카테고리 나눔		약한 classifiers을 조합하여 더 강한 classifier로 만듦
	Naïve Bayesian	최적화	SGD (Stochastic Gradient Descent)
	지도 학습. 독립. 비슷한 군집을 찾아주는 classifier		손실 함수를 최소화시키는 parameter 찾기

Quant (Quantitative Analysis in Financial Investment)

검증	Backtesting	자산 분배	Rebalancing

Data

Algorithms Ensemble

SVM + AdaBoost	Python Spyder	분류 + 강화
Naïve Bayesian + AdaBoost	Python Jupyter	
SVM + SGD	PySpark	분류 + 최적화

Quant Methods

Rebalancing	Java	자산 분배
Backtesting	Python, Java	실제 검증

종목명	날짜	수익률	SVM+Ada Precision	Naïve+Ada Precision	SVM+SGD Precision	Up/Down
삼성전자	2010.01.05	XXXXX	XXXXXX	XXXXX	XXXXX	0
삼성전자	2010.01.06	XXXXX	XXXXXX	XXXXX	XXXXX	1
	28일 실제 수익률		...			28일 예측치
삼성전자	2017.08.28	XXXXX	XXXXX	XXXXX	XXXXX	1
한미약품	2010.01.05	XXXXX	XXXXX	XXXXX	XXXXX	1
한미약품	2010.01.06	XXXXX	XXXXX	XXXXX	XXXXX	0
...						

(1) SVM&Ada + Naïve&Ada,

(2) SVM&Ada + SVM&SGD

(3) Naïve&Ada + SVM&SGD

(4) SVM&Ada + Naïve&Ada + SVM&SGD

종목명	날짜	수익률	SVM+Ada Precision	Naïve+Ada Precision	SVM+SGD Precision					Up/Down
삼성전자	2010.01.05	XXXXX	XXXXXX	XXXXX	XXXXX					0
삼성전자	2010.01.06	XXXXX	XXXXXX	XXXXX	XXXXX					1
삼성전자	2017.08.28	XXXXX	XXXXX	XXXXX	XXXXX					1
한미약품	2010.01.05	XXXXX	XXXXX	XXXXX	XXXXX					1
한미약품	2010.01.06	XXXXX	XXXXX	XXXXX	XXXXX					0
...										

종목명	날짜	수익률	Naïve+Ada Precision	Up/Down
삼성전자	2017.08.28	XXXXX	XXXXX	0
LG전자	2017.08.28	XXXXX	XXXXX	1
한미약품	2017.08.28	XXXXX	XXXXX	1
이마트	2017.08.28	XXXXX	XXXXX	1
		…		

(1)
먼저 up으로
예측한 값들을
추리고

(2) 그 중에서 가장 정확도가 높은
상위 4개 기업을 선택한다

Backtesting

a_k : k company

n_{a_k} : a number of stocks of k company

$P_{a_k,\ c-t}$: c-t time's price of k company

c : current

t : t days

$A_{c-t} = n_{a_x} * P_{a_x,\ c-t} + n_{a_y} * P_{a_y,\ c-t} + \ldots$: c-t time's amount invested

$$\frac{A_{n+c-t} - A_{n-1+c-t}}{A_{n-1+c-t}} = R_{n+c-t}$$: ratio of returns

$$\sum_{n=1}^{t} \frac{A_{n+c-t} - A_{n-1+c-t}}{A_{n-1+c-t}} = S_n$$: n+c-t time's cumulative returns

Rebalancing

A : asset

a_k : k company

n_{a_k} : a number of stocks of k company

p_{a_k} : price of stock of k company

[1st] $a_1 , a_2 , \ldots , a_t$

[2nd]

A/t : amount of Asset per stock

$Max(n_{a_k} * p_{a_k}) < A/t$

[3rd]

while

if(1),

$A - (n_{a_1} * p_{a_1} + n_{a_2} * p_{a_2} + \ldots + n_{a_k} * p_{a_k}) > Min(n_{a_1} * p_{a_1} , n_{a_2} * p_{a_2} , \ldots , n_{a_k} * p_{a_k}).price$

(else : out)

so,

$Max(n_{a_1} * p_{a_1} , n_{a_2} * p_{a_2} , \ldots , n_{a_k} * p_{a_k}) - Min(n_{a_1} * p_{a_1} , n_{a_2} * p_{a_2} , \ldots , n_{a_k} * p_{a_k}) = D$

$Min(n_{a_1} * p_{a_1} , n_{a_2} * p_{a_2} , \ldots , n_{a_k} * p_{a_k}).number + 1$

if(2),

$Max(n_{a_1} * p_{a_1} , n_{a_2} * p_{a_2} , \ldots , n_{a_k} * p_{a_k})^{\#} - Min(n_{a_1} * p_{a_1} , n_{a_2} * p_{a_2} , \ldots , n_{a_k} * p_{a_k})^{\#} = D^{\#}$

$D > D^{\#}$

(else : $Min(n_{a_1} * p_{a_1} , n_{a_2} * p_{a_2} , \ldots , n_{a_k} * p_{a_k}).number - 1$)


3
.2
뉴 스 데 이 터

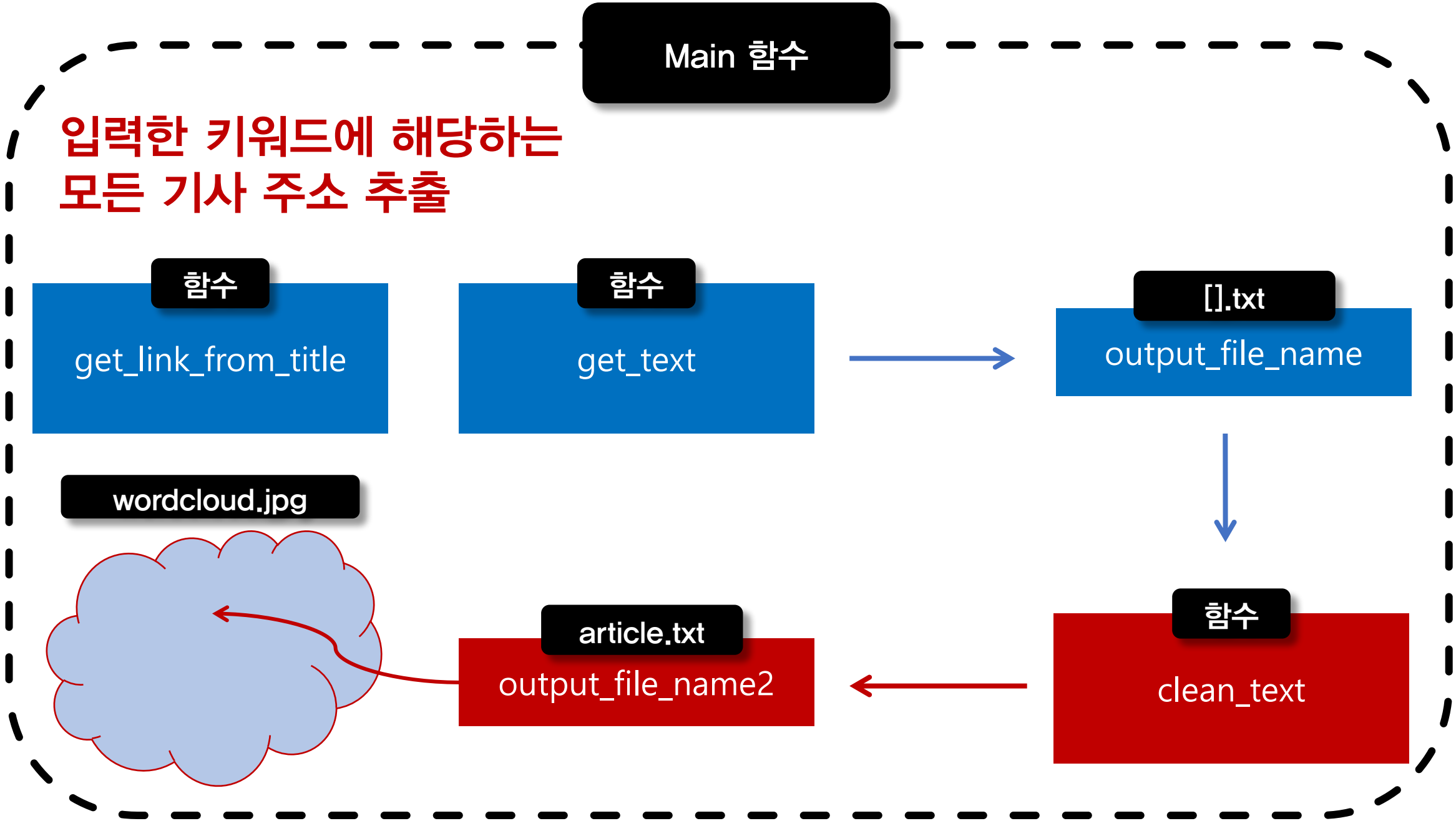
(1) 동아일보 Webpage Crawling

(2) 제외 단어 리스트 입력

```
# 워드 클라우드 생성
    wordcloud = open("article.txt")
    text = wordcloud.read()
    analysis = Twitter() #Twitter 분석기 사용
    nouns = analysis.nouns(text)
    #명사 분석 #1. 한 음절 제외 2. 시간 관련 제외 3. 정당, 정치색 짙은 단어 제외(정부정책은 제외) 4. 의미없는 단어 제외 5. 중요도 낮은 단어 제외 6. 연관성 
    stop_words = [#//한 음절 제외
                  '것','이','수','일', '등', '년', '위', '를', '창', '송', '만', '책', '활', '원', '석', '지', '치', '추', '날', '더', '번', '억', '조',
                  '더', '의', '고', '며', '웹', '및', '개', '달', '중', '줌', '명', '약', '도', '선', '시', '처', '용', '곳', '여', '윤', '사', '명', '
                  '씨', '심', '속', '방', '못', '께', '잠', '안', '내', '제', '최', '분', '뒤', '주', '코', '회', '말', '로', '점', '데', '차', '세', '
                  #//시간 관련 제외
                  '올해', '작년', '오늘', '어제', '내일', '모레', '이번', '이후', '오전', '오후', '저녁', '아침', '오시', '지금', '최근', '과거', '년반',
                  #정당, 정치색 짙은 단어 제외 #안철수 제외
                  '자유한국당', '더불어민주당', '더민주', '국민의당', '홍준표', '추미애', '정의당', '바른정당', '여당', '야당', '김무성', '유승민', '문재인
                  #//의미없는 단어 제외
                  '정말', '대신', '부형', '관련', '마리', '가장', '근시', '대해', '에서', '경우', '이상', '우리', '연주', '통해',
                  #//중요도 낮은 단어 제외
                  '기자', '나라', '교수', '색깔', '사설', '사장', '회사', '사업', '기준', '직원', '그룹', '회장', '상무', '뉴스', '대표',
                  #//연관성 없는 단어 제외
                  '고승도치', '비키니', '마녀사냥', '정성희', '도자기', '프라이어', '부친상', '우물쭈물', '버스',
                  #//자신이 검색한 keyword 제외
                  keyword] #제외 단어 리스트

    nouns = [each_word for each_word in nouns if each_word not in stop_words] #제외 단어 제거
    count = Counter(nouns) #nouns 계산
    tags = count.most_common(40) #빈도 수 내림차순 정렬
    taglist = pytagcloud.make_tags(tags, maxsize=80) #pytacloud 사용해서 워드클라우드 그리기
    pytagcloud.create_tag_image(taglist, 'wordcloud.jpg', size=(800, 500), fontname='korean', rectangular=False) #워드클라우드 생성 후 wordcloud.jpg로
    wordcloud.close() # wordcloud 종료


if __name__ == '__main__':
    main(sys.argv)
```


Main 함수
입력한 키워드에 해당하는
모든 기사 주소 추출
함수
get_link_from_title
함수
get_text
[].txt
output_file_name
함수
clean_text
article.txt
output_file_name2
wordcloud.jpg

4

분 석 결 과 (W E B)

5

서 비 스 활 용 방 안

Value Model	Customer Model	Financial Model
Who-What “개인투자자” “효율적인 투자 솔루션 제공”	**금융산업주요트렌드** “Convergence” “Safeguard Investment”	**수익창출** “Subscription Model”

로보어드바이저

‘진우(眞友)’

6

기 대 효 과

정보의 **비대칭성** 해소	개인투자자의 **투기성 투자** 방지	금융 시장 **하부구조**의 발전

개인 투자자를 위한
건전한 금융 투자 환경 구축