TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 11

전병진 FINGEREDMAN (fingeredman@gmail.com)

Part 10.

비정형데이터를 활용한 예측 모델링

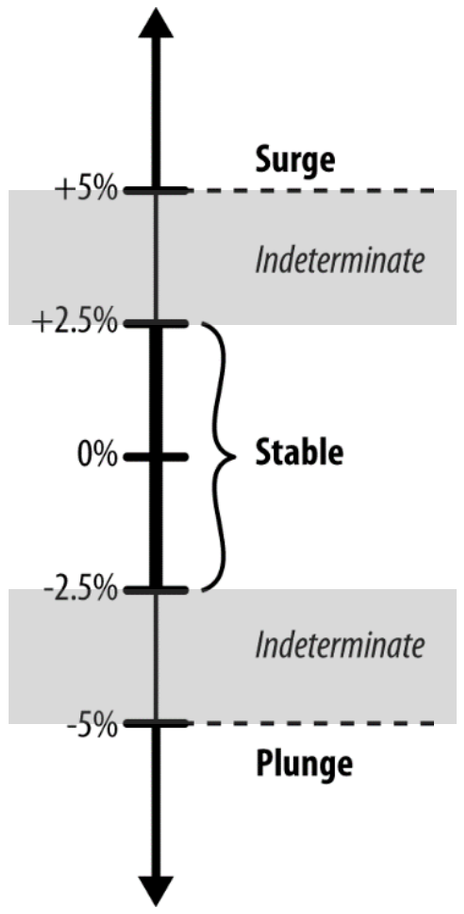
뉴스기사 마이닝

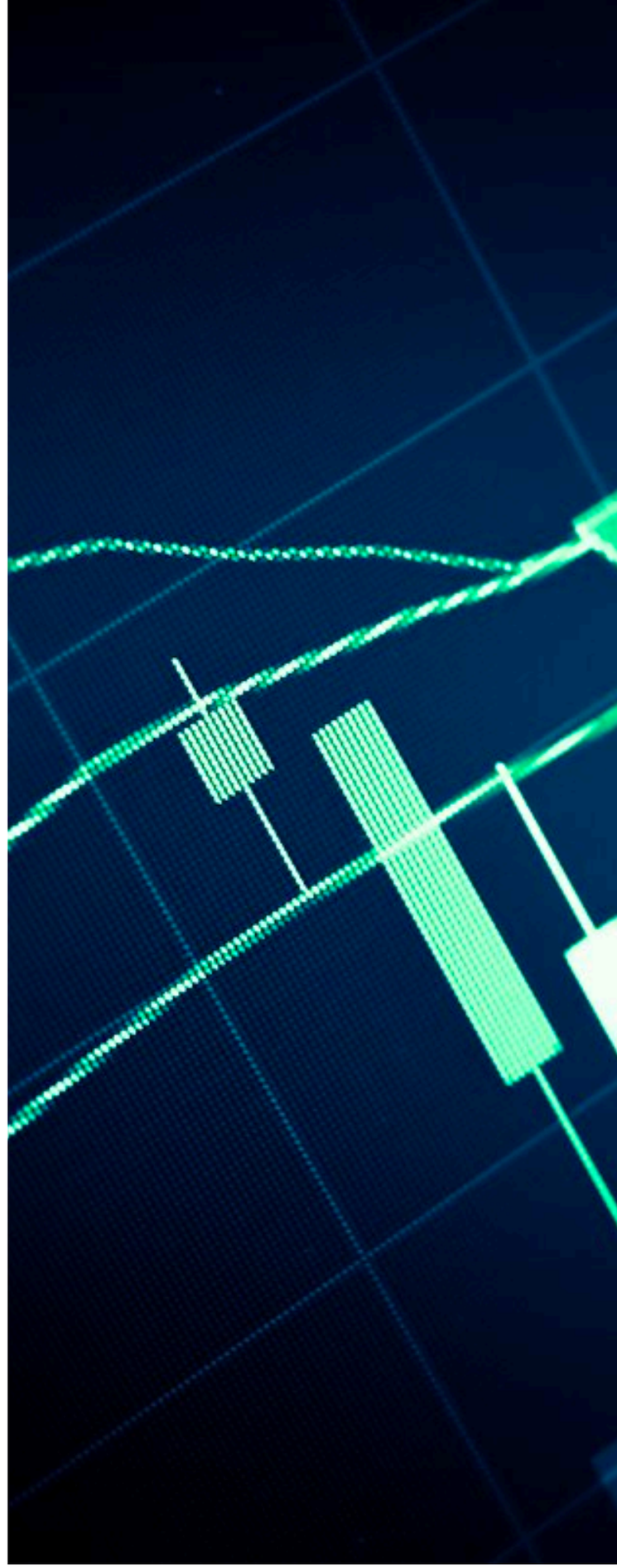
뉴스 기사에 들어 있는 텍스트에 기반해 주가 변동을 예측

- 실제 주가는 복잡한 요소들의 영향을 받아 변경되며, 모든 요소들이 뉴스에 기사화 되는것이 아니기 때문에 뉴스 추천을 목적으로 데이터 마이닝 목적을 축소
- 텍스트 마이닝을 통해 우리가 주의를 기울여야 할(주가 변동에 영향을 주는) 뉴스 기사를 예측

문제를 단순화하기 위한 가정

- 뉴스 기사가 '당일' 주가에만 어떤 영향을 미칠지 예측
- 주가의 변동을 '변동'과 '안정' 두 가지 경우로 단순화
- 비교적 변동폭이 큰 변화만 예측하고 '회색 지대'를 이용해 변동과 안정의 구분을 세분화
- 특정 증권을 언급하는 뉴스만 주가에 양향을 미침



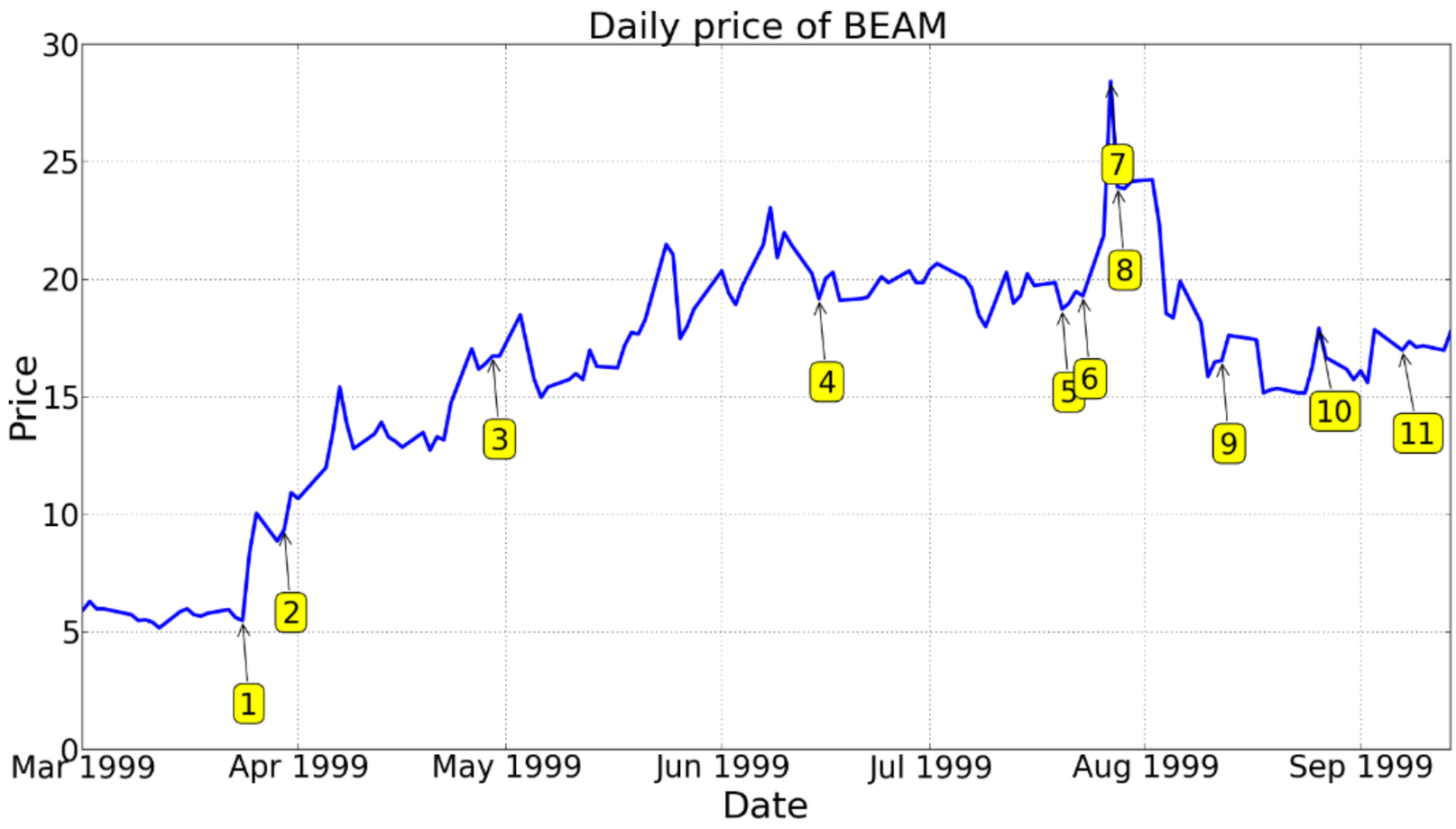


*Source : 포스터 프로보스트, 톰 포셋, Data Science for Business, 2014.6.30., O'REILLY.

뉴스기사 마이닝

데이터 소개

- 뉴스 기사와 일간 주가 변동을 나타내는 두 개의 시계열 (time series) 데이터
- 1999년 부터 뉴욕 증권 거래소와 나스닥에 상장된 증권들의 이력 데이터
- 주요 거래소에서의 증권의 시가 및 종가, 총 3만 6천 건의 연간 금융 뉴스 기사 요약본

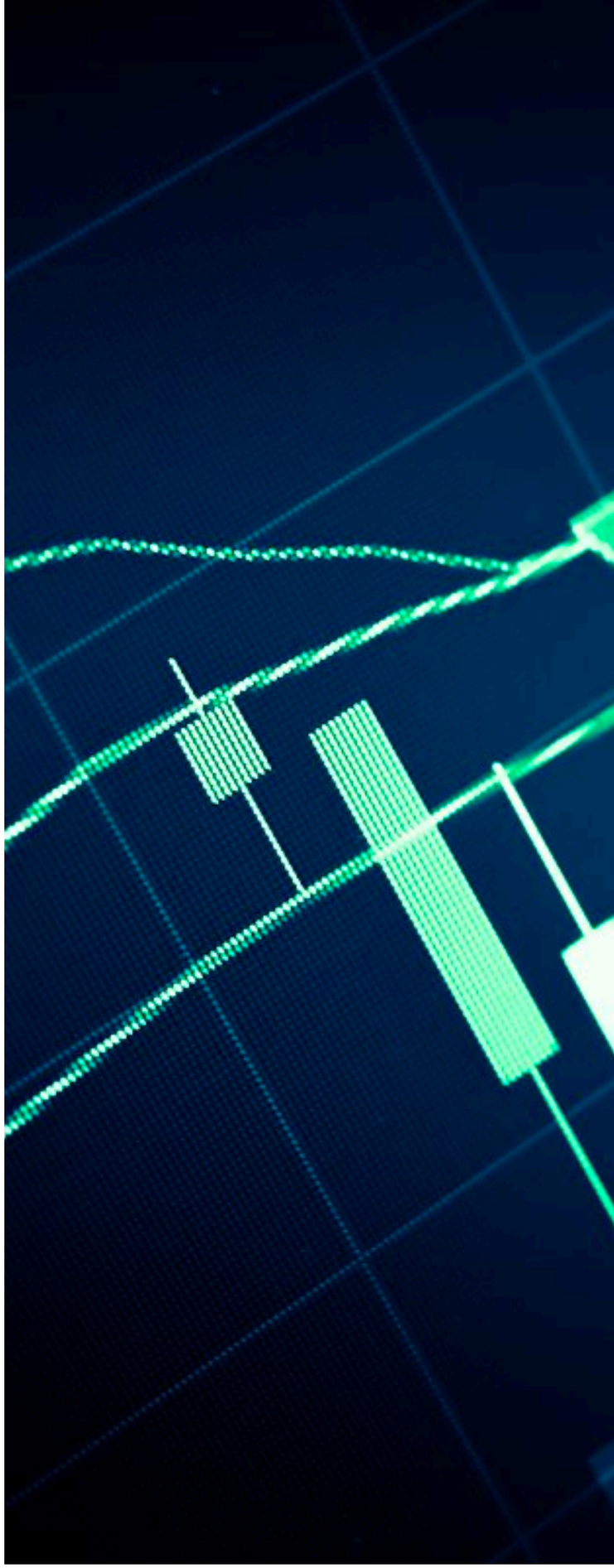


*Source : 포스터 프로보스트, 톰 포셋, Data Science for Business, 2014.6.30., O'REILLY.

뉴스기사 마이닝

데이터 전처리

- 상당수의 사건이 장이 폐장된 후에 발생하기 때문에 주가는 주식시장이 개장하는 9시 30분이 아닌, 10시에 측정
- 오후 4시 가격에서 오전 10시 가격은 뺀 값을 폐장 시각의 가격으로 나누어 당일 주가 변동률을 계산
- 기사와 기사에 영향을 받은 회사와의 긴밀한 관계만을 찾아내기 위해 3개 이상의 증권에 대해 언급하는 기사는 제외
- 각각의 기사를 TF-IDF 표현으로 축소 : 모든 단어를 소문자로 변경하고 어근을 추출해 불용어를 제거
- 바이그램을 만들어 모든 뉴스 기사를 한 단어 및 인접한 두 단어의 쌍으로 표현
- 가격 변동폭 기준에 따라 각각의 뉴스에 변동 또는 안정으로 레이블을 붙여 1만 6천개의 뉴스 기사로 데이터 최종 정리



*Source : 포스터 프로보스트, 톰 포셋, Data Science for Business, 2014.6.30., O'REILLY.

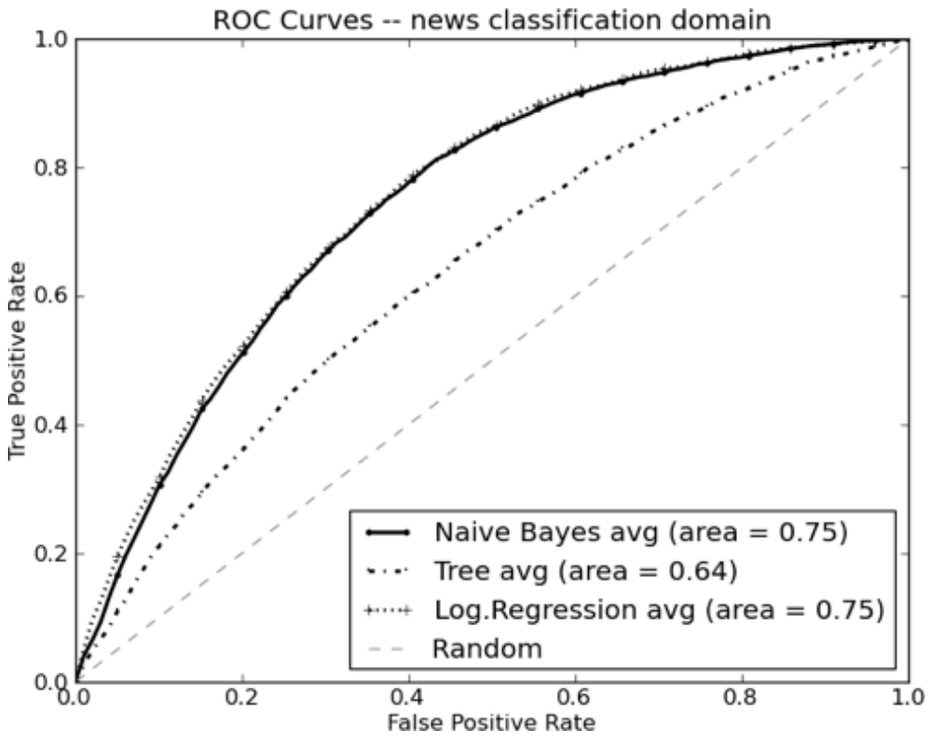
뉴스기사 마이닝

예측결과

- 분석의 목적이 뉴스 기사를 추천하는 것이므로, 의사 결정에 따른 비용과 효과에 대한 분석은 생략
- 로지스틱 회귀분석, 나이브 베이즈, 분류트리에 대한 ROC 곡선을 통해 결과 확인
- 곡선들은 변동을 양성, 안정을 음성 계층으로 분류하고, 10개의 폴드를 이용해 교차 검증한 평균값에 의한 그래프

ROC 곡선 분석 결과

- 세 가지 분류자 모두 무작위 분류자 보다 위로 많이 휘어 있고, AUC 값도 0.5 이상으로 예측력이 상당히 양호
- 로지스틱 회귀분석과 나이브 베이즈에 비해 분류 트리의 예측력이 떨어짐
- 곡선에서 두드러지게 뛰어난 부분이나 기형적으로 떨어지는 부분이 없으므로, 데이터 표현에 결함이 없음



*Source : 포스터 프로보스트, 톰 포셋, Data Science for Business, 2014.6.30., O'REILLY.

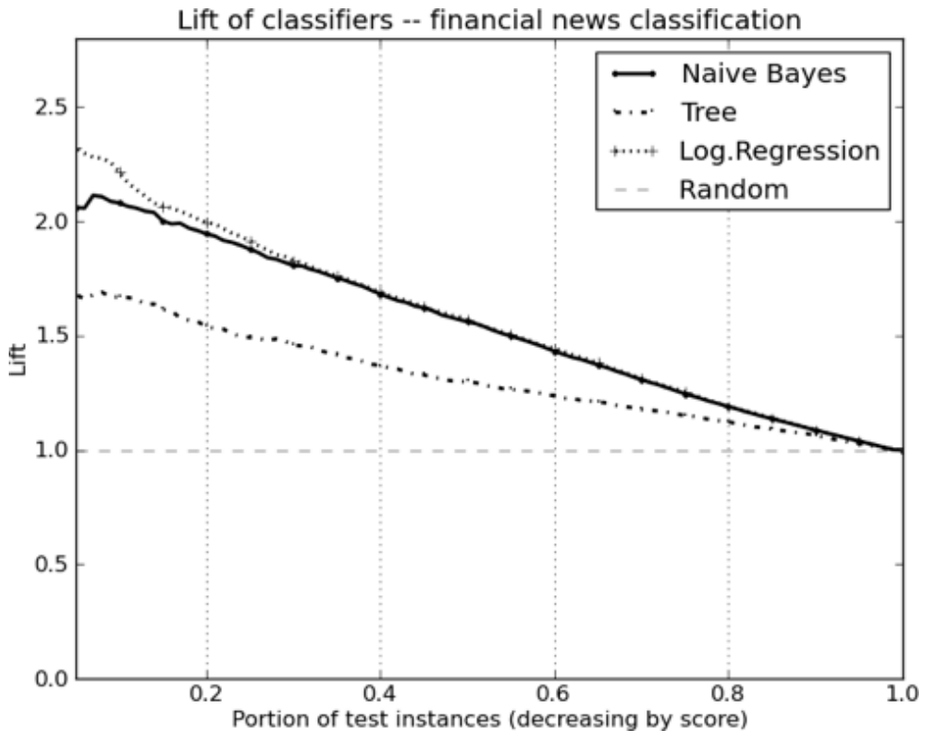
뉴스기사 마이닝

교차검증

- 분석의 목적이 뉴스 기사를 추천하는 것이므로, 의사 결정에 따른 비용과 효과에 대한 분석은 생략
- 로지스틱 회귀분석, 나이브 베이즈, 분류트리에 대한 ROC 곡선을 통해 결과 확인
- 곡선들은 변동을 양성, 안정을 음성 계층으로 분류하고, 10개의 폴드를 이용해 교차 검증한 평균값에 의한 그래프

향상도 (Lift) 곡선 분석 결과

- 각 곡선은 뉴스 기사에 점수를 매겨 순서대로 정렬했을 때 우리가 얻을 수 있는 임계치에 따른 정밀도를 보여줌
- 점 (0.2, 2)의 경우, 모든 뉴스 기사에 점수를 매기고 높은 순으로 정렬한 후, 상위 20%를 골라내면 모집단에서 무작위로 20%를 고르는 경우보다 양성객체 수가 두 배
- 모델이 분류한 뉴스 목록의 상위 20%를 골라내면 모집단에서 양성 객체의 비율이 25%이기 때문에 그 중 절반이 양성



*Source : 포스터 프로보스트, 톰 포셋, Data Science for Business, 2014.6.30., O'REILLY.

뉴스기사 마이닝

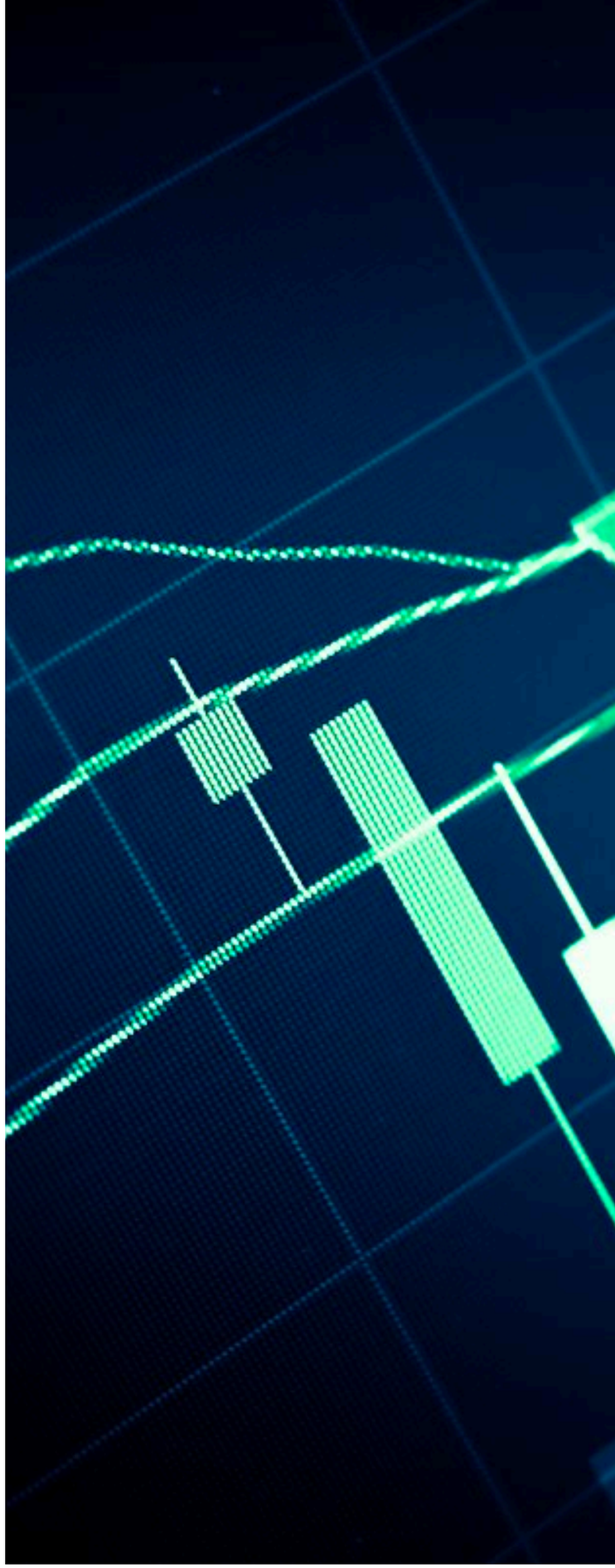
주가에 큰 영향을 미치는 단어들

- 다음의 단어나 어근(어근 뒤에는 괄호 안의 접미사가 붙음)들이 정보 전달력이 높음
- 동일한 코퍼스에 대해 맥스캐시 (Macskassy) 공저의 2001년 연구결과

alert(s, ed)
architecture
auction(s, ed, ing, eers)
average(s, d)
award(s, ed)
bond(s)
brokerage
climb(ed, s, ing)
close(d, s)
comment(ator, ed, ing, s)
commerce(s)
corporate
crack(s, ed, ing)
cumulative
deal(s)
dealing(s)
deflect(ed, ing)
delays
depart(s, ed)

department(s)
design(ers, ing)
economy
econtent
edesign
eoperate
esource
event(s)
exchange(s)
extens(ion, ive)
facilit(y, ies)
gain(ed, ing)
higher
hit(s)
imbalance(s)
index
issue(s, d)
late(ly)
law(s, ful)

lead(s, ing)
legal(ity, ly)
lose
majority
merg(ing, ed, es)
move(s, d)
online
outperform(s, ance, ed)
partner(s)
payments
percent
pharmaceutical(s)
price(d)
primary
recover(ed, s)
redirect(ed, ion)
stakeholder(s)
stock(s)
violat(ing, ion, ors)

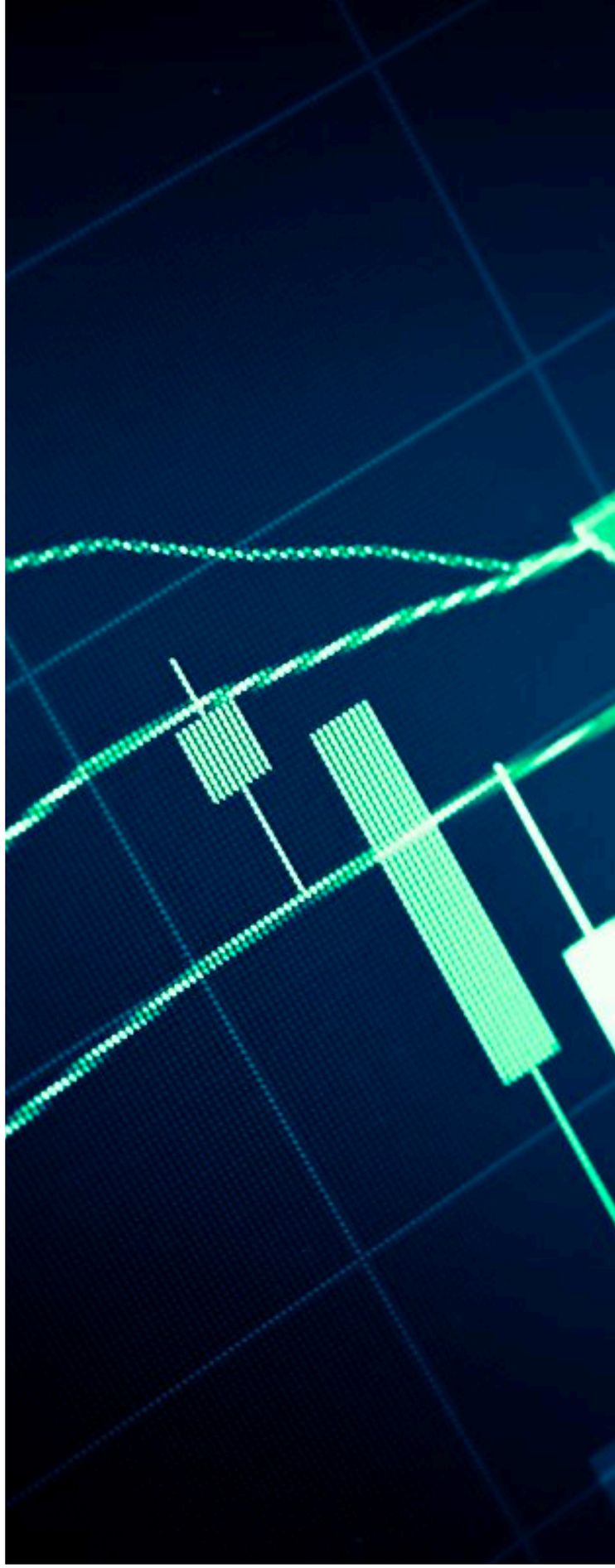


*Source : 포스터 프로보스트, 톰 포셋, Data Science for Business, 2014.6.30., O'REILLY.

뉴스기사 마이닝

결론

- 상당히 단순한 방법을 사용한 금융 뉴스 기사를 마이닝으로, 여러 방면으로 확대 가능
- 단어 주머니는 이 프로젝트에서 처음 하는 작업으로, 개체명을 인식하도록 만들면 기업이나 사람들을 더욱 잘 인식
- 뉴스 기사는 기업의 정적인 요소보다는 사건에 초점을 두고 만들어지므로 사건을 분석하면 훨씬 더 좋은 결과 도출 가능
- 객체가 사건의 주어인지 목적어인지 알기 어렵고 not, despite, except 등의 한정어는 수식하는 문장에 인접해 있지 않으므로 단어 주머니로 처리하기에 한계
- 주가 변동을 계산하기 위해 개장 및 폐장 시각의 가격만 고려했지만, 시장은 뉴스에 빠르게 반응하므로 주가와 뉴스 기사의 세세하고 정확한 시각 정보가 필요



*Source : 포스터 프로보스트, 톰 포셋, Data Science for Business, 2014.6.30., O'REILLY.

텍스트 데이터로 예측을?

실제 뉴스기사를 통해
코스피 지수를 예측해보자!

E.O.D