Chapter. 07

직접 해보기

주문서 정리하기

FAST CAMPUS
ONLINE
데이터 탐색과 전처리

강사. 안길승

I 문제 상황

- 월별로 시트가 구분되어 있는 매출 데이터를 요약하여 정리해야 함

					담당	팀장	사장
					(인)	(인)	(인)
1월 매출대장							
일자	지점	품명	수량	주문인 ID	수령 주소	주문 상태	결제 수단
2018.1.1	지점1	제품B	3	C-168	특별시 동작구	주문완료	인터넷뱅킹
2018.1.1	지점2	제품F	8	C-87	별시 서대문구	배송완료	신용카드
2018.1.1	지점4	제품B	2	C-158	별시 종로구 종	배송완료	휴대폰결제
2018.1.1	지점3	제품D	7	C-307	별시 서대문구	주문완료	휴대폰결제
2018.1.1	지점2	제품E	9	C-342	특별시 종로구	배송완료	인터넷뱅킹

⋮

					담당	팀장	사장
					(인)	(인)	(인)
12월 매출대장							
일자	지점	품명	수량	주문인 ID	수령 주소	주문 상태	결제 수단
2018-12-1	지점1	제품C	3	C-45	서울특별시	배송중	신용카드
2018-12-1	지점4	제품C	4	C-409	서울특별시	주문완료	인터넷뱅킹
2018-12-1	지점4	제품B	1	C-233	서울특별시	배송완료	신용카드
2018-12-1	지점4	제품D	7	C-115	서울특별시	배송완료	인터넷뱅킹
2018-12-1	지점2	제품B	9	C-49	서울특별시	주문완료	인터넷뱅킹

→

- 데이터 병합
- 포맷 통일 및 변수 추가
- 월별 매출 추이 파악
- 조건에 따른 판매 통계 분석
- 충성 고객 파악

Step 1. 데이터 병합

- 하나의 데이터가 12개의 시트로 구분되어 있어, 효과적인 분석을 위해 통합해야 함
- 또한, 분석에 불필요한 부분이 있어 이를 제거해야 함

					담당	팀장	사장
					(인)	(인)	(인)
1월 매출대장							
일자	지점	품명	수량	주문인 ID	수령 주소	주문 상태	결제 수단
2018.1.1	지점1	제품B	3	C-168	특별시 동작구 흑	주문완료	인터넷뱅킹
2018.1.1	지점2	제품F	8	C-87	별시 서대문구	배송완료	신용카드
2018.1.1	지점4	제품B	2	C-158	별시 종로구 종	배송완료	휴대폰결제
2018.1.1	지점3	제품D	7	C-307	별시 서대문구	주문완료	휴대폰결제
2018.1.1	지점2	제품E	9	C-342	특별시 종로구	배송완료	인터넷뱅킹

⋮

					담당	팀장	사장
					(인)	(인)	(인)
12월 매출대장							
일자	지점	품명	수량	주문인 ID	수령 주소	주문 상태	결제 수단
2018-12-1	지점1	제품C	3	C-45	서울특별시	배송중	신용카드
2018-12-1	지점4	제품C	4	C-409	서울특별시	주문완료	인터넷뱅킹
2018-12-1	지점4	제품B	1	C-233	서울특별시	배송완료	신용카드
2018-12-1	지점4	제품D	7	C-115	서울특별시	배송완료	인터넷뱅킹
2018-12-1	지점2	제품B	9	C-49	서울특별시	주문완료	인터넷뱅킹

일자	지점	품명	수량	주문인 ID	수령 주소	주문 상태	결제 수단
2018.1.1	지점1	제품B	3	C-168	특별시 동작구 흑	주문완료	인터넷뱅킹
2018.1.1	지점2	제품F	8	C-87	별시 서대문구	배송완료	신용카드
2018.1.1	지점4	제품B	2	C-158	특별시 종로구 종	배송완료	휴대폰결제
2018.1.1	지점3	제품D	7	C-307	별시 서대문구	주문완료	휴대폰결제
2018.1.1	지점2	제품E	9	C-342	특별시 종로구	배송완료	인터넷뱅킹
중략							
2018-12-31	지점1	제품B	4	C-17	시 영등포구 양	주문완료	인터넷뱅킹
2018-12-31	지점2	제품D	4	C-156	특별시 중구 을지	배송중	인터넷뱅킹

Step 2. 포맷 통일 및 변수 추가

- 일자 컬럼이 YYYY-MM-DD라는 포맷의 날짜와 YYYY.MM.DD라는 포맷의 날짜가 혼합되어 있어, 원활한 분석을 위해 **포맷을 통일**해야 함

- 품명과 수량 컬럼을 바탕으로 주문 금액을 계산하여, 주문 금액이라는 컬럼을 추가

> 주문 금액 = 수량 × 가격 × 1.1 (부가세)

- 제품에 따른 가격 정보는 제품별_가격정보.xlsx에 정의되어 있음

제품	제품A	제품B	제품C	제품D	제품E	제품F
가격	20000	5000	10000	8000	30000	15000

I 새로운 문법 (1/2)

- Series.to_dict(): Series의 index를 key로, data를 value로 하는 사전으로 변환

Index	Data
‘a’	1
‘b’	2
‘c’	3

Series.to_dict() → {‘a’: 1, ‘b’: 2, ‘c’:3}

- Series.replace(dict): Series에 있는 값 가운데 dict의 key의 값이 있으면 대응되는 value로 변환

Series

Data
1
2
3

Series.to_dict()

{1: ‘V1’, 2: ‘V2’}

Data
‘V1’
‘V2’
3

I 새로운 문법 (2/2)

- .T: 행과 열을 바꾼 전치 행렬을 반환

Index	Col1	Col2	Col3
Ind1	V11	V12	V13
Ind2	V21	V22	V23

df

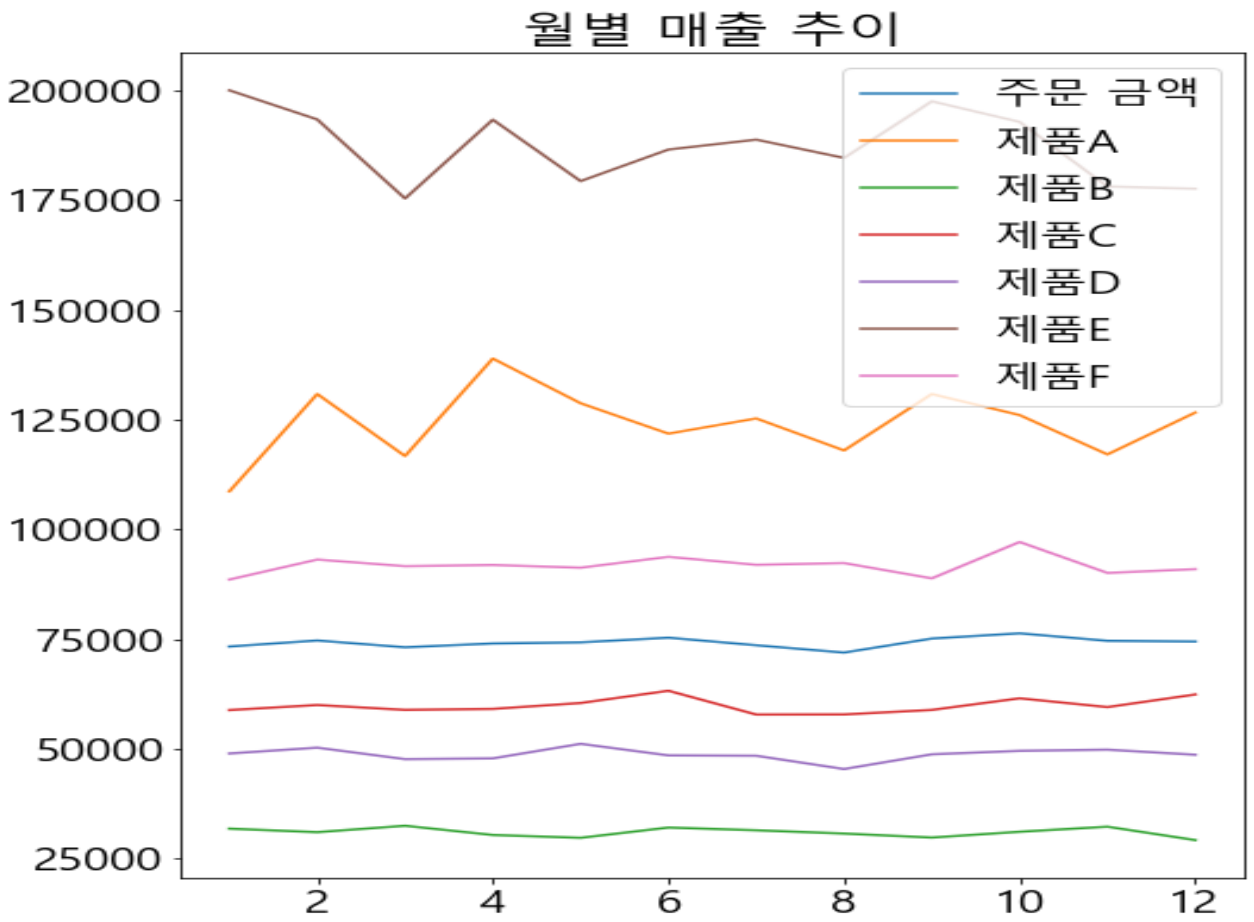
df.T →

Index	Ind1	Ind2
Col1	V11	V21
Col2	V12	V22
Col3	V13	V23

- DataFrame에 새로운 컬럼 추가: df[‘새로운 컬럼명’] = 연산 결과

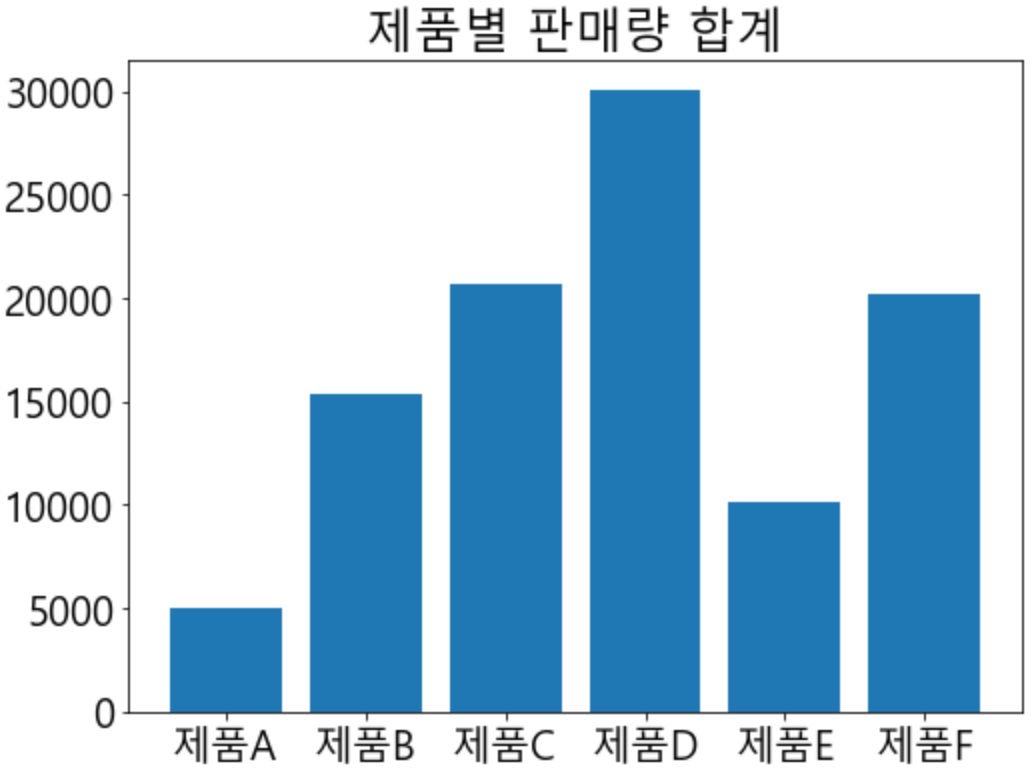
Step 3. 월별 매출 추이 파악

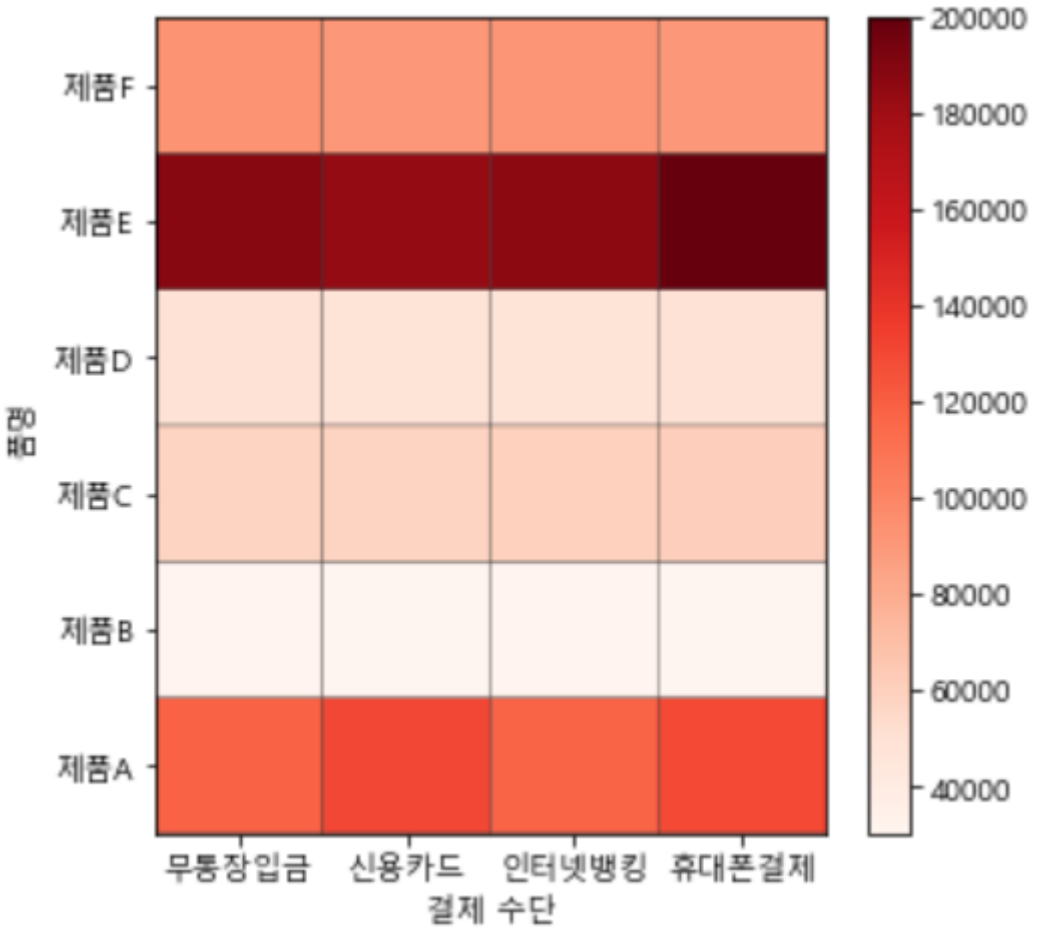
- 일자 변수를 바탕으로 월을 추출한 뒤, 월을 x축으로 주문 금액 및 품목별 주문 금액을 y축으로 하는 꺾은선 그래프를 그림



Step 4. 다양한 조건에 따른 판매 통계 분석

- 제품별 판매량 합계를 나타내는 막대 그래프와 제품과 결제 수단에 따른 히트 맵을 그림





Step 5. 충성 고객 찾기

- 주문 금액 합과 빈도가 각각 상위 10%안에 속하는 고객을 찾아 정리

주문인 ID	sum	count
C-450	4759700.0	51
C-288	4640900.0	50
C-320	4313100.0	63
C-106	4276800.0	46
C-189	4250400.0	49
C-389	4171200.0	47
C-475	4012800.0	46
C-439	3972100.0	50
C-100	3950100.0	46
C-317	3891800.0	48

Chapter. 07

직접 해보기

뉴스 기사 요약하여 정리하기

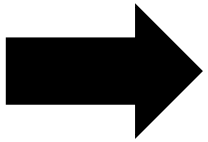
FAST CAMPUS ONLINE
데이터 탐색과 전처리

강사. 안길승

I 문제 상황

- 2018년 1월 1일부터 12월 31일까지 네이버 금융 – 뉴스의 주요 뉴스에서 크롤링을 통해 수집한 **뉴스 기사 제목**을 정리해야 함 (단, 실제 뉴스 기사 제목에는 단어가 정제 되어 있지 않지만, 이번 과제를 위해 형태소 분석을 통한 단어 정제를 수행함)

기사제목	신문사	작성일자
휴장일 슬쩍올빼미 공시 주의보	서울경제	2018-01-01
코스피 상단 28003100	서울경제	2018-01-01
한숨 동부대우전자 중동	서울경제	2018-01-01
금융투자업계 지도 생존 성장 운용자문사 대주주교체	파이낸셜뉴	2018-01-01
금융투자업계 지도 모험자본 확충 원년 초대형 질주	파이낸셜뉴	2018-01-01
미래에셋대우 2017년 주관 톱한투 2위	헤럴드경제	2018-01-01
한국기업글로벌 자본시장 가교 외국계 금융회사 한국	한국경제	2018-01-01
3개월來 최고치요즘 금값 이유	이데일리	2018-01-01
증시 1월 효과 있을까코스피 24502500 등락 전망	헤럴드경제	2018-01-01
첫주 증시 1월 효과 적 기대 호조세 전망	연합뉴스	2018-01-01



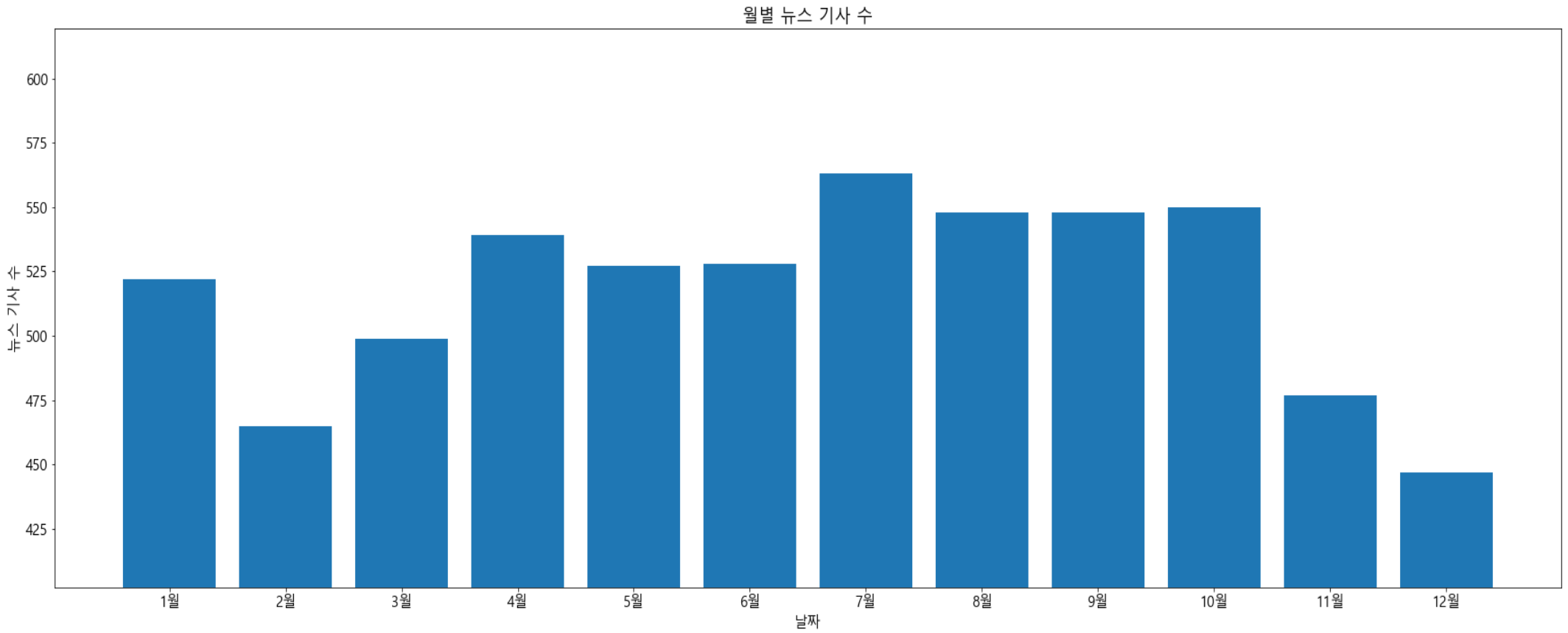
월별 뉴스 기사 수 추이 시각화

주요 단어 추출

월별 주요 단어 출현 빈도 시각화

l 월별 뉴스 기사 수 추이 시각화

- 월별 뉴스 기사 수의 합계를 바탕으로, 월별 뉴스 기사 수의 합계의 추이를 나타내는 막대 그래프를 그림

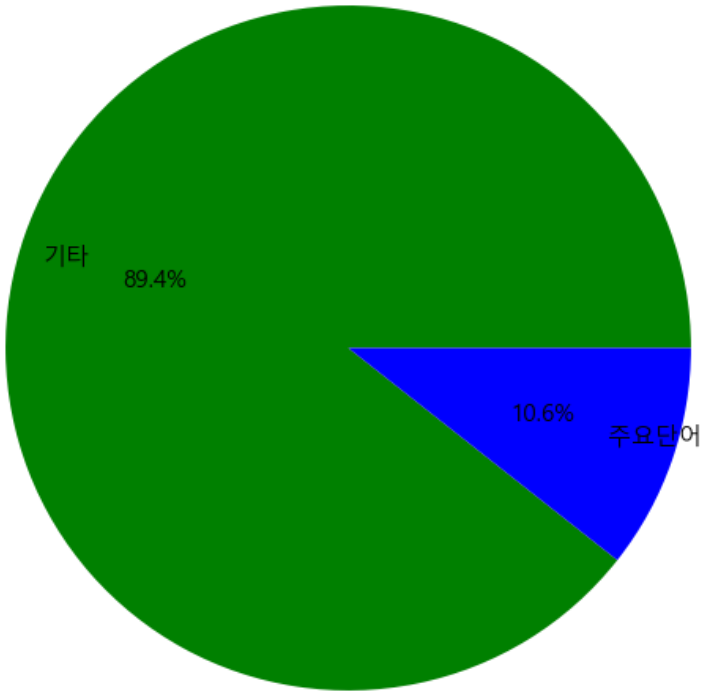


FAST CAMPUS
ONLINE
안길승 강사.

I 주요 단어 추출 및 시각화

- 기사 제목을 어절별로 구분한 뒤, 가장 자주 출현한 상위 10개의 단어와 그 빈도를 추출함
- 주요 단어와 그렇지 않은 단어의 출현 비율을 계산하여 시각화함

코스피	815
증시	745
상승	390
하락	350
코스닥	252
美	224
우려	215
마켓뷰	205
오전시황	190
환율	184



월별 주요 단어 출현 빈도 계산

- 뉴스 기사 별로 주요 단어의 출현 여부를 계산하여, 월별 주요 단어 출현 빈도를 계산함

	월	코스피_등장횟수	증시_등장횟수	상승_등장횟수	하락_등장횟수	코스닥_등장횟수	美_등장횟수	우려_등장횟수	마켓뷰_등장횟수	오전시황_등장횟수	환율_등장횟수
0	01	112.0	81.0	67.0	43.0	154.0	29.0	6.0	20.0	19.0	34.0
1	02	100.0	136.0	74.0	38.0	50.0	72.0	16.0	18.0	14.0	22.0
2	03	138.0	104.0	48.0	43.0	38.0	90.0	54.0	23.0	20.0	16.0
3	04	152.0	106.0	68.0	55.0	48.0	61.0	25.0	13.0	22.0	30.0
4	05	106.0	113.0	62.0	67.0	38.0	35.0	22.0	12.0	13.0	19.0
5	06	80.0	154.0	54.0	54.0	19.0	40.0	23.0	17.0	5.0	25.0
6	07	90.0	126.0	52.0	49.0	54.0	42.0	24.0	20.0	16.0	30.0
7	08	96.0	113.0	47.0	35.0	24.0	51.0	15.0	22.0	22.0	24.0
8	09	65.0	97.0	35.0	37.0	30.0	59.0	25.0	16.0	13.0	13.0
9	10	119.0	172.0	45.0	56.0	30.0	59.0	19.0	12.0	15.0	31.0
10	11	100.0	107.0	45.0	44.0	31.0	70.0	12.0	22.0	15.0	17.0
11	12	70.0	116.0	33.0	32.0	14.0	72.0	30.0	19.0	17.0	15.0