

# 빅데이터 통계 분석

20.09.19 토요일 회의

박채린, 이해승

# TO DO - 빅카인즈

< BigKinds 뉴스 단어 카운팅 >

- 검색어: 경기 침체
- 기간: 08.07.01 ~ 오늘
- 분류: 정치 경제 국제
- 검색어 처리: 형태소/바이그램
- 키워드 트렌드: 월간.

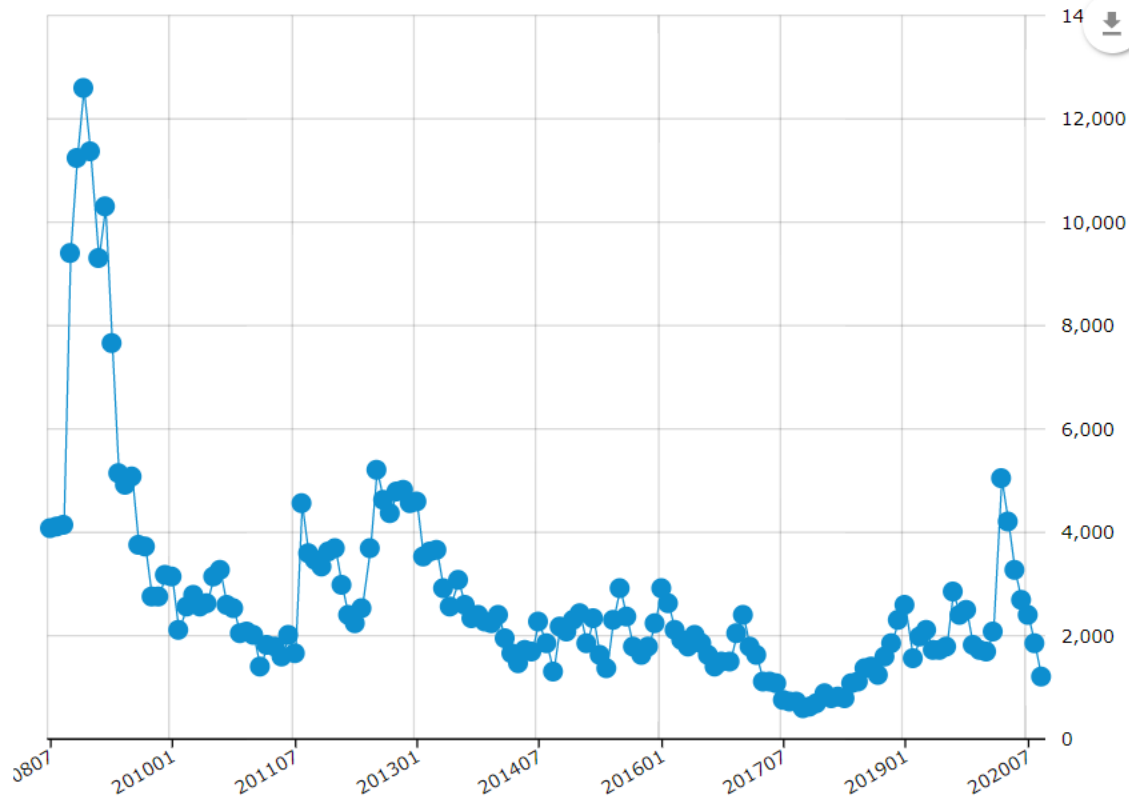
기간 선택:

차트 선택:

데이터 유형:

그래프 색상: 경기 침체

상관계수: ?



분석결과 저장

Elements Console Sources Network

```

new NewsDetail();

var search = new Search();
search.setValuesByParams({"indexName": "news", "searchKey": "국
회", "searchKeys": [], "networkNodeType": "OG", "startDate": "2020-09-
18", "endDate": "2020-09-19"});
new NewsResult(search, "2");
new DictSearch(search);
});
    
```

html body div#contents script (text)

Styles Computed Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls +

No matching selector or style

element.style { }

div, p { word-break: keep-all; application.scss:57 }

address, blockquote, body, caption, dd, div, dl, dt, fieldset, form, h1, h2, h3, h4, h5, h6, input, li, ol, p, select, table, td, textarea, th, tr, ul { margin: 0; padding: 0; box-sizing: border-box; offset.min.css:1 }

\* { -webkit-box-sizing: border-box; -moz-box-sizing: border-box; \_vendor-pre...xes.scss:75 }

Console

js?timestamp=201911081000:409:20 at b.a.inherits.b.fire (https://www.amcharts.com/lib/3/amcharts.js:1:1088) at Object.handleClick (https://www.amcharts.com/lib/3/amcharts.js:8:16380) at SVGCircleElement.<anonymous> (https://www.amcharts.com/lib/3/amcharts.js:8:15260)

# TO DO - 트위터

## <트위터>

- getoldtweets3 모듈 사용 유무 확인(robots.txt에는 \*에 대해 Disallow)
- 제대로 끌어오는 법 확인(현재 HTTP 404 error)
- 키워드: 경제  
동의어 : 경기, 생활  
포함어 : 가계, 수입, 소비, 지출, 살림살이, 생활형편  
배제어 : 운동, 수입품, 수출
- 크롤링 할 데이터: 날짜, 텍스트, 리트윗수(리트윗 수 = 공감 → 가중치)
- (-) 트위터 검색에 생각보다 이상한 트윗들이 많이 낀  
→ NLP로 공간에 벡터화해서 유사도 높은 것들만 사용. Or 배제어 활용..?

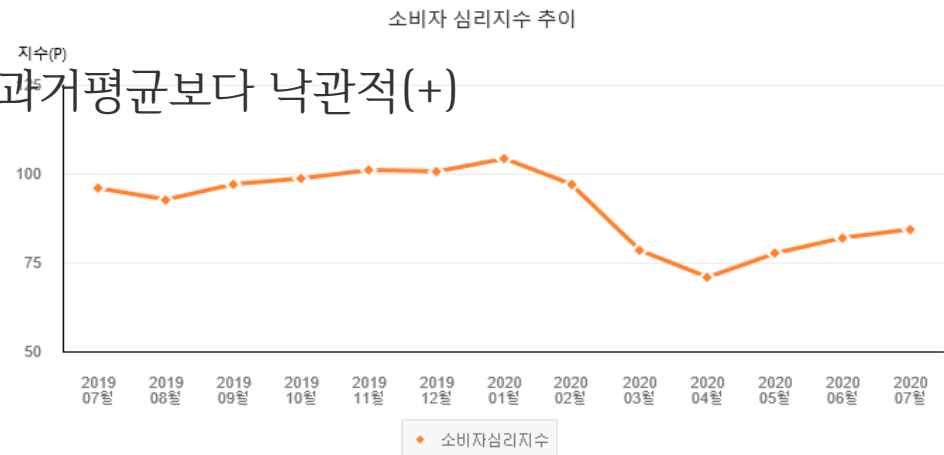
# TO DO - summary

1. 빅카인즈 키워드 카운팅
2. 트위터 크롤링 → 감정분석
3. 1번과 2번 결과를 소비자심리지수보다 더 선행지표로 활용

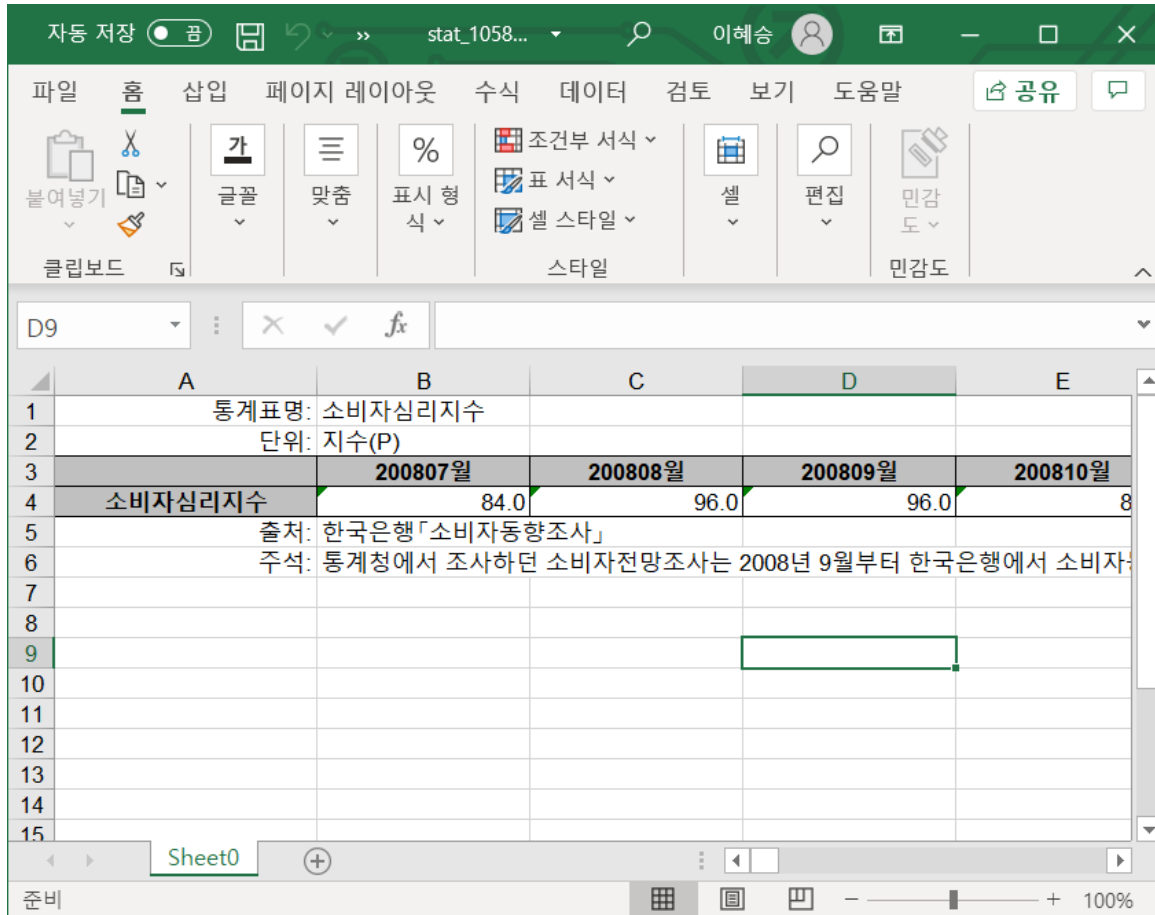
사실 키워드를 우리가 직접 찾아봐야하는게 웃긴..상황이라서,  
그게 얼마나 영향을 미치나? 보다 시간 선행이 얼마나 되나?를 알아보는 거에 중점을 맞추면 어떨까 함!

# 소비자심리지수란?

- CCSI·consumer composite sentiment index
- WHY important?
  - 소비자 심리가 경제에 반영되면서 경기의 방향을 바꿀 수 있음
  - 많은 사람들이 불황이라 느끼고 소비를 줄이면 경기가 침체될 가능성
  - 알고 싶은 것. 소비자(국민)들이 과연 불황이라 느끼는 것에 뉴스들이 얼마나 영향을 미칠지? 뉴스의 '경기 침체'라는 word count를 통해 알아보고자 함.
- 현재생활형편, 가계수입전망, 소비지출전망 등 6개의 주요 개별지수를 표준화하여 합성한 지수.
- 전반적인 소비자심리를 종합적으로 판단
- 지수 > 100: 경제상황에 대한 소비자의 주관적인 기대심리가 과거평균보다 낙관적(+)
- 지수 < 100: 비관적(-)



# 소비자심리지수란?



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E
1	통계표명: 소비자심리지수				
2	단위: 지수(P)				
3		200807월	200808월	200809월	200810월
4	소비자심리지수	84.0	96.0	96.0	84.0
5	출처: 한국은행「소비자동향조사」				
6	주석: 통계청에서 조사하던 소비자전망조사는 2008년 9월부터 한국은행에서 소비자-				
7					
8					
9					
10					
11					
12					
13					
14					
15					

- 통계표 > 시계열조회 > 기간선택 후, 엑셀테이블로 다운로드 가능
- API로 혹시 다운로드 가능한지?
  - Maybe..안되는 듯... 그냥 다운로드 받아서 사용해야 할 것 같다!
  - 더 알아봐야 할 듯

[http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx\\_cd=1058](http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1058)

# 소비자심리지수란?

## <알고 싶은 것>

1. 성장 = GDP성장  
가계 - 소비 / 기업 - 투자 / 정부 - 재정지출 / 해외 - 수출입  
가계 WHY important? 개인의 소비 증가 → 수요 증가 → 기업의 투자로 이어질 수 있음.  
GDP성장이 있을지, 가계의 소비, 개인의 심리지수를 미리 **선행하여** 알아보고 싶음. (선행지수)
2. 소비자 심리지수의 그 밑단에 있는 것을 알아보고 싶음(by 뉴스 카운팅, 트위터)
3. 어떤 뉴스나 어떤 키워드들이 이 소비자들에게 심리적 영향을 주는지
4. 키워드 카운팅 수와 소비자 심리지수와 매칭되는지.  
- Moving avg / time lag가 있는지 알아봐야 함!

## <연구목적>

**Time lag에 초점. 얼마나 더 선행되는 지를 알아보고자 한다.**

1. Chain reaction 이 일어나는 지 검증 (연구자체의 유효성을 증명)
2. **트위터, 뉴스가 얼마나 더 선행되는지** 확인해서, 실제 예측에 활용할 수 있게.(투자활동, 정부정책 제언 등..)
  - Time lag이 일정하진 않을 듯 → 평균과 표준편차를 구해보자.
  - 트위터, 뉴스가 소비자심리지수보다 더 선행  
→ 소비자 심리지수가 (최종적으로) 성장률gdp을 예측.  
→ 추가에도 영향을 미칠 것임.
3. (future work) Industry별로, 여행 / 가전 / 화장품 등등 경기에 민감한 제품들. 섹터별로 맞춰서 해보면 어떨까? 키워드를 분석 → 회사실적/ 추가에 어떻게 영향을 주는지???



〈부록 3〉

BSI/CSI 및 긍정/부정 어휘사전 예시

〈3-1〉 BSI/CSI 사전 예시

구 분	단 어
BSI 사전	국제무역, 계획경제, 신재생에너지, 자금세탁, 상업시설, 무역흑자, 코스닥시장, 시장점유율, 자본잠식, 가공무역, 상장회사, 무역흑자, 무역특화지수, 이머징마켓, 보복조치, 통신업, 농작물, 동맹국, 주문량, 선물옵션, 기간산업, 우대금리, 회계기준, 생산요소, 다국적 기업, 세일요일, 통상협정, 신용거래, 차익거래, 전자상거래
CSI 사전	소비자 보호법, 가스요금, 파트타임, 생활비, 궁핍, 노년, 운전자, 소시민, 청춘, 최저생계비, 출근시간, 기대수명, 가스요금, 임용고시, 양도소득, 잔여재산, 가난, 집세, 아르바이트생, 알뜰, 담뱃값, 인간관계, 노동시간, 교육비, 노점상, 무기계약직, 디딤돌대출, 궁핍, 누진세, 월셋값

〈3-2〉 긍정/부정 사전 단어 예시

구 분	단 어
긍정 사전	메리트, 상쇄, 흥행, 점증, 훈훈하다, 순조롭다, 성공사례, 활약, 대단하다, 순항, 고취, 깔끔하다, 도와주다, 드디어, 고밸류, 업그레이드, 대등, 점령, 밝아지다, 뚜렷하다, 준법, 풀어지다, 선명하다, 안전지대, 호감, 귀재, 첨가, 붐, 건강하다, 꺾찬
부정 사전	단절, 다급하다, 위험수위, 변질, 쪼그라들, 과소, 무분별, 슬프다, 자살, 담보, 불가항력, 구멍, 족쇄, 진퇴양난, 버겁다, 불리, 분주하다, 위태롭다, 불공평, 불충분, 가난, 고비, 당하다, 후유증, 문책, 멸렬, 아슬아슬, 내물렸, 농락, 얕다

# 감정분석 방법

표 1.センチメント 분석 주요 기법

분석기법	설명
Machine Learning Approach	- 사전에 긍정/부정으로 분류된 학습데이터로 텍스트의 긍정/부정 의견을 분류하는 방식 (SVMs(Support Vector Machines)이 주로 쓰임).
Lexicon-based Approach	- 사전에 정의된 긍정/부정 단어를 이용하여 텍스트에 포함된 긍정/부정 단어의 출현 빈도로 긍정과 부정을 판별하는 방식.
Linguistic Approach	- 텍스트의 문법적인 구조를 파악하여 극성을 판별하는 방식.

# 트위터 issue

- 찍어쓰기랑 이모티콘은 어떻게 할 건지?
- 오타, 줄임말?