<발표>

1. 안녕하세요. 데이터 사이언스 9조 프로젝트 발표를 맡은 김성윤입니다.
2. 본격적인 주제 소개에 앞서 가벼운 질문으로 시작하자면, 여러분들은 수업을 들으실 때, 종이 책과 pdf파일 중에 어떤 것을 사용하고 계신가요? 제가 느끼기로는 대부분의 사람들이 종이책보다 pdf파일을 사용하는 것 같습니다. 아는 선배가 이렇게 말합니다. '나 때는 두꺼운 전공책 들고 다니느라 죽는 줄 알았어~'
3. 이렇듯 저희는 디지털 시대로의 전환의 시기에 있습니다.
4. 이러한 전환의 시기에서 종이 책을 출판하는 출판회사들의 상황은 어떨까요? 종이책을 덜 사용하니까 그만큼 심각하지 않을까요?
5. 예상과 다르게 최근 출판시장은 점점 성장하고 있습니다.
6. 세계 혁신 지수(GII)에 따르면 2022-2026년 세계출판시장 규모는 192억 달러의 증가가 전망된다고 합니다. 왜냐하면, 종이책이 줄어들지만, 그만큼 전자책의 전망이 좋기 때문입니다. 이는 총매출액을 통해서 확인할 수 있는데요. 대한출판문화협회의 '2022년 출판시장 통계' 보고서를 보면, 최근 출판회사들의 매출은 전년 대비 2.8%(1396억원) 정도로 소폭 증가했다는 것을 알 수 있습니다.
7. 하지만, 총매출액과 달리 출판시장의 영업이익은 전년 대비 38.7%(1312억원)이나 감소했습니다. 이게 무슨 말이냐 함은 도서의 판매량은 많이 늘었지만, 실제로 출판사가 가져가는 돈은 줄어들었다는 것을 의미합니다. 마케팅과 인건비 등을 그 원인을 생각할 수 있죠. 또한, 총매출액과 관련 있는 판매량 역시 교육출판사만 높게 증가했다는 것입니다. ruff하게 생각하자면, 입시 책 때문에 총매출액이 많아보는 상황인 겁니다.
8. 지금까지 이야기를 정리하자면, 출판회사들은 계속 성장하는 출판시장 속에서 영업이익을 계속 늘려 나아가야 합니다.
9. 그러기 위해 총매출액이라도 높은 교육출판사와 달리 경제와 과학과 같은 지식을 다루는 출판사는 위험한 상황
10. 그러면 어떻게 영업이익을 증가할 것이냐!?
11. 그래서 준비했습니다. '독서 트렌드와 영향 요인 파악', 저희의 목표입니다.
12. 저희 기대효과는 이렇습니다. "독서 트렌드와 영향 요인을 파악하여 자동화 한다면, 이를 마케팅 전략에 활용하여 영업손실에 해당하는 마케팅 비용과 인건비 등을 최소화할 수 있을 것이다." 저희는 이러한 '독서 트렌드'와 '영향 요인'을 파악하기 위해서 다음 세 가지의 데이터를 분석해보았습니다.
13. '도서관 데이터', '서점 데이터', '뉴스 기사 수 데이터'. 독서 트렌드의 영향 요인에는 다양한 요인이 존재합니다. 그 중 '(뉴스 기사로 인해) 특정 소재가 이슈화된다면, 이는 독서 트렌드에 영향을 미칠 것이다.'가 저희의 가설이었습니다.
14. 각각의 데이터는 도서관 정보나루, 알라딘, 구글 뉴스 크롤링을 통해서 얻을 수 있었습니다.
15. 본격적으로 분석 과정을 살펴보면, 크게 3단계로 생각할 수 있습니다. step1 도서관 데이터와 서점 데이터 중 베스트 셀러와 대출 순위가 높은 상위 1000개의 데이터 중에서 '경제'에 해당하는 책의 제목을 받아옵니다. step2 이 제목들 중 tf-idf를 활용해서 키워드를 뽑아냅니다. tf-idf를 활용하면 단어의 상대적인 중요도를 파악할 수 있어서 점수가 높은 순서대로 키워드라고 해석했습니다. step3 이 키워드를 구글 검색창에 검색하고 뉴스 기사의 개수가 몇 개인지를 파악했습니다. 이 과정을 2020-2022년 3년 간 각 월별로 진행했습니다.
16. 그리고 다음은 저희가 얻은 시각화 결과에서 유의미하게 데이터를 선정해보았습니다.

17. <투자> - 대중적인 키워드이기 때문에 항상 높은 퍼센트를 보였다. 대체적으로 tf-idf 점수와 기사 수 퍼센트의 양상이 같은 모습을 보였다. <부동산> - 임대차법 발의 및 의결 시기인 20년 7월 이전으로 높은 기사 수와 점수를 보였다. 해당 법안의 효력이 발휘되는 시점인 21년도 하반기 말부터 다시 기사 수 와 점수가 오르는 모습을 보였다. <주식> - 투자와 같이 대중적인 키워드이기에 거의 높은 수치를 유지하고 있다. 상대적으로 경제 관심도가 떨어지는 연말 연초에는 기사와 점수 모두 하락하는 모습을 보였다. <부자> - 부자라는 키워드는 너무 추상적인 키워드라 뺄까 했지만 경제를 상징하는 단어라고 생각해 같이 분석을 해보았다. 부자 키워드의 특징은 기사 수 퍼센트에 해당하는 점수 시기가 약 두달가량 뒤로 밀린 형태가 나왔다. 키워드가 추상적이라 발생하는 특징인듯 하다. <투자> - 부동산, 주식, 가상화폐 등 경제의 대다수를 차지하는 키워드가 전부 투자 키워드와 관련이 있기에 항상 높은 관심도와 점수를 보였다. 마찬가지로 추상적인 키워드이기에 퍼센트와 점수 간극이 한달 정도 발생하는 모습이다.
18. 그래서 결론을 내리자면 ~ 라고 할 수 있습니다.
19. 이를 통해서 저희의 가설이 ~다는 것을 알 수 있었습니다.
20. 저희에게는 4가지 한계점이 존재합니다. ~ 입니다.
21. 이를 해결하기 위해 다음과 같이 진행할 예정입니다.~ 입니다.
22. 여기까지 저희의 핵심 발표를 맞치고요.
23. 부록으로 팀원별 역할입니다.
24. Q&A