안녕하십니까. 저는 이번 파이널 발표를 맡은__입니다. 저희 7팀은 품목과 가중치를 조정한 개인 맞춤형 체감물가지수 개발을 주제로 프로젝트를 진행하였습니다.

목차는 다음과 같습니다. 먼저 소비자물가지수에 대해 살펴보고, 저희가 개발하고자 하는 개인 맞춤형 체감물가지수를 소개합니다. 프로젝트 마무리에는 저희의 프로젝트가 시사하는 바와 한계점을 짚어보겠습니다.

자, 소비자물가지수란?

소비자물가지수는 통계청이 각 가정에서 생활을 위해 구입하는 상품과, 서비스의 가격 변동을 알아보기 위해 작성하는 통계지수입니다. 소비자물가지수를 구하기 위해서는 아래 3가지 요소, 품목, 가격 그리고 가중치가 사용됩니다.

품목에는 국민들이 일반적으로 주로 사용하는 458개가 존재하며, 가격은 기준년도인 2020년 대비 현재의 가격을 나타냅니다. 가중치는 해당 품목의 상대적 중요도를 나타내며, 가중치가 클수록 소비자물가지수에 더 많이 반영됩니다.

화면에 보이는 그래프는 전체 가중치 총합 1000에서 각 집합에 따라 얼마 정도의 가중치가 할당되어 있는지를 나타내고 있습니다. 가중치에 대해서 한 가지 알아두어야 하는 점은 단순히 가격이 비쌀수록 높아지는 것이 아니라, 우리 삶에 얼마나 더 밀접한지 또한 반영되어 있다는 점, 알아두시면 좋겠습니다.

이러한 소비자물가지수는 정부기관에서 발표하는 공식적인 수치인만큼, 현재 국내에서 가장 공신력있는 물가지수라 할 수 있습

니다. 그 때문에 언론에서 물가를 다룰 때 항상 인용되는 지표이기도 합니다.

하지만 저희는 소비자물가지수가 과연 소비자들이 실제 체감하고 있는 물가지수를 잘 나타내고 있는가에 대해 의문이 들었습니다. 당장 올해 1분기의 소비자물가지수 평균이 113.62인데, 이는 2020년 대비 물가가 겨우 13% 정도밖에 오르지 않았다는 것을 의미합니다. 여러분들도 과연 그렇게 생각하시나요?

통계청에서도 소비자물가지수가 실제 소비자가 체감하는 물가지수와 괴리가 있음을 인정하고 있습니다. 여러 이유가 있겠지만 그 중에서도 저희가 주목한 첫 번째 이유는, 너무 포괄적인 품목들이 포함되어 있다는 점입니다. 예를 들어, 비흡연자의 경우, 담배 가격이 얼마나 인상하던 본인이 체감하는 물가와는 아무런 관련이 없습니다. 마찬가지로, 1인 독신 가구의 경우 미성년자 자녀가 없으니 학원비 물가 역시 본인의 체감물가와는 무관합니다. 이처럼 소비자물가지수에는 각 개개인들과는 무관한 품목들도 모두 포함되어 있으므로, 체감물가와 괴리가 발생할 수 밖에 없습니다.

저희가 주목한 또 한 가지 이유는, 가중치 갱신 주기가 너무 길다는 점입니다. 소비자물가지수의 가중치는 연도의 뒷자리가 0,2,5,7로 끝나는 해에만 갱신이 됩니다. 그런데, 여기 계신 여러분들 모두 2020년 당시의 코로나 사태를 기억하실겁니다. 당시 어딜 가던 마스크는 품절이었기 때문에, 수많은 국민들이 마스크를 구입하기 위해 줄을 서야만 했으며, 심지어는 웃돈을 주고 마스크를 거래하는 일도 발생했습니다.

그런데 가중치는 2020년의 시작과 함께 갱신된 후, 다음 갱신까

지 무려 2년이나 기다려야 합니다. 이 때문에 마스크의 가중치는 이러한 마스크 대란을 전혀 반영하지 못했습니다. 당시 마스크를 쓰지 않으면 집 밖을 나갈 수도 없었음에도 말입니다. 이처럼 소비자물가지수는 가중치 갱신 주기 사이에 발생한 이슈에는 전혀 대응하지 못한다는 문제점이 있습니다.

그래서 저희는 이러한 문제점들을 해소한 개인 맞춤형 체감물가지수를 개발 프로젝트를 계획했습니다.

먼저 개개인의 소비습관에 맞춰 해당되지 않는 품목은 배제합니다. 거기에 추가로, 물가지수 산출 시점의 해당 품목이 얼마나 대중의 주목을 받았는지를 반영하여 가중치를 조정하고자 합니다.

이를 실현한 프로젝트 과정은 다음과 같습니다. 물가지수를 산출하고자 하는 대상자를 상대로 소비습관을 파악하기 위한 설문조사를 실시합니다. 이를 토대로 소비습관과 무관한 품목은 배제합니다. 그 다음으로는 검색 데이터 및 언론 데이터를 수집하여, 각 품목의 주목도를 산출하고, 이를 반영하여 가중치를 갱신합니다. 이렇게 품목과 가중치를 조정한 개인 맞춤형 소비자물가지수를 구하고자 하며, 이번 프로젝트에서는 서로 다른 성별과 세대의 3명을 대상으로 물가지수를 산출하였습니다.

먼저 설문지는 개인의 소비습관 및 세태를 파악할 수 있는 30개의 질문으로 구성하였습니다. 예를 들어, 자녀가 있는지, 술/담배를 하는지, 해외여행을 다니는지, 해산물을 구입하는 지 등을 파악하여 전체 품목들 중에서 대상자와 무관한 품목을 배제합니다.

가령 저희의 표본이었던 20대 남성의 경우 대학생이고, 월셋방에서 자취를 하고 있으며, 자녀가 없고, 술/담배를 좋아하지만, 반려

동물은 기르지 않습니다. 이밖에도 다양한 특징들을 토대로, 대상자와 무관한 상당수의 품목들을 걸러낼 수 있었습니다.

다음은 가중치를 조정하기 위해 수집해야 하는 데이터입니다. 저희는 올해 1월부터 4월까지의 물가지수를 산출하기 위하여, 2022년부터 28개월치의 검색 및 언론 데이터를 수집하였습니다. 참고로 2022년부터 데이터를 수집하는 이유는, 가장 최근에 가중치가 갱신된 해가 바로 2022년이기 때문입니다. 데이터 소스는 검색 데이터의 경우 네이버 데이터랩과 구글 트렌드로부터 수집하였으며, 언론 데이터는 지상파 3사 및 종편 채널에서 방영된 뉴스를 수집하였습니다.

먼저 네이버 데이터랩에서는 화면에 보이는 그래프와 함께 엑셀 파일로 지정한 기간 동안의 키워드 검색량을 제공합니다. 다만 이 검색량은 절대값이 아니라, 수집 기간 중에 고점이 찍힌 시점을 100으로 둔 상대값이 제공됩니다. 따라서 50이 찍힌 지점은, 실제 검색 건수가 50이라는게 아니라, 고점 대비 절반의 검색량이 발생했다는 것을 의미합니다.

구글 트렌드도 이와 정확히 같은 방식으로 데이터를 제공합니다.

또한 구글과 네이버에서는 동음이의어를 구별하여 데이터를 수집할 수 있습니다. 화면처럼 검색창에 사과를 입력하면 아래에 '과일'과 '행위' 중 무슨 의미로 검색을 한 것인지 선택할 수 있는 창이 뜹니다. 이를 이용하면 동음이의어를 구분하여 검색 데이터를 수집할 수 있습니다.

다음은 언론 데이터를 크롤링한 과정입니다. 처음에는 네이버 뉴스에서 제공하는 API를 이용하려 했으나, API 사용과 관련하여

여러 제약들이 존재하는 관계로, 저희가 직접 뷰티풀숲 패키지와 베이스 URL을 이용하여 크롤링하는 방법을 선택하였습니다.

저희가 어떻게 크롤링을 진행하였는지를 간단하게 보여드리기위해서, 베이스 URL을 가져왔습니다. 중요한 부분만 짚어보면요,
URL에서 쿼리값을 수정하여 검색어를 지정하고, photo값을 2로 두어서 동영상이 포함된 뉴스만을 가져옵니다. 참고로 이렇게 2로 지정해야만 지상파와 종편으로 방영된 뉴스들만을 가져올 수 있습니다. 그리고 ds와 de 값을 수정하여 수집할 기간을 설정합니다.

크롤링은 총 28개월을 4개월씩 7분할하여 진행하였습니다. 28개월을 한꺼번에 크롤링하면 사이트로부터 차단을 당하기 때문에, 이렇게 4개월을 사이사이에 두고 슬립을 걸어주었습니다.

이후 크롤링된 뉴스 데이터를 처리하는 과정에서 저희는 여러 시행차오를 겪었는데, 이 과정을 좀 소개드리고자 합니다.

먼저 저희는 챗 GPT API를 사용하여 기사에서 원하는 키워드에 알맞은 내용인지 판별하도록 하는 방법을 진행했었습니다.
하지만 많은 양의 데이터를 처리하며 빈번한 응답 지연이 발생했고, API 사용 비용 문제 등을 고려한 결과, 해당 방법으로 동음이의어들을 처리하기에는 불가능하다고 판단했습니다.

다음으로 시도한 것은 페이스북의 BART 모델입니다.
BART 모델은 문장 간의 관계를 이해하도록 설계되었기 때문에, 문장 내의 단어 의미를 구분하는 것보다는, 두 문장 간의 관계를 판단하는 쪽에 가까운, 동음이의어 구별 작업에 적합하다고 생각했었습니다.

하지만 실행 결과 성능이 기대에 미치지 못해 이 방법도 포기할 수 밖에 없었습니다.

마지막으로 저희는 기사를 전부 영어로 번역한 후, 동음이의어를 구분하는 방법을 사용하였습니다. 다만, 데이터 양이 너무 방대하여 데이터를 줄일 필요가 있었습니다. 그래서 동음이의어가 서로 다른 품사를 갖는 경우, 활용할 수 있는 조사가 달라지는 점을 응용하였습니다. 가령 '사과'의 경우, 과일 사과는 '사과하다', '사과해라', '사과하지' 등으로 사용될 수 없으므로, 이러한 형태로 사용된 기사는 전부 제거할 수 있습니다. 이 방법을 통해 데이터 양을 4분의 1로 줄일 수 있고, 이후 구글 번역 API를 활용하여 분류할 수 있었습니다.

이렇게 동음이의어 처리를 마친 언론 데이터를 토대로 각 키워드에 언론 주목도 점수를 부여하였습니다. 예를 들어, 1월 한 달 동안 사과가 언급된 기사가 다음과 같을 때, 각 기사별로 언급된 횟수에 루트를 씌운 후 모두 더하는 겁니다. 이렇게 하면 특정 키워드가 한 달동안 언론의 주목을 얼마나 받았는지 측정할 수 있습니다. 참고로 루트를 씌운 이유는 특정 기사에서의 과도한 언급 횟수에 패널티를 주어 전체적으로 균형을 맞추기 위함입니다.

이렇게 저희가 수집한 검색량 데이터와 언론 데이터는 프로펫 모델을 이용하여 처리하였습니다. 프로펫은 페이스북에서 만든 시계열 예측 모델입니다. 다만 저희는 프로펫을 이용하여 시계열을 예측하지 않았습니다. 저희가 프로펫을 어떻게 사용하였는지는 예시와 함께 설명드리겠습니다. 먼저 프로펫에 저희가 수집한 28개월치의 쌀 데이터를 모두 학습시킵니다. 그 후 프로펫은 2024년 1월의 쌀 검색량 76에 대하여, 추세 64.06, 주기성 9.01, 오차

항 2.93으로 분리합니다. 여기서, 주기성은 계절성처럼 특정 시기에 반복되는 패턴을 나타내고, 오차항은 설명할 수 없는 돌발 변수를 의미합니다.

그런데 수식을 이항해서 정리해보면, 추세는 실제 값에서 주기성과 오차항을 뺀 값으로 정의할 수 있습니다. 이는 실제 값에서 계절성과 예측 불가능한 변수의 영향을 제거하면 '추세'가 된다는 것을 의미합니다.

그렇다면 실제 값을 추세로 나눌 경우, 계절성과 예측 불가능한 변수가 그들 자신이 발생하지 않았을 때의 값 대비, 얼마 만큼의 영향을 주었는지 계산할 수 있게 됩니다!

이해를 돕기 위해 그래프를 준비했습니다. 그래프의 파란 점은 실제 값이고, 회귀선처럼 보이는 주황색 점은 추세를 나타냅니다. 파란색 점들은 어느 정도의 주기를 반복하며 서서히 내려가고 있는데, 주황색 선이 이 추세를 잘 잡아냈습니다. 파란색 점이 주황색 선보다 위에 있으면 계절의 영향이나 혹은 특수한 사건이 발생하여 평소보다 더 주목을 받았다고 해석할 수 있습니다. 실제로 화면에서 빨간색 선이 그어져 있는 지점은, 삼양사가 밀가루 가격을 1년 사이에 30퍼센트나 인상한 것에 대해 많은 논란이 발생한 시점입니다.

저희가 사용한 방법이 다소 낯설게 여겨질 수 있는 분들을 위해 위해 연구 사례 하나를 가져와봤습니다. 화면에 보이는 뉴스는, 5일 전에 국내 연구진이 지구 온난화가 발생하지 않은 가상의 '메타 지구'를 시뮬레이션하였음을 밝히고 있습니다. 만약 실제 지구에서 태풍이 발생했을 때, 메타 지구에서도 유사한 규모의 태풍이 발생했다면 이는 자연적인 태풍으로 해석할 수 있습니다. 하

지만 메타 지구에 비해 실제 지구에서 더 강한 태풍이 발생했다면, 이는 지구 온난화의 영향때문이라 추측할 수 있다는 겁니다. 저희의 방법도 이와 유사합니다. 추세와 실제 값의 차이가 크다면, 계절성과 돌발 변수의 영향이 크게 작용했다고 볼 수 있는 겁니다.

그리고 앞에서 저희가 수집한 검색량 데이터기 절대적인 값이 아니라는 점을 밝힌 바 있습니다. 화면에 보이는 그래프는 2022년부터 2024년 4월까지의 담배의 검색량인데요, 2023년 11월 3일이 고점으로 100이 찍혀있고, 6월 2일에 46이 찍혀있습니다. 그런데 저희가 데이터 수집기간을 2023년 10월까지로 당겨버리면,

---

6월 2일이 기간 내 최고점이 되어 46이 아니라 100이 된 것을 확인하실 수 있습니다. 이처럼 데이터 수집 기간을 어떻게 정하느냐에 따라 그 값이 변할 수 있다는 문제가 발생합니다.

그런데 앞에서 보여드린대로 실제값 나누기 추세를 계산하게 될 경우, 분자와 분모에 있는 고점의 실제 검색량 값이 서로 상쇄됩니다. 그에 따라, 저희가 구한 값은 시점으로부터 독립된 값을 갖게 됩니다. 실제값을 추세가 아닌 평균이나 중위값으로 나눌 경우 시점에 여전히 종속되기 때문에, 이 부분에서 저희의 방법이 장점을 갖는다고 할 수 있겠습니다.

그렇다면 다음으로 가중치를 갱신하는 과정을 보여드리겠습니다. 예를 들어 2024년 1월 도시가스의 구글 검색량과 네이버 검색량, 언론 주목도를 방금 보여드린 방법을 활용하여 각각의 추세 대 실제값 비를 구합니다. 이 값들을 사전에 구한 점유율로 가중평균하고, 통계청이 정한 도시가스의 기본 가중치와 곱하면 2024년 1월의 주목도를 반영한 새 가중치를 구할 수 있게 됩니다.

참고로 방금의 점유율은 문화체육관광부 여론집중도조사위원회가 내놓은 여론집중도조사 결과와 인터넷 트렌드라는 업체가 제공하는 포털 점유율 데이터를 이용하여 산출하였습니다.

이러한 방법을 토대로 각 품목에 대해 월별로 가중치를 갱신합니다. 몇 가지 사례를 보여드리자면, 1월의 라면은 가중치가 거의 변하지 않았지만, 비빔밥은 오히려 떨어졌습니다. 반면 4월의 고등어 가중치는 2배 가까이 늘어났고, 2월의 수신료 가중치는 기존의 2배를 훌쩍 넘겨버린 것을 확인할 수 있습니다. 참고로 이 시기에는 KBS 수신료 분리징수 관련 이슈가 있었습니다.

이렇게 갱신된 가중치는 국가통계포털에서 가져온 품목별 소비자물가지수와 함께 새로운 가중평균을 구하게 됩니다.

여기서부터는 통계청이 기존 소비자물가지수를 구하는 방법과 완전히 같습니다. 예를 들어 쌀의 가중치가 400, 햄버거가 200, 휘발유가 400이면, 각자 가중치와 가격지수를 곱한 후 가중치의 총합인 1000으로 나누어 가중평균을 구하게 되는 것입니다.

이러한 과정을 통해 저희는 조사대상 3명의 개인 맞춤형 물가지수를 구했습니다. 기간은 2024년 1월부터 4월입니다.
화면의 그래프를 보시면 20대 남성 대상자의 경우 현행 소비자물가지수와 거의 차이가 없지만, 40대 여성은 이보다 조금 높고, 70대 여성은 확연하게 높은 것을 확인할 수 있습니다.
그렇다면 이러한 차이가 발생한 이유는 무엇일까요?

화면에 보이는 품목들은 해당 대상자의 바스켓에만 존재했던 품목들 중 일부입니다. 20대 남성의 바스켓에 포함된 담배와 대중

교통은 최근 가격이 계속 동결되었던 품목인데 반해, 40대 여성의 과일이나 항공료의 경우 최근에 계속해서 인플레이션이 발생해왔던 품목들입니다. 마지막으로 70대 여성의 경우, 물가 인상폭이 넓은 보철치료, 수산물, 화초 등의 품목들이 포함되며 체감물가 상승을 견인했습니다. 이처럼 개인의 소비습관에 따라 실제 체감하는 물가가 상당히 달라질 수 있으며, 현행 소비자물가지수는 이를 제대로 반영하지 못하고 있음을 보여드렸습니다.

마지막으로 저희 프로젝트의 시사점과 한계점에 대해 짚어보겠습니다.

먼저 첫 번째는 현행 소비자물가지수의 한계점을 인식하고, 최신의 개인 맞춤형 물가지수를 내놓으려는 시도는 저희가 처음이라는 점입니다. 기존 소비자물가지수의 계산식을 수정하여 특정 집단에 더 적합한 물가지수를 내놓으려는 연구는 많았지만, 저희의 프로젝트처럼 각 품목의 주목도를 반영하여 매월마다 가중치를 갱신하고, 물가지수를 개인 맞춤형으로 내놓겠다는 시도는 기존에 없었음을 확인하였습니다.

두 번째는 같은 인플레이션 상황에서도 특정 계층에게 더 큰 영향이 갈 수 있음을 보였다는 점입니다. 앞의 그래프에서 개인의 소비습관에 따라 체감물가가 달라짐을 확인하였고, 그 중에서도 특정 집단이나 계층이 주로 사용하는 품목의 가격이 급증하면 이들은 소비자물가지수와의 괴리가 심화될 수 있습니다.

마지막으로 향후 금융업계의 마이데이터 서비스와 연계하면, 전 국민을 대상으로 개인 맞춤형 체감 물가지수를 제공하는 것도 가능하다는 점입니다.

예를 들어 KB 페이나 네이버페이, 토스와 같은 어플 등에서 결제내역이나 소비습관 데이터를 제공받으면, 별도의 설문조사를 진

행하지 않고도 서비스 이용 고객들을 대상으로 손쉽게 개인 맞춤형 물가지수를 제공할 수 있습니다. 또한, 추가로 특정 품목을 더 자주 구입하는 경우, 이에 더 세부적으로 가중치를 조정하는 것도 가능할 것입니다.

다음은 저희 프로젝트의 한계점입니다. 첫째, 원래부터 소비자물가지수 산출 과정에서 누락된 품목들을 저희 역시 새롭게 추가할 수는 없었다는 점입니다. 왜냐하면, 저희의 프로젝트는 통계청이 제공하는 가격지수와 가중치를 가공하여 새로운 물가지수를 만드는 것이기 때문에, 통계청이 이를 제공하지 않으면 저희 역시 이들 품목을 반영할 수 없다는 문제점이 있습니다.
두 번째는 개개인마다 각 품목별로 느끼는 중요성의 차이도 모두 다른데, 이를 가중치에 담아내지는 못했다는 점입니다. 가령 매일 사과를 하나씩 먹는 사람은 일주일에 한 번씩 먹는 사람보다 사과의 가중치가 높아야 체감물가를 더 잘 반영할 수 있을 것입니다. 하지만 저희의 설문조사는 이러한 경향을 담아낼 수 없다는 점에서 명확한 한계를 가집니다.

이상으로 발표를 마무리하며, 질문을 받도록 하겠습니다.