zum internet

zum Research Lab

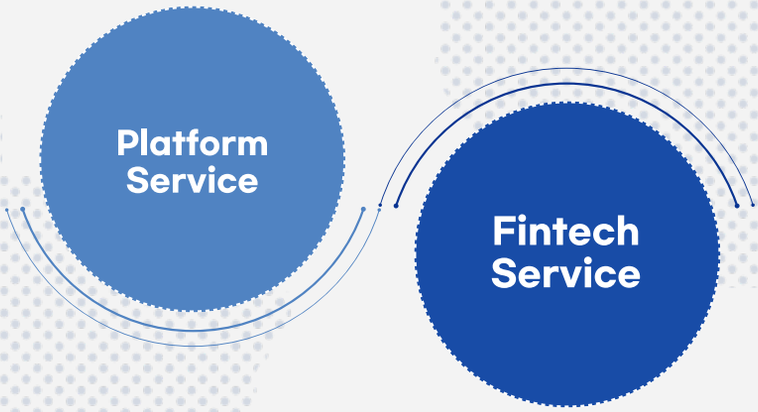
금융과 콘텐츠 경험의
판을 바꾸다

Contents

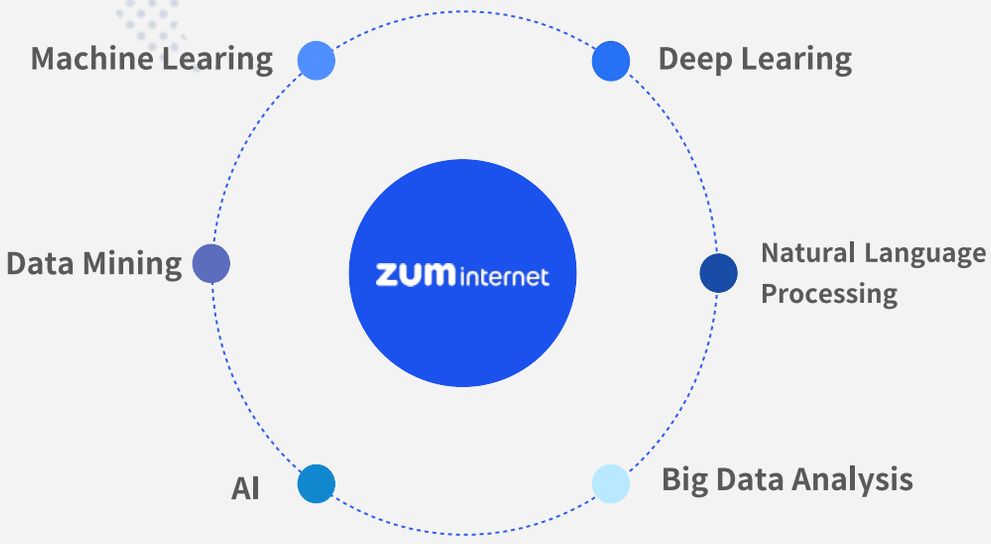
zum internet

줌인터넷은 포털 서비스 'ZUM'을 10년 이상 제공해 오면서 사용자 경험 설계, 빅데이터 처리 및 분석 능력 등에 있어 높은 역량을 보유하게 되었습니다. 특히 줌인터넷의 AI 추천 기술은 맞춤형 광고 및 콘텐츠 제공에 특화되어 있습니다.

줌인터넷은 멈추지 않고 도전합니다. 다양한 금융 정보와 투자 기회 제공을 위한 플랫폼 기업으로도 성장하고 있습니다. 선도 포털 사업자로서의 역량을 기반으로 'ZUM 투자', 'GET STOCK' 'Investing View' 등 고객 가치 증대를 위한 혁신적인 금융 서비스를 제공하고 있습니다.



줌 부설 연구소는 **AI를 바탕으로 한 최고의 서비스 기업**을 목표로
Search, Finance, Language 등 다양한 분야의 AI 기술을 연구하고 있습니다.



Case Study

Portal Service

셀럽NOW

SANS (Safe form Ausive and Nasty Searches)

가짜 뉴스 판별 서비스

DMP

DeepCat

셀럽 NOW

셀럽 NOW는 유명 연예인들의 **뉴스와 사진을 고품질 이미지 순 혹은 타임라인 순**으로 제공하는 서비스입니다.

셀럽 NOW는 포털 줌닷컴에서 유명 연예인의 최근 근황을 사진으로 묶어 보여주는 이미지 검색 서비스입니다. 줌닷컴 검색창에 유명 연예인을 입력하면 '셀럽Now' 섹션에서 촬영된 시간을 기준으로 검색어와 관련된 사진을 보여줍니다. 또한, 사진은 단순히 촬영 시간으로만 나열되지 않고 사진에 대한 간단한 설명과 함께 상황별 해시태그로 구분되어 관심 있는 연예인(셀럽)의 근황을 일일이 찾아보지 않고도 사진으로 한눈에 살펴볼 수 있습니다.

셀럽NOW 서비스는 줌인터넷이 자체 보유한 딥러닝 기반 이미지 분석 기술과 자연어 처리 (NLP, Natural Language Processing) 문서 분석 기술이 융합된 AI 서비스로 개발되었습니다. 해당 기술을 통해 사진이 촬영된 상황의 키워드를 자동으로 분류하고 관련된 고품질의 연예인 사진을 검색 결과로 제공합니다.

이러한 AI 기술을 활용해 문서 이해와 이미지 분석을 위한 딥러닝 기술을 지속해서 연구 개발하고 있습니다.

Service UI

< '셀럽 NOW'의 실제 서비스 화면 >

SANS Safe from Abusive and Nasty Searches

SANS는 머신러닝을 이용하여 성적으로 **유해한 어휘를 자동으로 인식하여 필터**하는 기능으로, 효과적인 포털 운영에 기여하는 서비스입니다.

기존에는 포털 서비스 ZUM의 안전한 이용 환경 조성을 위해 서비스 운영 담당자가 직접 유해한 검색어를 차단해왔습니다. 그러나 많은 수의 유해 검색어를 직접 모니터링하기 위해서는 많은 시간적 비용이 발생하며 유해 검색어를 차단하기 전에 사용자에게 노출될 수 있다는 한계가 존재했습니다. 이 한계를 해소하기 위해 줌인터넷 부설연구소는 자연어처리 기술을 이용하여 유해한 어휘를 인식하고 조정하는 자동화 시스템을 고안했습니다.

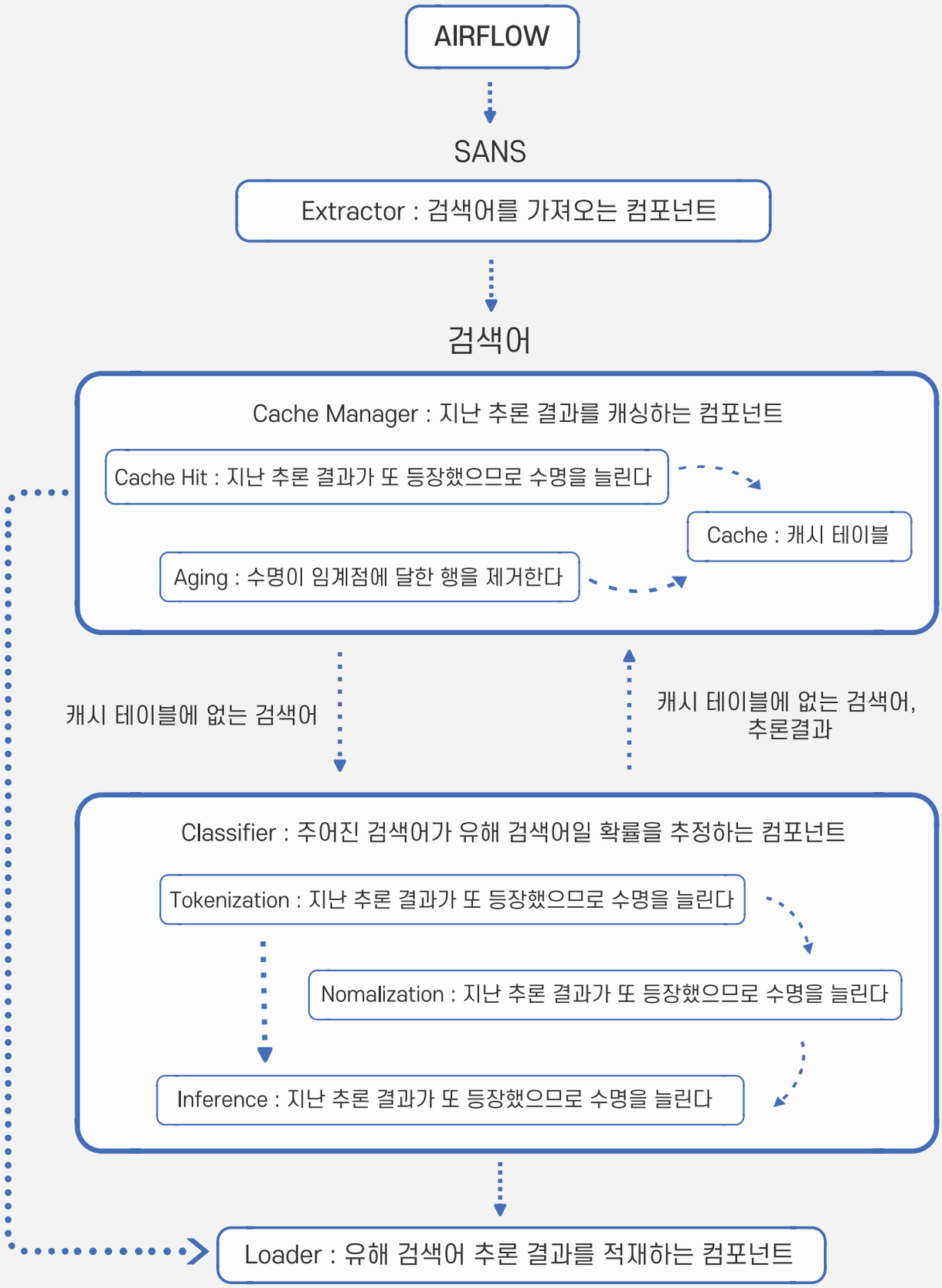
Research Process

SANS의 연구 개발 과정에서 많은 수의 유해 검색어에는 셀럽 이름이 등장하며, 특히 남성보다 여성이 훨씬 더 많다는 점을 파악했습니다. 즉, 셀럽 이름과 성별은 유해 검색어 인식에 결정적인 역할을 합니다. 하지만 현재에도 새로운 셀럽이 꾸준히 등장하고 변화하고 있기 때문에 매번 같은 속도로 새로운 모델을 배포하는 것은 비효율적이라고 판단했습니다. 이런 비효율성을 개선하기 위해 셀럽 이름을 그 셀럽의 성별로 바꾸기로 결정했습니다.

이를 위해서 토큰화 단위는 형태소로 하되, 줌인터넷의 데이터 웨어하우스에 있는 셀럽 이름의 목록을 형태소 사전에 추가했습니다. 물론 인터넷에 등장하는 유해한 어휘에는 필터링을 피하기 위해, 단어를 음절 혹은 음소 단위로 변형을 가하거나 한국어를 영타로 변형하는 등 다양한 패턴이 발생하므로 형태소 단위의 토큰화가 부적절할 수 있습니다. 그러나 SAN의 유해검색어 조정 자동화 영역은 검색어 자동 완성과 연관 검색어 영역이기 때문에 해당 영역에서는 앞서 말한 필터링 회피 패턴의 빈도수가 매우 낮습니다. 따라서, 형태소 단위의 토큰화는 등장하는 자연어를 의미있는 단위로 충분히 분해할 수 있을 것이라고 판단했습니다.

하지만 셀럽 이름을 성별로 바꾸기만 해서는 셀럽 이름이 불러올 수도 있는 중의성을 해소할 수 없다고 판단했습니다. 따라서 셀럽 이름을 추가하기 전의 형태소 사전에 이미 동일한 표층형을 가진 어휘가 존재하는 경우에 주목했습니다. 그 결과, 셀럽의 성별로 바꾸기 전과 후 두 검색어를 Test-Time Augmentation에 사용해서 해당 중의성을 해소하고자 했습니다.

마지막으로 검색어는 보통 그 길이가 짧으며 한국어는 복합어와 교착어의 특징을 가지고 있습니다. 그래서 Transformer보다 훨씬 작으면서 character n-gram 단위로 임베딩을 학습해서 Word2Vec보다 복합어와 교착어에 상대적으로 더 강건한 fastText를 선택했습니다.



< 'SANS'의 모델 구현 과정 >

가짜뉴스 판별 서비스 Fake News Detection

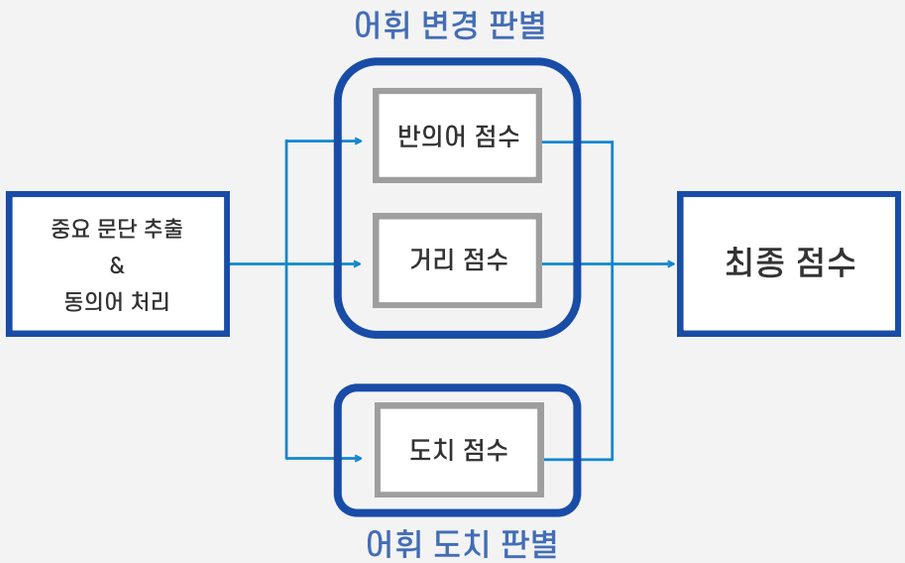
뉴스 제목과 본문이 연관성이 없는 가짜뉴스를 판별하여 뉴스 소비자의 의사결정에 도움을 주는 서비스입니다.

'가짜뉴스 판별 서비스'는 본문 내용과 관련이 없으며 조작된 정보를 기사 제목을 포함하는 가짜뉴스를 판별하는 서비스입니다. 해당 서비스는 뉴스 기사에서 가장 중요하다고 인식되는 중요 문단을 식별한 후, 중요 문단에서 스팸 여부를 판별할 점수 3개(반의어, 거리, 도치 점수)를 산출해내고 통합된 점수를 바탕으로 스팸 여부를 판별하는 구조입니다.

Research Process

우선 제목에 포함된 단어와 본문 각 문단에 나타난 단어의 개수를 단순 비교하여 가장 많은 단어를 포함한 문단을 '중요 문단'으로 선별합니다. 그리고 제목과 중요 문단을 비교해, 제목에는 나와 있지만 중요 문단에는 나와 있지 않은 단어를 선별합니다. 이런 '중요 단어'가 많을수록 해당 뉴스가 가짜뉴스일 확률이 올라갑니다.

중요 단어를 추출한 후, 세 가지의 점수를 추출하는 과정을 거치게 됩니다.



< 제목과 본문이 다른 가짜뉴스 탐색의 도식도 >

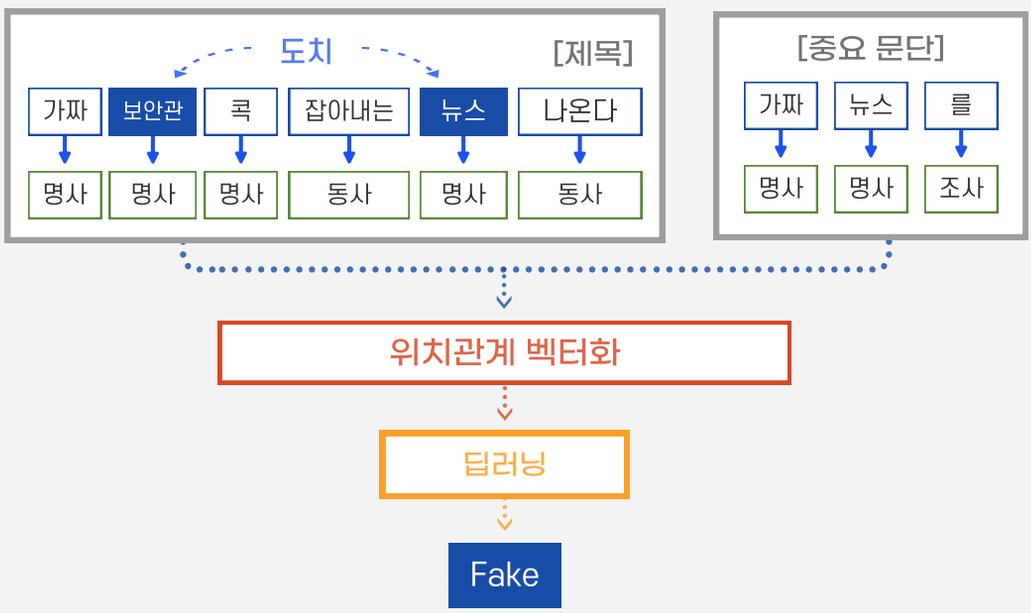
첫 번째 과정인 '반의어 점수'는 Word2Vector와 반의어 임베딩, 두 가지 요소를 사용하여 계산합니다. Word2Vector는 단어 그대로를 벡터화시킨 지표로, 비슷하게 사용되는 용어들끼리 가깝게 위치하게 됩니다. 이러한 Word2Vector 모델을 통해서 중요 단어와 중요 문단의 단어가 반의어 관계일 때 높은 점수로 측정되어 가짜뉴스일 확률이 증가합니다. 단, 반의어 임베딩을 통해서 이미 단어 관계들의 반의어 관계를 사전에 학습시켜 두었습니다.

두 번째 과정인 '거리 점수'는 제목에 포함된 단어 간 거리를 점수로 사용합니다. 각 단어 사이에 몇 개의 단어가 배치되었는지를 거리로 수치화하여 해당 수치에 비례한 점수를 부여합니다. 따라서 이 점수가 올라갈수록 중요 문단에 없고 제목에만 임의로 삽입한 단어가 많은 것으로 인지되어 가짜뉴스 확률이 증가하게 됩니다.

< 거리 점수 >

마지막으로 '도치 점수'는 단어들의 상대적 위치 관계를 벡터화한 후, 임의의 두 명사를 도치한 학습 데이터를 이용하여 제목에 있는 단어가 도치되어 나타날 확률을 점수로 계산합니다. 해당 점수가 높게 나오면 제목에 추가적으로 삽입한 단어는 없지만, 순서를 꼬아서 본문 내용과는 다른 제목을 선정하였다고 판단합니다.

위와 같은 세 단계를 거치면서 받은 점수를 통합하여, 점수가 높다면 해당 뉴스가 가짜뉴스라고 판단하게 됩니다. 이는 모두 제목과 중요 문단의 내용이 다른 경우에 가짜뉴스로 판별한다는 점에서 같은 대원칙을 공유하고 있습니다.



< 도치 점수 >

DMP Data Management Platform

DMP(Data Management Platform)는 사용자별 광고 타겟팅을 위하여 줌닷컴 로그 기반으로 **사용자의 관심사를 분류하는 시스템**입니다.

DMP(Data Management Platform)란 다양한 소스에 산재해 있는 데이터를 수집하여 체계화하고 활성화한 뒤 사용할 수 있는 형태로 가공하는 플랫폼입니다. DMP는 많은 유형의 데이터 수집 및 관리가 가능하지만 통상적으로 개인 식별이 되지 않는 정보를 다룹니다.

줌인터넷은 줌닷컴 로그 기반으로 사용자의 관심사를 분류하여 사용자별 광고를 타겟팅합니다. 그와 동시에 분석한 사용자 행동 로그를 반영하여 사용자에게 특화된 콘텐츠와 경험을 제공하고 있습니다.

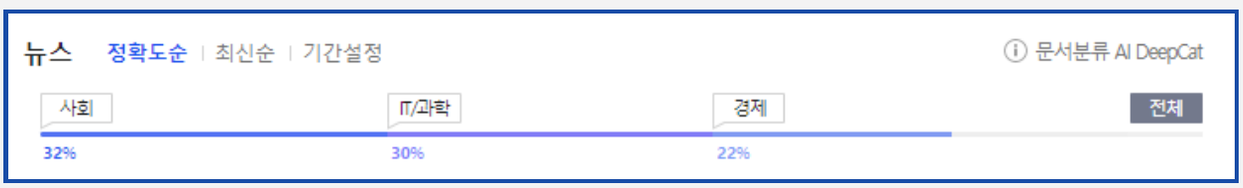
DMP 서비스 개발은 사용자의 다양한 기록을 분석하는 것부터 시작했습니다. 수집된 줌닷컴의 사용자 로그를 토대로 DataMart를 형성했습니다. 이러한 DataMart를 기반으로 Text Classification 모델을 생성하여 각종 카테고리로 분류하여 데이터를 가공했습니다. 그리고 KOBART 모델을 사용하여 사용자별 광고를 타겟팅하기 위해 가공한 데이터를 이용합니다. 해당 결과값은 카테고리별로 점수화하여 저장되며 사용자에게 특화된 콘텐츠와 경험으로 제공되고 있습니다.

DeepCat

DeepCat은 뉴스, 블로그 문서를 주제별로 분류하여 모든 검색 결과들을 정확하고 다양하게 보여주는 문서 분류 AI 서비스입니다.

DeepCat은 카테고리에 분류된 문서 비율을 그래프로 표현하여 주제를 직관적으로 파악할 수 있습니다. 처리 속도가 빠른 카운트 기반의 Bow(Bag of Words) 및 단어를 벡터로 바꿔주는 알고리즘인 Word2Vec 방식으로, 단어를 임베딩한 후 문서를 분류하는 원리로 개발되었습니다.

해당 서비스는 사용자가 다양한 관심 분야의 검색 결과를 바로 확인할 수 있도록 딥러닝을 활용한 문서 분류 AI 엔진입니다. 또한, 입력한 검색어에 대해 다양한 주제별 문서 비율을 직관적으로 그래프 형태로 표현합니다. 그 결과 각 문서에 카테고리를 태깅(Tagging)하여 문서가 주제별로 분류되고 있다는 점 또한 한눈에 확인할 수 있습니다.



< DeepCat의 실제 서비스 화면 >

Case Study

Fintech Service

AweSum News

뉴스 추천 서비스

호·악재판별 서비스

댓글 감성 분석

AweSum News

AweSum News는 인공지능(AI)을 통해 **기사 본문을 자동으로 요약하는 '세 줄 요약'** 서비스입니다.

AweSum News는 기사 요약 AI 기술이 국내 언론에 최초로 도입된 사례로, 현재 연합뉴스 홈페이지에서 확인할 수 있습니다. 해당 연구는 자연어 처리(NLP) 분야 중 하나인 문서 요약 연구로 볼 수 있습니다. 문서 요약은 전체 문서에 포함된 글자와 문장들을 분석하여 요약문과 같이 글의 특징(Feature)을 뽑아내는 시스템입니다.

요약 알고리즘 개발에는 구글이 도입한 자연어처리 딥러닝 언어모델을 사용했으며, 알고리즘이 다량의 데이터를 읽고 스스로 학습하며 사람의 단어, 문맥 이해 방식을 모사하는 자연어 처리 기술입니다. 사람이 요약한 자료를 알고리즘에 사전 학습시켜 AI 스스로 긴 기사를 짧은 문장으로 요약할 수 있도록 했습니다. 또한, AI가 요약한 결과물은 점수로 환산하여 사람이 다시 재평가하는 과정을 통해 AI 학습 알고리즘을 정교화했습니다. 이렇게 완성된 요약 알고리즘은 API(응용프로그래밍 인터페이스)로 체계화해 손쉽게 관리, 활용할 수 있도록 했습니다.

해당 서비스는 추후 줌인터넷에서 제공하는 투자 서비스에 맞춰 AI 요약 알고리즘을 고도화하고, 요약 기술이 필요한 각종 보도자료, 금융 서비스에 적극적으로 활용될 예정입니다. 이러한 서비스는 사용자가 종목별 투자 동향을 파악하는 데에 유용하며, 각종 투자 관련 이슈를 요약하여 사용자의 투자 판단에 도움을 줄 수 있습니다.

Service UI

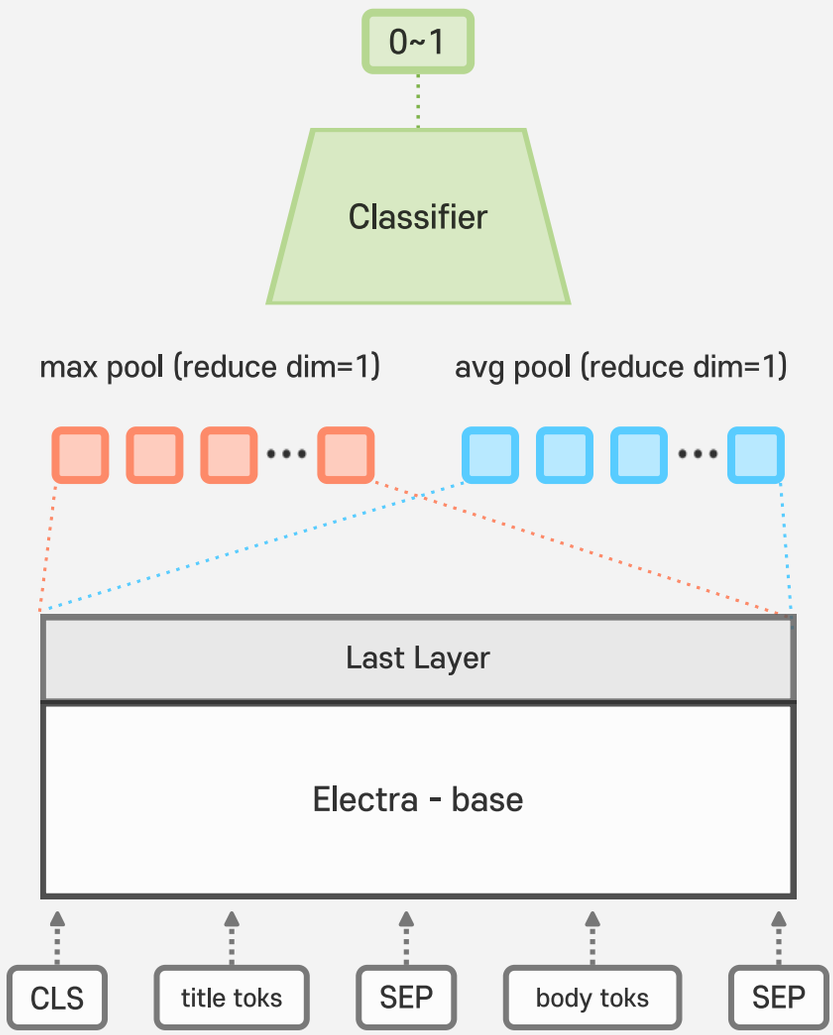
< AweSum News의 실제 서비스 화면 >

Research Process

문서 요약(text summarization)은 크게 추출적 방식(extractive approach)과 추상적 방식(abstractive approach)으로 접근합니다. 추출적 방식은 텍스트로부터 요약에 해당하는 문장 혹은 문장의 위치를 추출하는 방식으로, 추출된 문장을 순서에 맞게 이어 붙여 간단하게 요약문을 생성해 낼 수 있습니다.

뉴스는 블로그나 다른 웹문서에 비해서 문법적 오류가 적고 본문의 경우에는 중복이 적으며 전문 기자가 육하원칙을 기반으로 형식을 갖춘 형태를 따르고 있습니다. 또한, 기사 제목은 본문 내용의 주제를 대표하는 등의 특징을 가지고 있습니다. 따라서 뉴스를 요약할 때는 문장 단위의 추출적 방식으로 제목과 본문의 문장 사이의 관계를 고려하면 그 문장이 제목에서 다루는 토픽을 얼마나 잘 표현하는지 추정할 수 있으며 이를 바탕으로 요약문을 구성할 수 있을 것이라고 판단했습니다.

이러한 과정을 거쳐, 제목 토큰과 문장 토큰을 모델의 입력으로 넣었을 때 그 결과를 0과 1 사이의 값(확률)으로 출력하도록 학습시켰으며, 이때 사용된 프리트레인 모델로는 구글의 ELECTRA입니다. 이를 기반으로 뉴스 본문을 구성하는 문장들에 대한 각 확률(스코어)을 얻고, 확률이 가장 높은 순으로 k개의 문장을 추출함으로써 요약문이 구성되도록 설계하였습니다.



< AweSum News의 모델 구현 과정 >

뉴스 추천 서비스

뉴스 추천 서비스는 사용자의 관심 분야를 파악하여
사용자에게 최적화된 뉴스 콘텐츠를 추천해주는 서비스입니다.

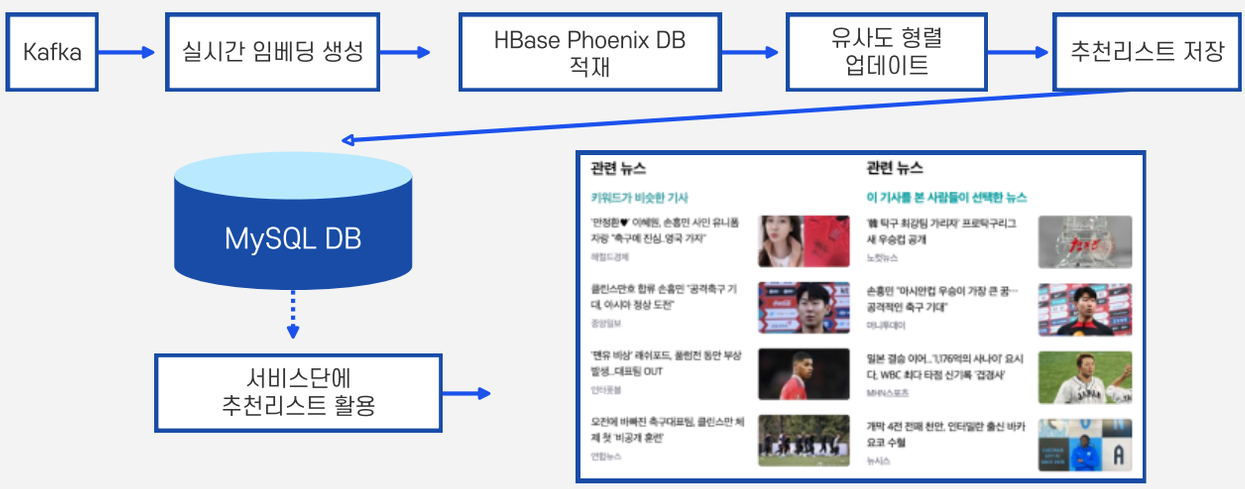
뉴스 추천 서비스는 실시간으로 유입되는 사용자 로그를 통해 사용자에게 최적화된 뉴스, 종목, 투자 정보 등 다양한 콘텐츠를 제공하는 서비스입니다. 줌 서비스 도메인으로부터 로그 데이터 현황 파악 및 추천 알고리즘 연구를 진행했고, 콘텐츠 기반 추천(CBR)을 통해 사용자가 소비한 뉴스 콘텐츠와 유사한 뉴스를 추천해주는 로직을 구현했습니다. 또한 아이템 기반 협업 필터링(CF)를 바탕으로 사용자의 콘텐츠 소비 습관을 분석하여 개인화 추천 콘텐츠를 제공하고 있습니다. 뉴스 추천 외에도 줌인터넷이 제공하는 다양한 투자 콘텐츠에 해당 추천 서비스를 확대·적용할 예정입니다.

Research Process

뉴스 추천 서비스는 Content-Based와 개인화 추천으로 나눌 수 있습니다.

우선, Content-Based 추천은 실시간으로 생성되는 뉴스 데이터를 딥러닝 NLP 모델, SentenceBERT에 임베딩하는 모델입니다. 뉴스를 임베딩하면 해당 뉴스를 숫자로 변환한 벡터값을 얻을 수 있으며, 이러한 벡터값을 기반으로 뉴스 간 유사도를 계산하는 동시에 유사한 뉴스를 클러스터링(clustering)할 수 있습니다.

그러나 Content-Based 모델만 사용할 경우 유사한 주제의 뉴스 기사들만 추천할 수 있다는 한계점이 존재했습니다. 이를 보완하기 위해 Collaborative-Filtering 알고리즘을 사용하여 폭넓은 뉴스 기사를 추천할 수 있게 되었습니다. Collaborative-Filtering 알고리즘은 개인화 추천의 방식으로, 사용자가 활동한 로그를 바탕으로 사용자와 취향이 유사한 유저들이 많이 소비한 뉴스를 추천하고 있습니다. 또한, 특정 뉴스와 함께 소비되는 뉴스 기사를 추천할 수 있습니다.



< 뉴스 추천 서비스의 모델 구현 과정 >

호·악재 판별 서비스

호·악재 판별 서비스는 AI 기술을 통해 **뉴스 기사의 호악재 여부**를 추론하여 제공하는 서비스입니다.

호·악재 판별 서비스는 특정 종목에 대한 뉴스가 해당 종목에 대한 호·악재 점수를 추론하여 호재인지 악재인지에 대한 여부를 제공하는 서비스입니다. 해당 서비스는 ChatGPT와 같은 뿌리를 두고 있는 딥러닝 NLP 모델 BERT를 줌인터넷의 방대한 경제 뉴스 데이터로 Fine-Tunning 해서 제작한 모델입니다.

호·악재 판별 서비스 모델은 문장에 대한 전체적인 이해와 문맥 파악 측면에서 큰 장점을 가지고 있으며 종목별 줌 AI의 한마디 요약, 긴급 알람 PUSH 서비스, 줌 AI의 추천 종목 등 다양한 서비스에 활용하여 사용자에게 투자 동향을 제공, 투자 판단에 도움을 줄 수 있습니다.

< 호·악재 판별 서비스의 실제 서비스 화면 >

댓글 감성 분석

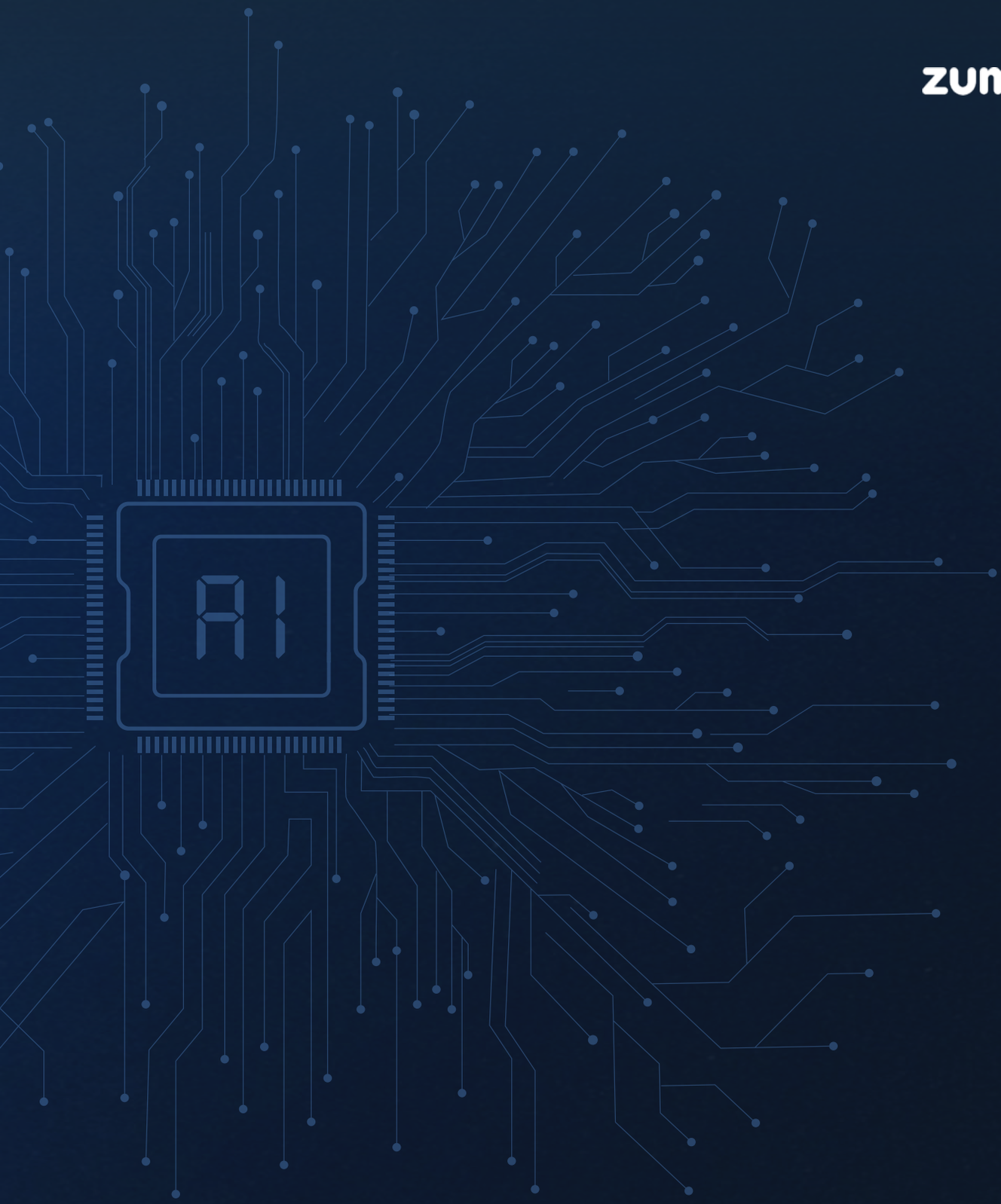
댓글 감성 분석 서비스는 핀테크 콘텐츠 내 사용자 댓글을 감성 분석하여 **사용자의 감성지수(긍정/부정)를 제공하는 서비스**입니다.

댓글 감성 분석 서비스는 사용자 댓글의 감성지수를 산출하여 사용자의 투자 판단에 도움을 주고자 제공되는 서비스입니다.

줌에서 제공하는 핀테크 서비스 내 사용자 댓글을 학습하여 댓글 데이터를 라벨링합니다. 이때 댓글 라벨링은 긍정, 부정, 삭제 대상으로 분류되며 댓글 데이터 중 이모티콘, 영어, 숫자, 특수기호, 띄어쓰기와 같은 요소들을 전처리합니다. 이러한 전처리 과정을 거친 후, KcElectra 모델을 사용하여 긍정과 부정을 추론하는 분류기를 생성하여 최종적인 댓글 감성을 산출합니다.

해당 서비스는 특정 뉴스에 대한 사용자의 댓글 성향을 파악하기에 용이하기에, 사용자가 다른 사용자들의 투자 동향을 쉽게 파악하도록 도울 수 있습니다. 이러한 강점을 살려 줌인터넷에서 제공하는 금융 서비스에 적극적으로 활용할 예정입니다.

zum internet



Case Study

02·583·4640
서울특별시 서초구 반포대로 3, 7층