<주요 Q&A>

Text Analysis 기술을 이용한 VoC 처리 지능화 적용 사례

Q1. 노이즈를 필터링하는 것도 음성분석시 필요한데 노이즈 분석에 대해서는 어떻게 처리하시는지요?

음성인식은 STT(Speech To Text)로서 텍스트로 만들기 이전 단계로 분류되어 또 다른 영역의 기술입니다.

Q2. 감성분석에서 수치는 결국 빅데이터로 분석하는 걸까요?

빅데이터로 학습된 딥러닝 모델에 예측 대상 데이터를 넣으면 감성에 대한 정형화된 수치를 예측해줍니다.

Q3. 감성사전 구축 자동화를 위한 기술도 연구되고 있나요?

어떤 분야이든 사전 구축은 아직까지는 노동집약적인 과정이라 구문 수집과 사전 단어 구축에 많은 시간과 노력이 들어가고 있습니다.

Q4. 텍스트 분석에서 수집된 고객 텍스트 데이터 분석의 정확성을 높이기 위한 고도화 방안은 무엇인지요? 예측 결과를 높이기 위한 텍스트 데이터 사전 정제 작업 등 중요 작업은 무엇인지요?

목적에 따라 크게 전처리, 임베딩, 모델 선택이 중요합니다. 전처리는 토큰화, 불용어 제거, 정규식 적용, 오탈자 제거 등을 필요에 따라 적용합니다. 임베딩은 W2V, FastText 등이 있습니다. 모델은 정말 다양해서 사전학습모델을 기반으로 할 수도 있고, RNN이나 CNN등을 활용할 수도 있고, Random Forest나 Logistic Regression같은 전통적 머신러닝 기법을 사용할 수도 있습니다.

Q5. VoC 분석 모델을 기반으로 서비스 모델 적용된게 무엇이 있나요?

고객사를 명시할 수는 없지만 CRM, 자동응답, VoC 통계분석 서비스가 개발되고 있습니다.

Q6. 키워드 추출이나 자연어 처리를 위한 NLU는 어떤 것을 사용하고 있나요?

목적과 모델의 구동 환경에 따라 여러가지를 시도해보고 가장 성능이 좋은 알고리즘을 적용합니다. 일반적으로는 삼성SDS의 KoreALBERT를 사용하는데, 다른 딥러닝 모델이나 전통적 머신러닝 기법을 쓰기도 합니다.

Q7. 키워드 추출을 통해 현재 빅데이터 수집을 하여 따로 진행중인 프로젝트가 있나요?

발표자료의 사례 4번에 소개됩니다.

Q8. 현재 기술로 문맥을 이해 할 수 있는 단계까지 도달했는지 궁금하고 문맥을 이해하기 위해서는 어떤 기법을 사용하는지요?

2021년 11월 24일 기준, KorQuAD 기계 독해 평가에서 삼성SDS의 모델이 1위를 기록하고 있는데, 이미 사람보다 나은 수준을 나타내고 있습니다. 문맥을 이해하기 위한 기술은 많이 공개되어 있습니다. 요즘 가장 핫한 BERT기반 모델에 도메인 데이터로 fine-tuning만 하더라도 문맥을 이용하여 여러 응용을 할 수 있습니다.

Q9. 토픽 모델링을 할 때에도 유사도 분석을 통한 벡터 확보가 필요했을 것 같은데요. 정확도 높은 결과를 위해서 평균 어느 정도의 학습이 있었는지 궁금해요

높은 정확도를 위해 정해진 학습 데이터 양은 딱히 정해진 것은 아니고, 저희는 최소 1만건 이상이 필요하다고 판단했습니다.

Q10. 해당 내용의 학습 단계에서 워드임베딩 차원수를 어느 정도로 설정했을 때 최적으로 나왔는지 궁금합니다.

어떤 사례에 대한 질문인지 모르겠지만 토크나이저를 어떤 것을 쓰냐에 따라서도 다른데, 시도하는 차원 수는 W2V 기준으로 70~200 입니다.

Q11. 토픽모델링, 키워드 추출 등의 성능을 직접 확인해보고 싶으면 어떤 절차를 진행할 수 있는지 궁금합니다.

토픽모델링과 키워드 추출은 정답과 비교 방법을 정의하는 것이 굉장히 다양하기 때문에 정해진 것은 없고 정성적으로 평가하기도 합니다.
텍스트의 요약이 미리 정해져있다는 가정 하에 ROUGE라는 방법을 사용할 수 있습니다.

Q12. 별도의 fine-tuning을 하지 않고 pretreaing만 된 상태에서 cls vector만 가지고 similarity가 높은 문서를 뽑으면 모종의 공통점을 발견할 수 있을까요? 왠지 시도해보셨을 것 같아서 여쭤봅니다.

대부분의 프로젝트가 성능을 높이는 것이 목적이기 때문에 사전학습 모델을 그대로 사용하는 경우는 없었습니다.

Q13. 한글 텍스트에서 오타나 줄임말(약어)등은 전처리 단계에서 어떻게 처리하는지 궁금합니다. 솔루션이 있는지? 분석가가 개입하여 처리하는지?

오탈자 처리는 프로젝트 여건에 따라 symspell, py-hanspell과 같은 솔루션을 참고하고, 일반적으로 입력 데이터가 많기 때문에 분석가가 직접 개입하기는 어렵습니다.

Q14. 지금 발표하시는 Text Analytics 포함해서 SDS가 제공하는 AI서비스는 모두 클라우드에서 제공하는지요?

클라우드로 제공하는 서비스도 있고, 온프레미스 형태로 제공하는 서비스도 있습니다.

Q15. 기업이 고객 관리를 효율적으로 하려는 경우 Text Analysis 를 활용하는 방안에 대해서 문의 드립니다. 이와 관련하여 중점적으로 고려해야 할 사항들에 대해서 설명해 주세요.

고객 관리도 굉장히 넓은 분야인데, 대내 시스템에서는 입력되는 VoC의 토픽이 무엇인지 빠르게 뽑아내는 것과 기간 별로 등장하는 키워드가 무엇인지 뽑아낼 수 있겠고, 대외 시스템에서는 고객들의 감성 반응을 수집하는 것에 이용할 수 있겠습니다. 분석이 잘 되기 위해서는 양질의 데이터를 누락 없이 저장하는 것이 중요합니다.

Q16. 공공기관에서 활용된 예시도 있을까요

공공 프로젝트에 적용된 사례는 아직 없습니다.

Q17. 텍스트를 인식할 때 사람들이 글을 쓰며 반어적인 문구를 사용하는 경우에 그런 부분들을 찾아낼 수 있는 방법도 있나요?

최근 등장하는 사전학습모델 기반 자연어 처리 기술들은 학습 데이터만 많다면 반어, 역설이 들어가는 문구들까지도 문맥 학습이 가능합니다.

Q18. 키워드 추출시 오류등을 방지하기 위한 적절한 필터링은 어떻게 되는지요?

먼저 불용어 제거, 정규식 적용, 오탈자 제거 등을 하고 토큰화 기법에 따라서도 차이가 납니다.

Q19. 유사도 분석에는 Sentence to Vector 기술이 적용되어있는 건지 아님 워드 임베팅만 활용하는지요?

토큰화를 거친 워드 임베딩을 활용했습니다.

Q20. 서비스를 운영하면서 외부 챗봇서비스를 통해 VOC를 처리하고 있습니다. VoC처리 지능화를 위해서 빅데이터를 구축하려면 기존 외부 챗봇으로는 적용이 불가능할까요? 즉 데이터를 통해 학습시키려면 챗봇을 자체 구축해야 할까요?

들어오는 텍스트는 전부 수집 가능해야 하고, 외부 챗봇 서비스의 응답 모듈에 접근이 가능한지가 관건일 것 같습니다. 응답 모듈에 적절한 추천 응답을 넣는 수준으로만 가능하다면, 자체적으로 학습 모델을 만들고 예상되는 입력에 대한 예측 답변을 넣는 방향으로 가는 것이 어떨까 합니다.

Q21. VoC의 데이터를 효과적으로 분석할 수 있는 방안에 대해서 질문 드립니다. 이를 통하여 데이터를 효율적으로 분석하고 배포하는 방식에 대해서 질문 드립니다.

목적에 따른 양질의 데이터를 확보하는 것이 가장 중요합니다. 분석의 목적이 질문에 드러나진 않지만 효율적인 분석과 배포 방식은 삼성SDS의 Brightics ML을 소개드립니다.

Q22. VoC를 음성으로 접수하여 텍스트로 변환하는 것이 현업적용 시 필요한데 이런 사례도 있을까요?

음성인식은 STT(Speech To Text)로서 텍스트로 만들기 이전 단계로 분류되어 또 다른 영역의 기술입니다.

Q23. 표와 같은 특정양식에 담겨있는 정보를 추출하는 기능도 필요할 것으로 생각되는데, 표의 왼쪽의 구분에 따른 오른쪽의 값을 연관지어 분석할 수 있는 방법은 어느정도 연구되고 있을까요?

테이블(표)의 서로 다른 컬럼(구분)을 연관 짓는 것은 카테고리로 분류할만큼 깊이있는 기술은 아닙니다. 기존 텍스트 분석 결과와 다른 컬럼의 정형화된 값을 매칭하여 머신러닝으로 학습하는 방법이 적절한 예시로 생각됩니다.

Q24. 예를 들어. 제조사에서 운영하는 VOC 채널에 들어온 내용을 판단해서 해당 솔루션으로 연결할 수 있으면 좋을 것 같습니다. (챗봇기능도 연계)

감사합니다. 챗봇에도 텍스트 분석 기술이 응용되어 들어갈 수 있습니다.

Q25. 활용 방안에 대한 질문이 많은 거 같은데 주로 어느 분야(ex. 금융)에서

많이 활용되고 있는지 궁금합니다. 그리고 기술 비교를 위한 대조 상품(or 기술)이 있다면 어딘지 궁금합니다.

적용 현장의 모수가 굉장히 많지는 않아서 분야를 특정하기는 어렵지만, 불특정 다수의 고객을 상대하는 곳이라면 어디든 활용 가능합니다. 응용 분야가 많은 기반 기술이기 때문에 특정 상품이나 기술을 언급하기는 어렵고, 탑재를 예상할만한 챗봇이나 자동응답 서비스를 생각해볼 수 있습니다.

Q26. 간혹 공문서 추천시, 오타나 부적절한 용어를 탐지해서 제거하는 것이 가능한지요? 혹시 학습단계에서 이미 가능성이 없어지는 것인지 궁금합니다.

오탈자 처리를 위해 참고하는 외부 패키지들이 있습니다. 학습 전에 전처리로 걸러내야 합니다. 만약 학습 과정에서 방대한 데이터에 비슷한 오탈자가 많이 있다면 그 조차도 문맥으로 학습되기도 합니다.

Q27. STT 지원하는 언어는 어떻게 되나요?

음성인식은 STT(Speech To Text)로서 텍스트로 만들기 이전 단계로 분류되어 또 다른 영역의 기술입니다.

Q28. 연구개발과정에서 한국어 데이터를 이용했을 때, 특별한 이슈가 있었는지도 궁금합니다.

저희는 대부분 한국어를 대상으로 분석합니다. 토크나이저나 사전학습모델을 사용할 때, 한국어가 지원되는 것으로 해야 하는데 수준 높은 선택지가 많아서 큰 이슈는 없었습니다.

Q29. 텍스트 데이터 분석을 통한 마이닝 시 필연적으로 개인정보 노출/보안에 신경쓰이는 부분이 있으실 텐데요. 이 부분은 어떻게 처리를 하셨는지 궁금합니다.

개인정보는 제거가 원칙입니다. 개인정보가 포함된 성질의 데이터는 잘 사용하지 않고, 데이터에 일부 포함된다 하더라도 정규식 적용과 토크나이징 시 고유명사 제거 등으로 전처리 합니다.

Q30. 수행시간 예측 기반 시스템 적용에 최소 비용과 개발에 걸리는 시간은 최대 얼마만큼 소요가 되는지 궁금합니다

비용과 시간은 모델의 정확도 및 크기에 따라서 천차만별이라서 특정하기가 힘들지만 해외 연구 논문과 적용 기술을 참고 한 결과 AI 모델을 이용해서 정확하게 수행시간을 예측하는 모델이 아직은 없기에 상당한 시간과 비용이 소요된다 라고만 말씀드릴 수 있을 것 같습니다.