자기소개글 필터링 모델

작성일 : 2022.10.17.(월)

작성자: 정보기술연구소 김건희 주임

1. 모델 설명(학습데이터 수)

(1)무성의 필터링 모델

- 학습 데이터 : 정상(32,796개), 무성의(32,380개)로 총 65,179건

(2) 연락처 필터링 모델

- 학습 데이터 : 정상(30,116건), 연락처(24,127건)로 총 54,245건

2. 검증 방법

(1) 2022년 10월 17일 기준 데이터베이스 내에서 여보야 맞선 프로필에 노출되어 있는 회원 자기소개글, 가족소개글로 검증 (79,081개)

- 자기소개글, 가족소개글은 cs에서 검증이 완료된 글로 해당 글은 모두 정상글이라는 가정하에 정확도 판별

(2) 2022년 9월 1일 ~ 2022년 10월 17일 관리자 삭제된 데이터로 검증 (2,681개)

- 삭제된 데이터글은 약 2,500개로 직접 검수하여 정확도 판별함

1. 소개글 모델 검증(22.10.17) - 자기소개글

2022년 10월 17일 월요일 기준 프로필 노출되어 있는
79,087명의 회원 자기소개글 검증

- 총 79,087건 중 **무성의 모델**이 '무성의'로 **총 316건 예측**
  ○ 316개 중 실제 무성의글 38개, 정상글 278개

=> accuracy : 99.6%, precision : 100%, recall :99.6%, f1-score: 99.7%

- 총 79,087건 중 **연락처 모델**이 '연락처'로 **총 76건 예측**
  ○ 76개 중 실제 연락처글 53개, 정상글 13개

=> accuracy : 99.9%, precision :100%, recall:99.9%, f1-score:99.4%

무성의 모델

	실제 정상글	실제 무성의글
정상글로 예측	78,771	**0**
무성의글로 예측	**278**	38

연락처 모델

	실제 정상글	실제 연락처글
정상글로 예측	79,021	**0**
연락처글로 예측	**13**	53

2. 소개글 모델 검증(22.10.17) - 가족소개글

2022년 10월 17일 월요일 기준 프로필 노출되어 있는
79,087명의 회원 가족소개글 검증

- 총 79,087건 중 **무성의 모델**이 '무성의'로 **총 640건 예측**
  ○ 640개 중 실제 무성의글 530개, 정상글 110개

=> accuracy : 99.8%, precision : 99.8%, recall :100%, f1-score: 99.8%

- 총 79,087건 중 **연락처 모델**이 '연락처'로 **총 480건 예측**
  ○ 480개 중 공백데이터 422건 제외 58건으로, 58건만 봤을 때 실제 연락처글 17개, 정상글 41개

=> accuracy : 99.9%, precision :99.9%, recall:100%, f1-score:99.4%

무성의 모델

	실제 정상글	실제 무성의글
정상글로 예측	78,447	**110**
무성의글로 예측	**0**	530

연락처 모델

	실제 정상글	실제 연락처글
정상글로 예측	78,607	**41**
연락처글로 예측	**0**	439

3. 소개글 모델 검증(22.10.17) - 삭제내역

2022년 9월 1일 ~ 10월 17일 월요일 기준 삭제된
2,681명의 회원의 글 확인

- 총 2,681건 중 **무성의 모델**이 '무성의'로 **총 388건 예측**
  - ○ 388개 중 실제 무성의글 372개, 정상글 16개
  - ○ 나머지 2,293개 정상글 중 무성의글 2개

  => accuracy : 99.3%, precision : 99.9%, recall :99.3%, f1-score: 99.6%

- 총 2,681건 중 **연락처 모델**이 '연락처'로 **총 151건 예측**
  - ○ 151개 중 실제 연락처글 150개, 정상글 1개
  - ○ 나머지 정상글 2,530개 연락처 0개

  => accuracy : 99.9%, precision :99.9%, recall:100%, f1-score:99.4%

무성의 모델

	실제 정상글	실제 무성의글
정상글로 예측	2,291	**16**
무성의글로 예측	**2**	372

연락처 모델

	실제 정상글	실제 연락처글
정상글로 예측	2,530	**1**
연락처글로 예측	**0**	150

모델이 잘못 예측한 건

[1] 가장 치명적인 오예측건 (무성의글) case 1

회원 번호 : 1667930
등산캠핑좋아하고활동적이고책임감이강한성격입니다.언제나변함없고,노년에같이여행,캠핑다닐사람을만나고싶습니다.같이라는단어와 함께라는단어가어울리는여성을만나두손잡고잘놀다가노라말할수있는삶을살고싶습니다.다들행복하세요.인연은있겠죠.

회원 번호 : 1330142
안녕하세요 대구달서구에살고있는44살미혼에남자입니다착하고좋은분만나서행복해지고싶습니다많관심부탁드립니다가진건별로없구요 아파트작은거영구임대지만이런저라도괜찬으시면연락주세요전거짓말하고싶지않아요사진없음x제혼x거짓은정말×

회원 번호 : 391293
키160체중57나름대로열심히살고있어요긍정적인사고방식으로요긍정적이고가정정인진실한사람을만나고심어요미래의행복을추구하고이 해심이많은분이면좋겠어요가벼운마남피해가세요진실로노후를보낼분만요자기관리잘되있는분요남은인생행복하게잘살각오가되있분요

회원 번호 : 1691757
제2에인생이새로시작하는시기라고보고대화잘통하는사람기대합니다문자교환하면서만나천천히서로알아갔으면하네요서로소중함을느끼 고꿈꾸며살고싶습니다주변환경보다둘만의사랑이더중요할거같아요음악노래운동좋아하고ㅡ낙천적성격입니다착하단소리많이들었네요항 상일할수있고건강한것이감사하며살고있어요마음맞는인연이있을지궁금합니다^^~~혼자생활하며화물매일출퇴근합니다기독교초보닙다

회원 번호 : 199321
안녕하세요.^^좋은인연찾아요~한번실패는했지만아픔을딛고반쪽만나고싶어요~제아이도아껴주고소중히생각해주실분~진짜진심으로다가 와주실분~장난은사절입니다.천천히알아가고싶어요.마음이따뜻한내반쪽어디계신가요.?좋아하면거리지역상관없이잘와주실분^^아이랑 함께볼수있으신분,나이차이8살까지,거짓말제일싫어해요,아저씨스타일,머리까지신분죄송,키작으신분,외모,조건따지는분,흡연하시는 분no,변태사절(성적),성격급하신분쪽지보내지마세요~!없으면좋겠지만아이가1명까진괜찮아요~!이해심많고배려해주고매너있으신분환 영요※여자분들제발프로필보지마세요※

=> 문장이 길지만 띄어쓰기가 없는 경우 모델이 몇 건은 무성의 글로 잡음

모델이 잘못 예측한 건

[1] 가장 치명적인 오예측건 (무성의글) case 2

회원 번호 : 1736562
안녕하세요광고보고 가입해요 잘 부탁드려요 같이 탈퇴해요

회원 번호 : 1941051
안녕하세요 안녕하세요 ^^ 안녕하세요 그냥 광고보고 가입해봤어요

회원 번호 : 1787318
안녕하세요~~햄스터집사입니다 햄스터는 너무 귀여운거같아요~^^

회원 번호 : 1947075
가입합니다.가입해요 가입했어요 가입했네요 가입을 원행요

회원 번호 : 1892953
안녕하세요. 광고보고 왔어요. 좋은 인연 기대합니다. 하하호호

회원 번호 : 1423300
안녕하십니까~ 좋은 인연 만들어 봐요 잘 부탁 드리겠습니다~♡

회원 번호 : 1953930
안녕하세요. 이런건 처음 입니다 . 잘부탁 드립니다 .

회원 번호 : 1534342
안녕하세요. 광고보고 가입했습니다. 양양에 살구요 가까운 근처면 좋겠네요

⇒ '안녕하세요', '광고', '가입했어요', '잘부탁드립니다' 등
안녕, 광고, 가입 단어만 소개글에 입력되어 있는 경우 무성의글로 예측함

모델이 잘못 예측한 건

[1] 가장 치명적인 오예측건 (무성의글) case 3

회원 번호 : 1163602
망망한인해속에서 ,만나게된 것은,좋은인연이며! 대자연, 신선한, 공기, 아름다운,화,초,목,운동,을 ,좋아합니다 !Joyisthebestofwine.Tolove ,andtobeloved,isthegreatesthappinessofexistence.Toloveitisandlovealonethatlifeorluxuryisknown.

회원 번호 : 1673544
Introverted, driven, friendly, and stubborn.
A free-spirited and emotional personality.
Graduated from Peking University.

회원 번호 : 1389322
I feel like being alone in the world . need someone on my side. I will be on your side of course.

⇒ 긴 영어 문장으로 작성되어 있는 경우 모델이 무성의로 예측함

모델이 잘못 예측한 건

[1] 가장 치명적인 오예측건 (무성의글) case 3

회원 번호 : 1163602
망망한인해속에서 ,만나게된 것은,좋은인연이며! 대자연, 신선한, 공기, 아름다운,화,초,목,운동,을 ,좋아합니다 !Joyisthebestofwine.Tolove ,andtobeloved,isthegreatesthappinessofexistence.Toloveitisandlovealonethatlifeorluxuryisknown.

회원 번호 : 1673544
Introverted, driven, friendly, and stubborn.
A free-spirited and emotional personality.
Graduated from Peking University.

회원 번호 : 1389322
I feel like being alone in the world . need someone on my side. I will be on your side of course.

⇒ 긴 영어 문장으로 작성되어 있는 경우 모델이 무성의로 예측함

모델이 잘못 예측한 건

[2] 가장 치명적인 오예측건 (연락처글) case 1

회원 번호 : 1664664
저는 1남1녀를둔 아빠입니다 좀 빨리 결혼을 해서
지금 남19 여22살 남매를 둔 아빠입니다

회원 번호 : 1921439
육남일여로 다섯째입니다.제가 아직 혼자라 제일 부족하게 느껴지네요~.

회원 번호 : 1251035
2남 3녀중 장남. 부모님은 안계시며 여동생둘은 일본시민 누님은 전주에 거주 여조카는 아랍 에미레이트 승무원 남동생은 GM 연구실 부장...딸1아들1 딸은 출가했고 아들은 엄마하고 거주. ㅡ저는 솔로거주.

⇒ 가족소개글에서 주로 발생함
O남O여 패턴이 연락처 글을 쓰는 소개글 패턴과 유사해서 해당 글쓴 몇 사례들이 연락처 글로 모델이 예측함

+ 추가

- 프로필에 노출되어 있지만 **무성의 글**이거나,
  연락처가 노출되어 있는 회원 자기소개글 및 가족소개글이 있어 후처리 필요

[1] 자기소개글 무성의글/연락처 노출 중으로 처리 필요 : 92건
[2] 가족소개글 무성의글/연락처 노출 중으로 처리 필요 : 113건

=> 첨부된 회원 번호 및 해당 소개글 파일 참고