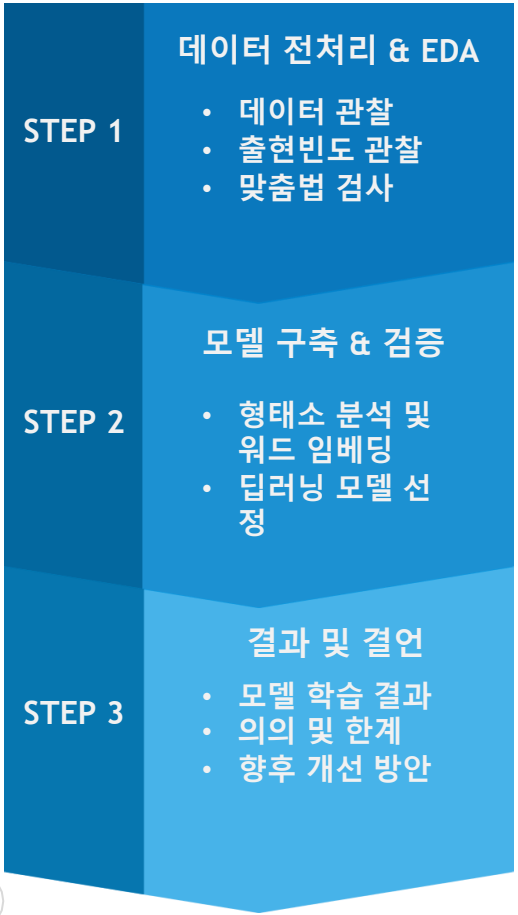
Dacon 14회 금융문자 분석 경진대회

스미스 요원
(김명수, 김백두, 정성현)

목차



1. EDA - 데이터 관찰

XXX 고객님항상 저희 XXX은행 미XXX점에 베풀어 주시는 성원과 애정에 깊은 감...	0
XXX 고객님안녕하십니까? 저는 여러해XXX 고객님과 함께 한 XXX XXX은행 개...	0
믿고 거래해주셔서 진심으로감사합니다 행복한하루되세요XXX은행신해운대XXX	0
사랑과 행복이 가득한 설날 되세요XXX서강 XXX올림	0
XXX 고객님그동안 감사드립니다꾸뻑금번 인사이동으로 제가 1월16일부터 영도지점장으...	0
내점하신 대출관련하여 전XXX드립니다.XXX은행XXX올림	0
시장동향(0130 마감 기준)-KOSPI: 2083.59pt(0.81%)126 마감...	0
연휴 잘 보내셨나요?활기찬 한주 시작하는 즐거운하루되세요XXX드림	0
명절선물(과일)발송했습니다약소하지만 성의껏준비했습니다XXX육둥 XXX	0
XXX은행XXX지점XXX대리입니다만기관련으로 전XXX드립니다	0
새해 복많이 받으시고 축복과 행운이 가득하기를 기원합니다 국XXX XXX 올림	0
XXX 고객님안녕하십니까고객님과 첫만남이 바로 엊그제같은데 벌써 아쉬운 인사를 드리...	0
지금이순간가장행복한XXX고객님이셨으면좋겠습니다행복한주말되세요XXX올림	0

(광고)XXX신용관리 그것이 알고 싶다나의 신용과 재무상태는 직접 관리해야지 누군가...	1
(광고)안녕하세요.잠깐의 여XXX을 내어 읽어 주신다면 도움이 되실거라 생각되어 저...	1
(광고)XXX신년특별빠르게천만원월39000원통화1번(무료거부XXX-XXX-XXX)	1
(광고)한국citi bank나의 대출한도와 금리는?대출때문에 고민하고 있다거나 높은...	1
(광고)한국XXX XXX이글을 읽는데 2분의 시간만 투자하시면 많은 도움이 되실껍니...	1
(광고)한국 (XXX XXX)2분의 시간만 투자하시면 월200만원 SAVE(절감)되...	1
(광고)한국citi bank나의 대출한도와 금리는?대출때문에 고민하고 있다거나 높은...	1
(광고)한국citi bank나의 대출한도와 금리는?대출때문에 고민하고 있다거나 높은...	1
(광고)XXXBXXX추운겨울 따뜻한금융 XXXBXXX따뜻한 금융이 누구보다 XXX ...	1
(광고)XXX(광고)지만 혜택되는 정보!!!고객을 최우선으로 생각하는 XXXBanX...	1
광 고) XXX이 글을 읽는데 2분만 시간을 투자하세요!매월 나가고있는 불입금을 줄...	1
(광고)XXXBXXX더 가까이 더 큰 혜택 XXXBXXX하고 싶은 것을 하고 좋은 ...	1
(광고)한국citi bank나의 대출한도와 금리는?대출때문에 고민하고 있다거나 높은...	1

스미싱 문자메시지와, 비 스미싱 메시지를 관찰한 결과 메시지 유형이 뚜렷히 달라짐을 확인

스미싱 : **불특정 다수 대상, 불법적 이익 취득 목적 문자메시지**
비 스미싱 : **기존 고객 대상, 안부, 명절 인사,**

1. EDA - 빈도수 관찰

[정상 문자]

- xxx : 기존 거래 고객님의 실명, 메시지 전송 직원명 혹은 지점명이 다수
- 올림, 감사, 고객, 부탁과 같이 공손한 표현 및 지점, 전담직원 등 구체적 단어 포함

[스미싱 문자]

- xxx : 불법 대환대출 업체명이 대다수
- 상담, 상품, 한도 많이 출현
- 등급, 금리, 신용, 부채 등 자극적인 표현 확인

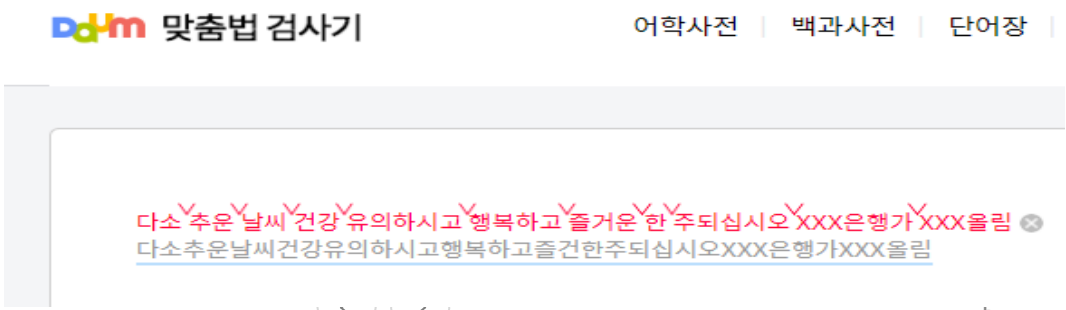
단어 출현 빈도에서 메시지 성향 차이 재확인

1. EDA - 맞춤법 검사

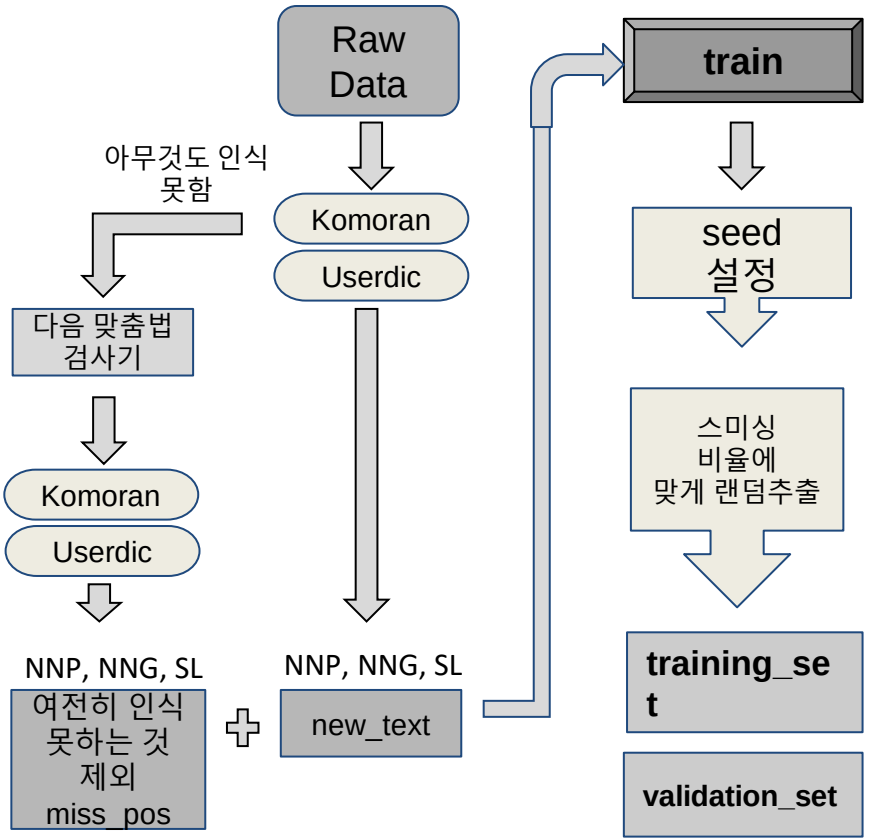
	id	year_month	text	smishing	pos
262	294	2017-01	하남힐즈110일2차중도금보증료157220원입금부탁드립니다XXX은행하남풍산지점	0	
438	481	2017-01	다소추운날씨건강유의하시고행복하고즐건한주되십시오XXX은행가XXX올림	0	

사람이 작성하는 메시지 특성상,
띄어쓰기와 맞춤법이 잘 지켜지지 않는 특성 존재
→이는 이후 이어지는 형태소 분석에 악영향

맞춤법 검사기 도입



2. 데이터 전처리 및 형태소 분석



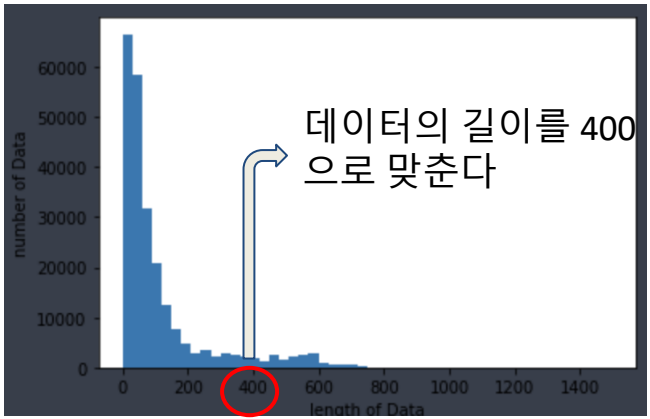
- 연산속도 빠르고 유의미한 명사 추출 성능 월등한 Komoran사용
- NNP : 고유명사
  NNG : 일반명사
  SL : 외국어 추출
- Komoran이 명사로 인식 못하는 것은 사용자 사전(serdic) 추가
  -Citi, 밸류, 서민지원 등
- EDA 결과 'XXX'의 빈도수가 매우 높은 것으로 확인되어 영어도 추출함 (tag : 'SL')
- 원활한 검증을 위하여 스미싱 비율에 맞춰 추출

2. 워드 임베딩

- word_vocab : 전처리 단계에서 획득한 단어들의 고유 번호를 매긴 스미스요원 팀 고유의 단어사전
- 원활한 학습을 위하여 Input data 의 길이를 400으로 맞춘다.
  -> EDA 시 400보다 긴 문자는 100여개로 파악됨
- pad_sequences 를 통해 워드 임베딩 진행
  -> 400 이하 문자들은 앞에서부터 0으로 채워 넣음

```
{'xxx': 1,
 '은행': 2,
 '고객': 3,
 '올림': 4,
 '감사': 5,
 '지점': 6,
 '대출': 7,
 '상품': 8,
 '행복': 9,
 '상담': 10,
 '금리': 11,
 '거래': 12,
 '하루': 13,
 '금융': 14,
 '부탁': 15,
 '한도': 16,
 '주말': 17,
 '전화': 18,
 '오늘': 19,
 '안내': 20,
```

Word Embedding



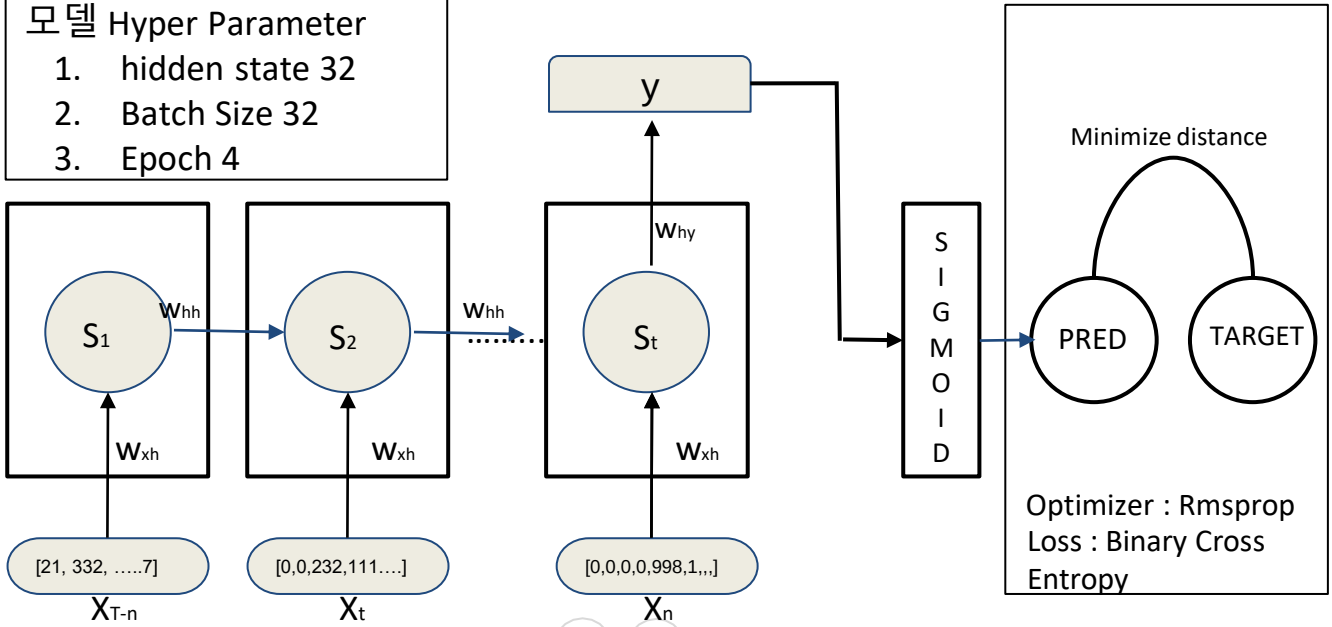

```
[   0,    0,    0, ..., 1442,    1,    4],
[   0,    0,    0, ...,    1,    1,   89],
[   0,    0,    0, ...,    1,   43,  104],
```

2. 딥러닝 모델 선정

RNN 순환신경망

- RNN은 시퀀스 데이터를 학습하는 데 최적화 된 모델
- 이전까지의 단어에 대한 기억을 바탕으로 새로운 단어를 이해하면서 스미싱 문자인지 아닌지 예측할 수 있다.

모델 Hyper Parameter
1. hidden state 32
2. Batch Size 32
3. Epoch 4



3. 의의 및 한계점

의의

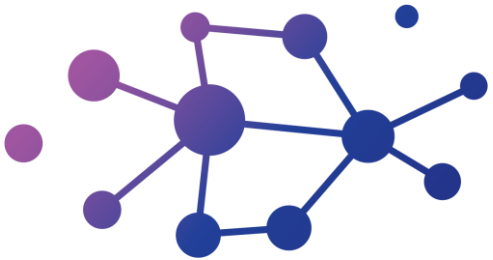
1. 맞춤법 검사기와 user_dic을 통하여 데이터를 정제하여 분류 모델의 성능을 최대한 끌어올림.
2. 머신러닝기법을 사용했을 때보다 딥러닝 RNN 기법을 사용하였을 때 분류 모델의 성능이 더 뛰어남.

한계점

1. 맞춤법 검사를 한 뒤에도 komoran이 8개의 행에 대해서 NNP, NNG, SL을 가져오지 못함 : 자음, 모음 숫자만으로 구성된 문자거나, 기호가 뒤에 붙어있어 인식을 못하는 경우
2. 딥러닝을 통해 모델을 만들었으나 이 모델이 왜 더 잘 학습했고 분류를 잘하는 지는 명확히 설명하기 힘듦

3. 향후 개선 방안

1. 딥러닝을 알고리즘을 통해 학습된, 정밀도가 높은 맞춤법 검사기 도입
   a. PykoSpacing 등
2. 형태소 분석시, 명사 외 타 형태소 적용
   a. 어간, 의존명사 등
3. Tokenizer에 의해 생성된 Word-Index에서 불용어의 적절한 제거
4. State-of-Art Word Embedding 모델 사용
5. State-of-Art NLP Model 사용
   a. bi-rnn, lstm, bi-lstm, cnn 등



THANK YOU