# 6주차 수행내역

2017103758 조문기

지난 2주간 수행했던 내역을 정리합니다.

**10/02경**(이쯤에 진행을 했던 것 같은데 파일을 찾을 수 없어 최근에 다시 진행했음.)

이미 나와있는 한국어 형태소 분석기(KoNLPy 라이브러리에 있는, Hannanum, Komoran, Mecab, Kkma, Okt)를 비교 분석하는 활동을 함.

이 때, 미리 구해 두었던 한국어 욕설 댓글 데이터들을 사용함.

수행 결과, Komoran과 Mecab이 우수한 성능을 내는 것을 확인함(수행 시간).

이 둘 중, 받침 마저도 분해하지 않는 분석기인 Mecab을 사용하는 것이 낫다고 판단함(ex. '난' -> '나' + 'ㄴ')

**10/03**

미리 구해 두었던 한국어 욕설 댓글 데이터들을 바탕으로 TF-IDF 벡터화 방식을 사용해 봄. 벡터화 한 데이터를 바탕으로 가장 기본적인 Classifier인 MultinomialNB를 이용해 욕설 댓글인지 아닌지 분류해 봄.

**10/04**

mecab을 사용해 훈련용 욕설 데이터들을 형태소 분석함. 분석 과정에서 stopword(불용어)를 따로 지정해주어 제외할 단어들을 선정함.

이후, 이를 tensorflow에 내장되어 있는 Tokenizer와 pad_sequences 라이브러리를 활용하여 토큰화 및 패딩 작업을 진행함.

패딩 작업을 완료한 데이터들을 가지고 언어 모델 중 단순한 형태의 BiLSTM 모델을 적용시켜 봄.

Test 정확도는 85.23%가 나옴.

**10/05**

이번에는 위에서 했던 것을 1D CNN에 적용시켜 봄.

테스트 정확도는 70.55%가 나옴. BiLSTM 기법보다 정확도가 굉장히 낮아졌음.

**10/07**

mecab으로 분석한 형태소들을 바탕으로 Word2Vec을 적용시켜 봄.

이 때, Phrase라는 라이브러리를 사용해 단어를 하나만 보는 것이 아니라 bigram, trigram으로 두 개, 세 개로 묶어서 진행해 보는 방법도 해 보았음.

이렇게 만든 w2v 모델을 바탕으로 위에서 좋은 성능을 보였던 BiLSTM을 적용시켜 봄.

'안녕하세요? 시발 개빡치네'라는 문장에 대해 79.91%의 확률로 욕설이라는 것을 진단하는 것을 볼 수 있었음.

**10/08**

훈련 데이터들을 더 수집하고자 크롤링 코드를 만들어 인터넷 커뮤니티 중 하나인 디씨인사이드의 게시글 제목들을 수집함(10000개).

이 때 수집한 데이터들은 훈련 데이터 셋으로 추가함(라벨링이 된).

지금까지의 데이터 상황: 임베딩 모델을 만들기 위한 unlabled 데이터 약 20만개, 훈련을 위한 labeled 데이터 약 2.4만개

labeled 데이터의 분포를 확인해 보면 욕설로 분류되지 않은 데이터들이 월등히 더 많음을 알 수 있다. 이것이 나중에 어떤 영향을 끼칠까?

**10/09**

이전에 수집했던 unlabeled 욕설 데이터 20만개를 가지고 W2V, FastText 모델을 만듦.

이 때, 앞서 사용했던 trigram 기법을 사용함.

W2V와 FT모델에 대해 most_similar 메서드를 사용해 유사도를 예측하고자 하니, w2v는 모르는 단어에 대해서는 결과값을 내지 못하고, FT는 결과값을 내는 것을 확인함.

=> FT모델을 사용하는 것이 낫다고 판단함.

**10/10**

앞선 모델은 단어 그대로를 사용했다면, 이번에는 단어를 자음과 모음으로 분리한 형태에 대해서 모델을 만듦(ex. 단어 -> ㄷㅏㄴㅇㅓ)

자음과 모음을 분리함으로써 오타와 같은 노이즈에 더 강해질 것이라고 판단함.

**10/11**

자모를 분리한 단어를 사용한 FT모델을 사용해 BiLSTM 모델을 적용시켜 봄.

**10/12**

자모를 분리한 단어를 사용한 FT모델을 사용해 1D-CNN 모델을 적용시켜 봄.

**10/15**

모델 적용 시 욕설 문장 자체는 잘 찾으나 positive한 문장에 대해서도 욕설로 판별하는 경우가 자주 발생. 이를 해결하기 위해 FastText 모델을 다시 수정해보기로 함. 그러면서 새로운 데이터 셋을 찾음(aihub). 데이터 셋의 크기가 너무 커(약 1억개) 합리적인 크기의 범위로(약 1천만개) 줄이고, 이를 원래 unlabel 데이터에 합쳐서 임베딩 모델을 만들기 위한 새로운 unlabeled data로 구성함.

새로운 데이터들을 토대로 새로운 FT 모델을 만들 예정.

**지금까지의 진행 상황 요약**

약 1천만개의 unlabeled data와 약 2.4만개의 labeled data를 수집했다.

형태소 분석기는 Mecab을 사용한다.

형태소 분석 이후 토큰화된 단어들을 자음과 모음으로 분리해서 사용한다.

단어 임베딩 모델은 FastText를 사용한다.

단어 임베딩 모델을 사용해 labeled data들에 대해 벡터화를 진행하고, 이 벡터화된 데이터들을 Input data로 사용해 모델들을 학습시켜봤다. (1DCNN, BiLSTM)

**향후 진행 방향**

1. 데이터들을 좀 더 구해볼까? -> aihub 데이터 양이 너무 커서 이것으로도 해결될 듯?

2. 언어 모델(BiLSTM, GRU, 1D-CNN) 모델들을 Sequential하게 쌓을 때 날을 잡고 좀 더 다양하게 쌓아보자.

3. 조금 더 심화된 언어 모델도 한번 학습을 해볼까? (BERT 등)

4. 임베딩 모델의 파라미터들도 다양하게 변화시켜보자. (FastText를 사용하는 것은 고정하는 것으로...)

학습했던 자료들은 개인 GitHub에 꾸준히 업로드 중입니다. 단순히 적용중인 모델들은 GitHub Repository의 Practice 폴더에 업로드 중이고, 실제로 적용될 것 같은 학습 내용들은 Works 폴더에 업로드 중입니다.

GitHub Link: https://github.com/siryuon/DataAnalysis_CapstoneDesign