**FastText 정보조사**

1. 기본개념

Facebook AI 팀에서 SNS단어 및 위키백과 단어들을 수집하여 벡터 공간에 각 단어를 -1부터 1까지의 실수로 임베딩변환(임베딩벡터)

기존 임베딩 방식인 1:1 임베딩(단어 1개당 임베딩벡터 1개 매치)의 한계점을 개선하고자 사용

단어를 N-Gram으로 나눠(주로 3~7) 각 단어에 다양한 임베딩 벡터 부여

ex) 3-Gram사용시

"hello" => "<he", "hel", "ell", "llo", "lo>"

             "<hel", "hell", "ello", "llo>"

               "<hell", "hello", "ello>"

               "<hello", "hello>"

               "<hello>"

위 예시처럼 단어를 나눠 각각의 단어에 임베딩 벡터 부여

🡪 기존 임베딩 방식의 Out Of Vocabulary 문제 해결가능

1. Use-Case

#### 김인환, 장백철, “트위터와 단어 임베딩을 사용한 인플루엔자 감지” 한국통신학회논문지, vol. 45, no. 1, pp. 96-104, Jan. 2020.

트위터 데이터와 FastText에서 질병과 관련된 트윗 및 단어들을 추출하여 질병 예측모델 구현

#### 이동호, 김유리, 김형준, Seung-Myun Park, 양유준 “Fake News Detection Using Deep Learning” 한국정보처리학회논문지, vol. 15, no. 4, pp. 119-130, 2019.

한국어 뉴스 데이터를 수집해 Bi-LSTM과 BCNN으로 구성된 다층 신경망을 쌓은 후 FastText 한국어 단어 임베딩을 적용하여 가짜뉴스를 식별하는 모델 생성

(데이터셋은 FastText에 게시된 한국어 위키 사전 데이터를 사용)

####

이외에도 다양한 국가에서 FatText와 LSTM/BERT 모델을 활용하여 비속어 탐지 모델을 개발한 것 확인

1. DarkWeb html 데이터셋 적용 방안

기존 eda\_data 데이터셋 내 토큰화된 단어 전처리 방안

🡪 FastText 파일 내 단어와 매치되는 것들만을 기준으로 모델 학습 및 테스트 과정 진행했음

- 버려지는 단어 약 45,000개(전체 단어 약 90,000개)

🡪 전체적으로 주소(.onion, https, .org, www, …) / 숫자 및 통화단위($240, 0.000039….) / 띄어쓰기 되지 않은 단어들(worksfaqaboutget 🡪 works, faq, about, get) / 다크웹 내 고유명사 및 은어

따라서, 3가지 진행방안 고려

1. 주소 및 숫자/통화 단위에 대해서는 최대한 정규식으로 전처리해서 단일 토큰으로 통일하는 작업을 수행한 뒤 FastText 내 단어 필터링 과정을 수행할 수 있음

🡪 다크웹 고유명사/은어를 여전히 학습하지 않게 되는 단점

1. 위 정규식으로 전처리 하지 않고 버려지는 단어와 은어파일을 FastText에 추가하여 학습

🡪 11/29부로 테스트단계

1. 1번 방안처럼 주소/숫자/통화 단일토큰으로 전처리한 다음, FastText를 사용하지 않고 다른 전처리 방안을 탐색(TF-IDF, CountVectorizer, Bag Of Words……)

🡪 FastText 필터링에 비해 모델 성능이 검증되지 않았음