AI,
이세계로 가는
길을 알려줘!

JUDY CHOI

오늘의 ~~오덕후~~ 발표자

- 이름 : 최민주 (Judy Choi)
- 직업 : NLP Engineer
- 연구분야 : 기계번역
- 좌우명 : 덕업일치
- TMI : 대학원 중도탈출 후 취업해서 어제 첫 출근함

평범한~~오덕후~~ 개발자
C씨,
현생은 따분한 일의
연속이다.

그러던 어느 날,
알 수 없는
유튜브 알고리즘에 이끌려
2D의 세계에
눈을 뜨는데...!


avex music
犬夜叉
奇架篇 戏


괴도 키드
©2010 GOSHO AOYAMA / SHOGAKUKAN・MAGIC KAITO COMMITTEE


하이큐!!
TO THE TOP

C씨는 급기야 현생을
버리고
2D속 이세계로 떠나기로
결심한다.

B.U.T

C씨는 이세계로
가는 방법을
모른다.

하지만 개발자 C씨는
포기하지 않고 좋은
방법을 떠올렸다.

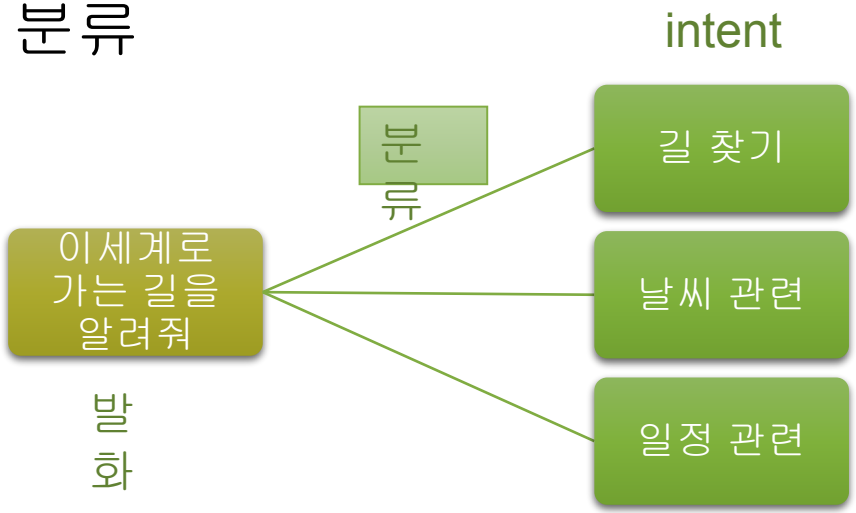
그것은 바로
이세계로 안내하는
AI를 직접 개발하는
것이었다.

AI,

이세계로 가는 길을 알려줘!

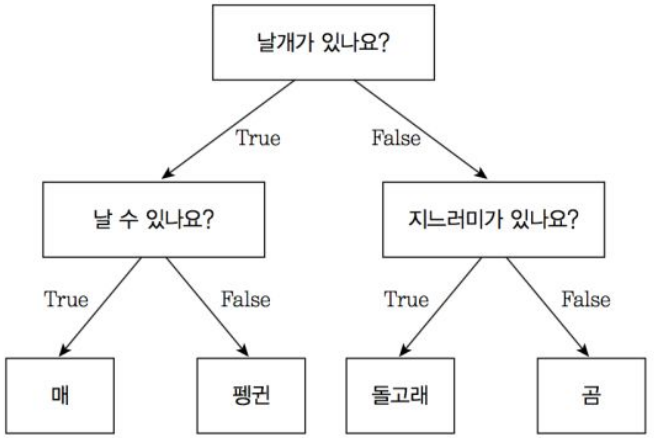
AI 설계

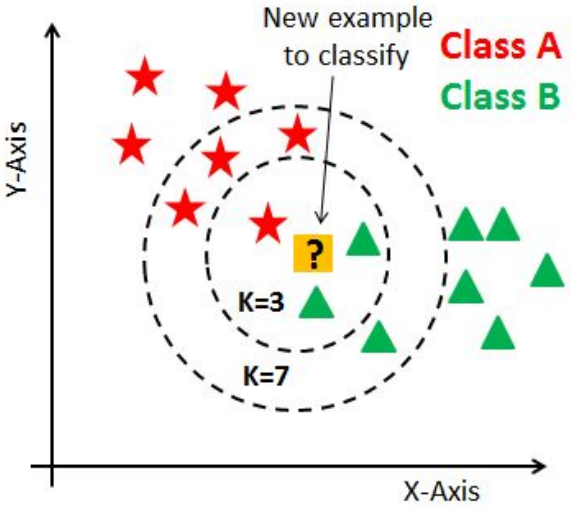
- 내가 하는 말을 AI가 이해해야 한다
  - 사용자 의도 파악 (User Intent)
    - ex) 이세계로 가는 길을 알려줘 == 경로 안내 요청
  - 사용자의 발화를 미리 정한 의도별로 분류 (User-Intent Classification)

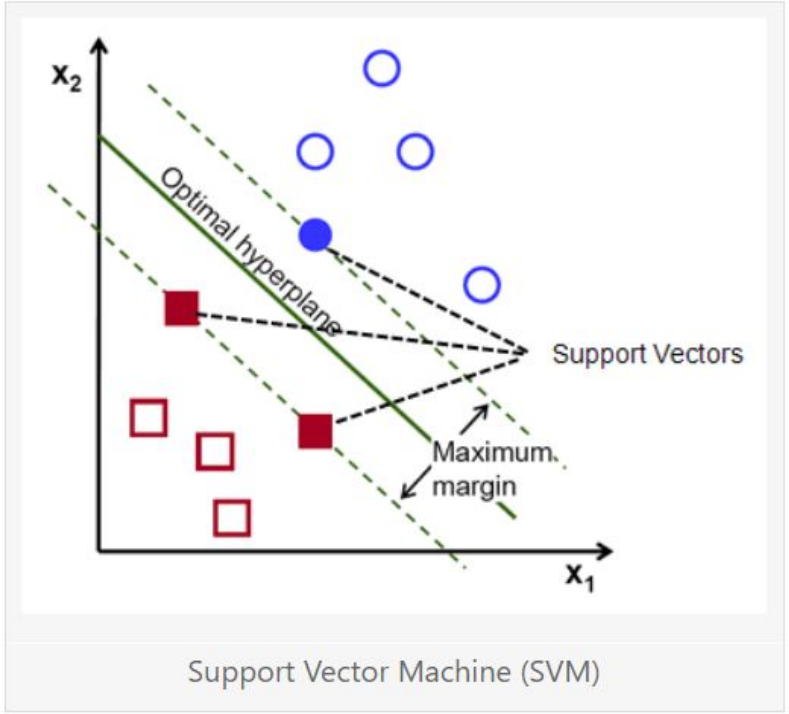


발화 분류 모델 결정

- Naive Bayes
- KNN (K-Nearest Neighbor)
- Decision Tree
- ✔ **SVM (Support Vector Machine)**
- …



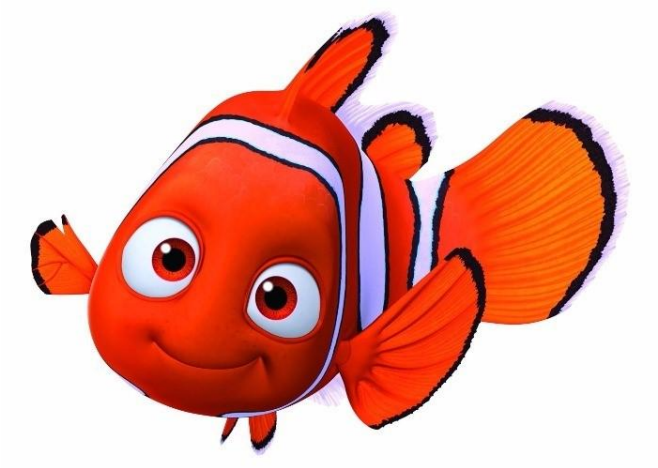




https://velog.io/@khsfun0312/KNN
https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-4-%EA%B2%B0%EC%A0%95-%ED%8A%B8%EB%A6%ACDecision-Tree
https://bioinformaticsandme.tistory.com/304

자연어처리 (NLP)

- 사람의 발화는 '자연어(Natural Language)' 에 해당
- 자연어를 컴퓨터가 처리하도록 (Natural Language Processing)



전처리 (Preprocessing)

- 1. Tokenize (토큰화, 형태소 분석)
  - ex) “이 방송은 보고 계시는 스폰서의 제공으로 보내드립니다”
    - ['이', '방송', '은', '보고', '계시는', '스폰서', '의', '제공', '으로', '보내', '드립니다’]

    - 형태소 원형 복원 (Stemming) 할 경우
      - ['이', '방송', '은', '보고', '계시다', '스폰서', '의', '제공', '으로', '보내다', '드리다’]

    - 불용어 제거 (Stopword)
      - ex) (영) the, a, an... (한) 그, 그리고, 아이구, 좀, 소인, 흐흐, 헉...
        - https://www.ranks.nl/stopwords/korean

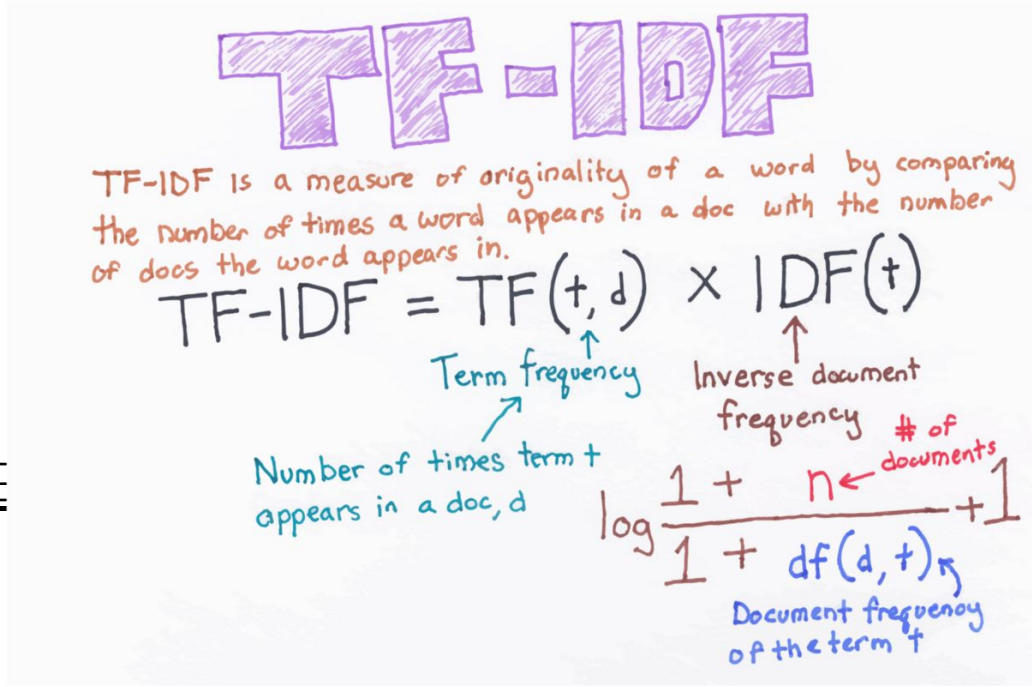
전처리 (Preprocessing)

- 2. TF-IDF vectorize
  - Term Frequency-Inverse Document Frequency
    - 단어의 빈도 수 (X) 중요도(O) 를 가중치로 줘서 주요 단어 추출
    - 특정 문서(문장, 발화) 내에서 자주 등장하는 단어일수록 TF-IDF 값 ↑
    - (참고) https://chan-lab.tistory.com/24

    - 적용 예시
      - 문서의 유사도를 구하는 작업
      - 검색 시스템에서 검색 결과의 중요도를 정하는 작업
      - 문서 내에서 특정 단어의 중요도를 구하는 작업

https://wikidocs.net/31698

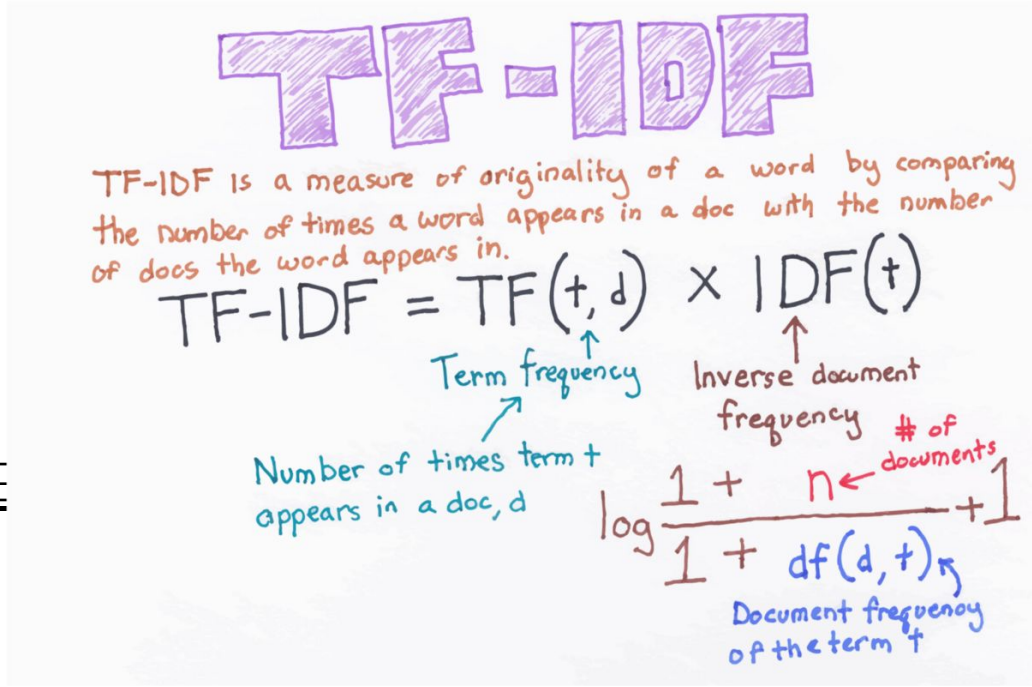
전처리 (Preprocessing)



- 2. TF-IDF vectorize
  - TF (Term Frequency, 단어 빈도)
    - 1개 문서(문장) 안에서 특정 단어의 등장 빈도
  - DF (Document Frequency, 문서 빈도)
    - 특정 단어가 나타나는 문서(문장) 수 (문서(문장) 빈도)
  - IDF (Inverse Document Frequency)
    - DF 에 ln 역수를 취해서 → 많이 등장하는 단어에 패널티

https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558

전처리 (Preprocessing)

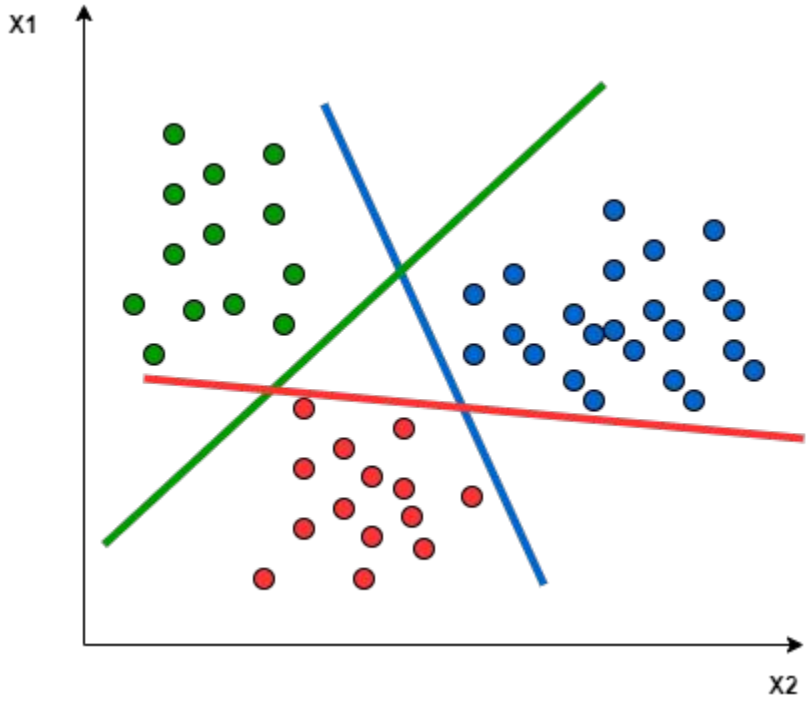


- 2. TF-IDF vectorize
  - TF (Term Frequency, 단어 빈도)
    - 1개 문서(문장) 안에서 특정 단어의 등장 빈도
  - DF (Document Frequency, 문서 빈도)
    - 특정 단어가 나타나는 문서(문장) 수 (문서(문장) 빈도)
  - IDF (Inverse Document Frequency)
    - DF 에 ln 역수를 취해서 → 많이 등장하는 단어에 패널티

https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558

텍스트 분류

- SVM in Scikit-learn
  - 문장(발화) 벡터를 각 label 로 가장 잘 분류하는 초평면 결정
  - ex) 3개의 label, 즉 intent
    - 길 찾기 - 이세계로 가는 길 알려줘
    - 날씨 - 이세계의 문은 언제 열려?
    - 일정 - 이세계에는 지금 비가 올까?



https://www.baeldung.com/cs/svm-multiclass-classification

☆실 습☆

과연 ~~오덕후~~ 개발자
C씨는
이세계로 갈 수
있을까?