

HanBert 패키지

2020. 8

주식회사 투블럭에이아이



TwoBlock Ai



Contents

- 도입 필요성
- HanBert 소개
- HanBert 패키지
- 문의처



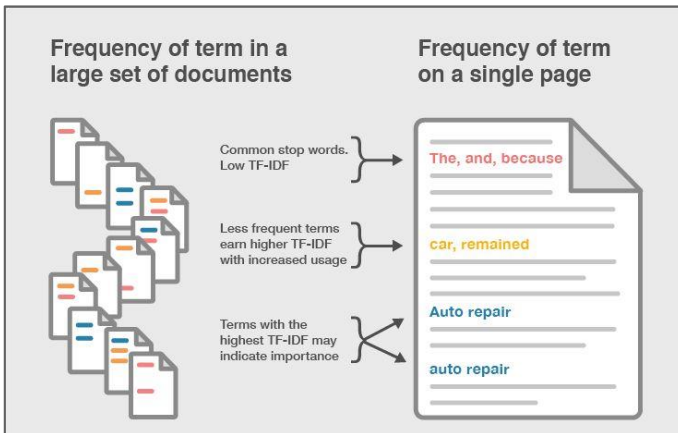
도입의 필요성 : 키워드 분석 수준의 70점짜리 NLP의 한계

정보량의 법칙(tf/idf)의 한계

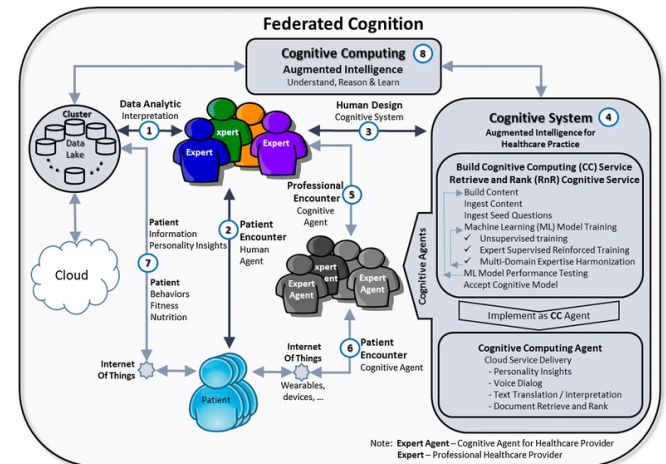
- 문서에서 중요한 단어 추출하여 문서의 의미를 대신 하는 방법
- "문서를 자동으로 요약하고 싶은데, 지금 수준으로는 못쓰겠어요."
- "게시판 댓글을 분석하려는데, 시스템을 구축하는 비용이 부담스러워요."

Classical NLP의 구축 절차와 비용

- 사전, 문법, 동의어 사전, 온톨로지, 의미 분류 등의 지식 베이스 구축이 필요
- 패턴 수집과 규칙 작성의 일관성과 상호 간섭을 고려하기가 어려움
- 어쨌든 시장에서 기대하는 성능을 만족시키지 못하고 있음



NLP 성능 리비히의 법칙





도입의 필요성 : BERT의 놀라운 성능

- 과거 기술을 통한 언어 이해 성능을 혁신적으로 추월

예) 네이버 영화평 DB에서 긍정과 부정의 판별

적용 기술	성능	속도	학습 비용
임의로 긍/부정 판별	50점	0ms	0원
Symbolic NLP 기술	75점 수준	50ms ~ 1000ms	1억원 수준
범용 딥러닝 기술을 활용	85점 수준	10ms ~ 30ms	100만원
BERT에 추가학습	91점	15 ~ 20ms	10만원

예) 논문에 발표된 BERT와의 성능 비교

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT_{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1



HanBert 소개

- 한국어 분석기 Moran 탑재
- 빠른 처리 속도
- 높은 품질
- 다양한 크기의 모델
- Vocab에 충분한 여유 공간

모델 코드	설명
54k vs 90k	Vocab의 크기 54k : 54000표제어 90k : 90000표제어
S/N/ML	모델의 크기와 학습량 S : 6층, 히든 : 768, 3.8M 학습 N : 12층, 히든 : 768, 5M 학습 ML : 18층, 히든 : 1024, 10M 학습



HanBert 패키지

- **Basic**

- 한국어 분석기 Moran 제공
- 모델 사용 라이선스 제공
- 용도에 맞게 모델 선택

- **Premium**

- Task 별 Fine Tuning
- Task별 추가 학습 기술지원

- **기술이전**

- 한국어 분석기 튜닝
- Vocab 튜닝 기술 지원
- BERT 학습용 코퍼스 레코드 제공
- BERT 학습 기술지원
- Domain 별 BERT 구축

패키지	비용
Basic	10천만원/4GPU + License Fee/ Year
Premium	Basic + 3천만원/Task
기술이전	별도협의

(부가세 별도)



HanBert 도입 방법 : Basic

- **Basic 패키지** : HanBert에 Task의 추가 학습을 스스로 하는 경우
 1. 오픈소프를 통해서 모델 다운로드 : <https://github.com/tbai2019/HanBert-54k-N>
 - 자체적으로 Task 데이터로 학습 수행
 2. 도입 요청을 이메일로 통보 : info@tbai.info
 1. 상업적 사용의 Task를 알려주면, 적합한 모델을 추천해 드림
 2. 도입을 원하는 HanBert 모델 협의
 3. 운영 장비 규모를 고려한 도입 라이선스 숫자를 통보
 3. 세금 계산서 발행 및 비용 납부
 4. HanBert 모델을 제공 : 다운로드 링크 제공
 - 최종 제공할 HanBert 모델과 라이선스 제공
 5. 도입한 규모의 범위에서 BERT 모델 활용
 - 자체적으로 Task 학습
 - 라이선스 비용협의
- 기술 지원 범위 : Task에 적합한 모델 선택, API 서비스 방법 지원, Docker 이미지 제공



HanBert 도입 방법 : Premium

- **Premium 패키지** : HanBert에 Task의 추가 학습을 의뢰하는 경우
 1. 협의 요청을 이메일로 통보 : info@tbai.info
 1. 상업적 사용의 Task를 알려주면, 학습 데이터에 대한 컨설팅 제공
 2. 적합한 HanBert 모델 선택
 3. 학습 데이터는 고객 측에서 준비
 2. Premium 패키지 계약
 3. Task 학습 수행
 - 협의에 따라 학습용 장비와 학습 주체 선택
 - 학습 후 성능 확인된 모델에 라이선스 제공
 4. 도입한 규모의 범위에서 BERT 모델 활용
 - Basic 모델과 동일하게 활용하는 GPU의 개수에 따라 비용 산정
 - Basic 모델에 대한 라이선스 제공
- 기술 지원 범위 : Task에 적합한 모델 선택, Task 학습 , Moran 튜닝 + Basic 패키지



HanBert 도입 방법 : 기술이전

- **기술이전 패키지** : 자체적으로 BERT를 만드는 경우
 1. 협의 요청을 이메일로 통보 : info@tbai.info
 - 자체 보유 코퍼스의 규모와 BERT 활용 목적을 통보
 2. 기술이전 패키지 세금 계산서 발행 및 비용 납부
 3. Moran 튜닝 및 Vocab 결정
 - 활용 분야에 따라, 한국어 분석기의 기본 사전을 튜닝
 - 활용하는 코퍼스에 따라 Vocab을 결정
 4. BERT 학습용 코퍼스 데이터 제공
 - Vocab을 적용하여 BERT 학습을 위한 학습 데이터 형태로 제공
 5. 자체적으로 BERT 학습
 - 고객이 보유하고 있는 코퍼스를 BERT 학습용 데이터로 변환
 - BERT 학습용 데이터로 BERT 모델 학습
- 보통 총 6~7개월의 기간이 소요됨
- 기술 지원 범위 : Moran 튜닝, Vocab 설정, BERT 학습용 데이터 변환, BERT 학습 기술지원

thank you

우리 회사에 가장 적합한

심층언어모델은 무조건, "HanBert"입니다.



<https://twoblockai.com>

문의: 070-8728-2233

Email: youngkim@tai.info

담당자: 김영훈 이사