오픈소스로 여는 뉴노멀

2020 공개SW 페스티벌

과학기술정보통신부
Ministry of Science and ICT

nipa 정보통신산업진흥원
National IT Industry Promotion Agency

자연어처리 커뮤니티를 위한 데이터 큐레이션

조원익, 송영숙,문상환

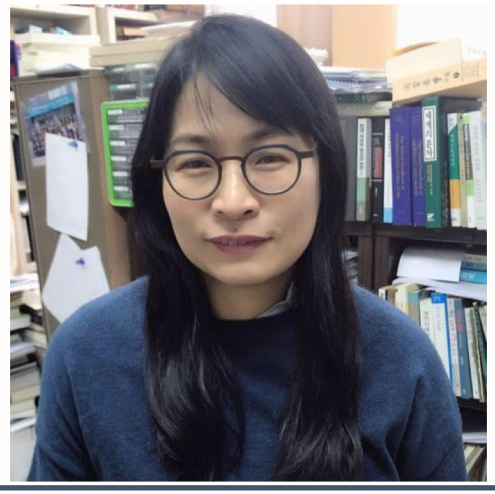
목차

1. 발표자 소개

발표자*

조원익*
서울대학교
전기정보공학부

송영숙*
경희대학교
국어국문학과

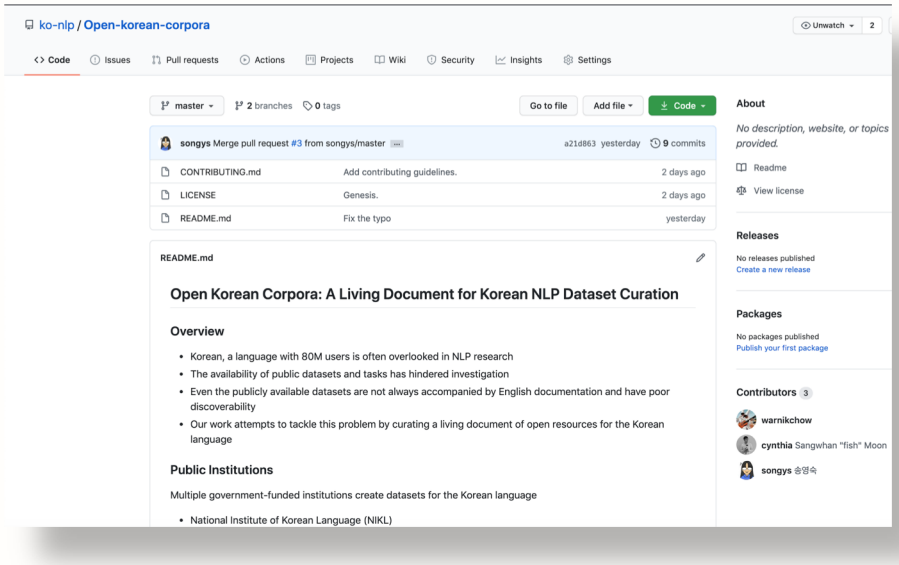


문상환
동경공업대학교
& 오드컨셉

2. 데이터 큐레이션

데이터 큐레이션이란?

- 데이터 큐레이션: 데이터 구축과 생성뿐만 아니라 데이터의 활용 가치를 높이는 모든 활동을 포함
- 활동 배경: 각자 데이터를 만들고 Github에 공개하는 경험을 하면서 체계적 데이터 관리의 필요성을 실감



https://github.com/ko-nlp/Open-korean-corpora

한국어 자연어처리와 데이터

한국어 자연어처리가 어렵다?

- 어디서부터 시작해야 할지 모르겠다
- 데이터셋을 구하는 것이 어렵다
- 누군가 데이터를 공개했지만 검색이 잘 되지 않는다
- 한국어의 특성을 파악하기가 어렵다

결론) 공개 데이터의 목록을 만들고 쉽게 접근할 수 있는 방법을 찾아보자!

한국어 공개 데이터셋

국가적 규모에서 구축하고 공개한 데이터

- 국립국어원 모두의 말뭉치
- NIA AI HUB
- ETRI 언어자원

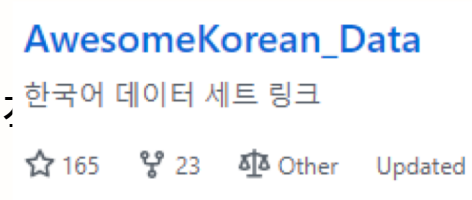
개인 및 기업이 구축하고 공개한 데이터

No	Dataset	Typical Usage	Provider	Docu	License	Volume	Goal	Lang	Description
1	Question Pair	Paraphrase detection	Academia	DOC	com/red	10K (p)	-	ko	유사 문장
2	KorNLI	NLI	Industry	INT	com/red	1,000K (p)	-	ko	자연어 이해를 위한 데이터 세트
3	KorSTS	STS	Industry	INT	com/red	8,500 (p)	-	ko	자연어 이해를 위한 데이터 세트
4	ParaKQC	STS	Academia	INT	com/red	540K (p)	-	ko	Parallel dataset of Korean Questions and

헉! 의외로 많았습니다. 정리를 해야 하는데 ...

Open Korean Corpora: A Practical Report, 작은 시작

2019년 12월 15일 여기저기 저장되어 있던
오픈 데이터 세트의 링크를 한 곳에 정리하기 시



2020년 8월 21일 조원익님, 문상환님이
함께하면서 분석적으로 정리하고
NLP-OSS 2020@EMNLP에 제출

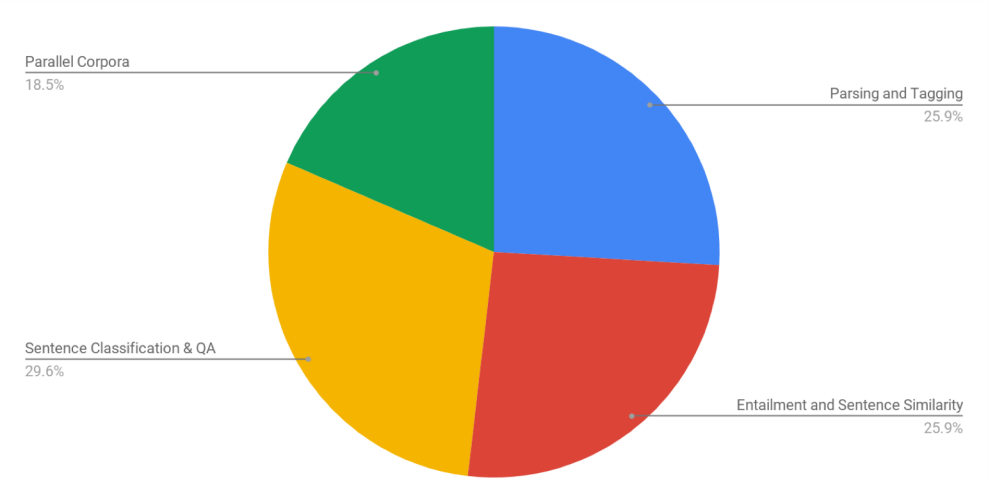
On behalf of the NLP-OSS 2020 Program Committee, I am delighted to inform you that the following submission has been accepted to appear at the conference:
Open Korean Corpora: A Practical Report

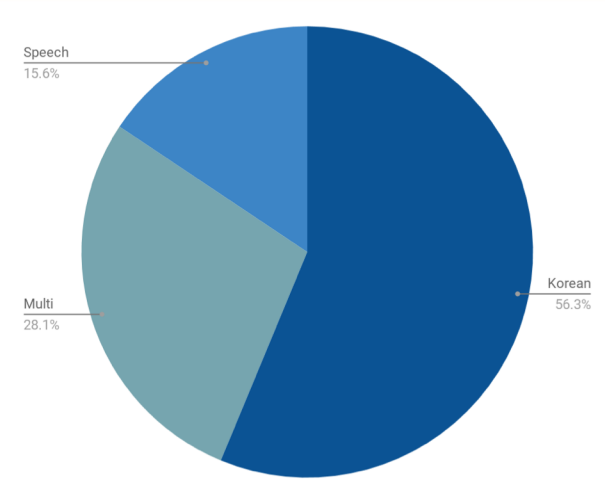
한국어 자연어처리 오픈 데이터 체계화의 contribution을 인정받아
게재 수락!

데이터 로더팀과 연결되어 지속 가능한 오픈소스의 체계화

조금 더 분석적으로 들여다보기!

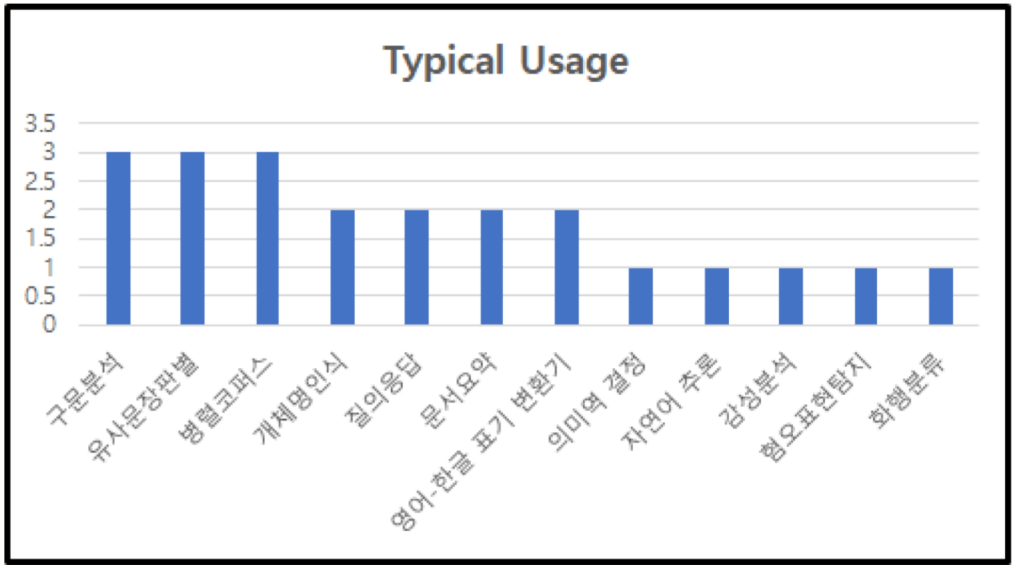
Statistics of the surveyed datasets

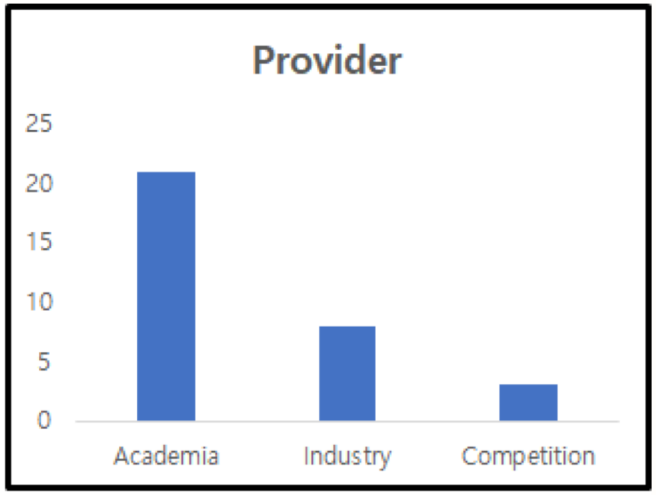




조금 더 분석적으로 들여다보기!

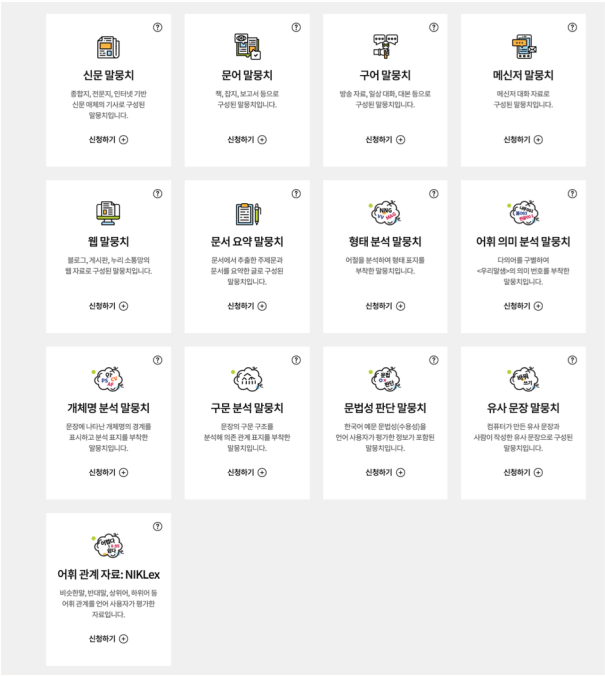
Statistics of the surveyed datasets



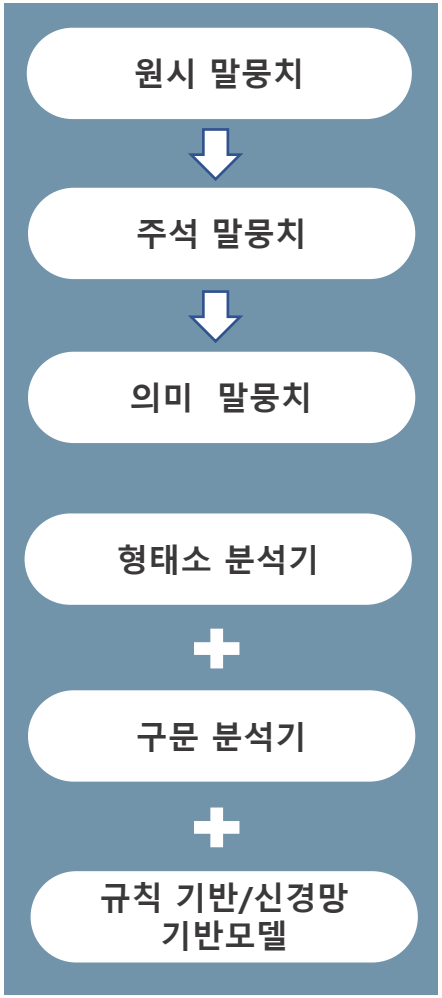


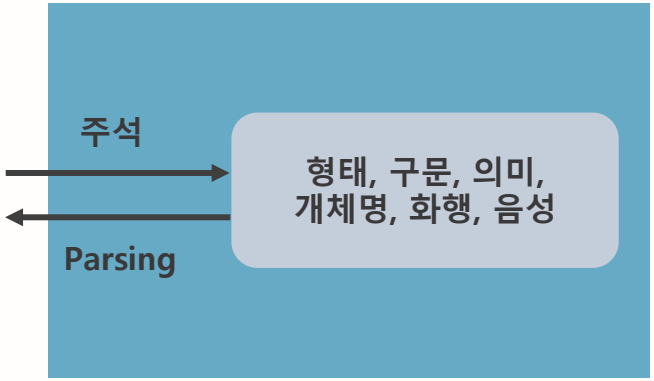
정교하고 분석적으로 구축된 말뭉치

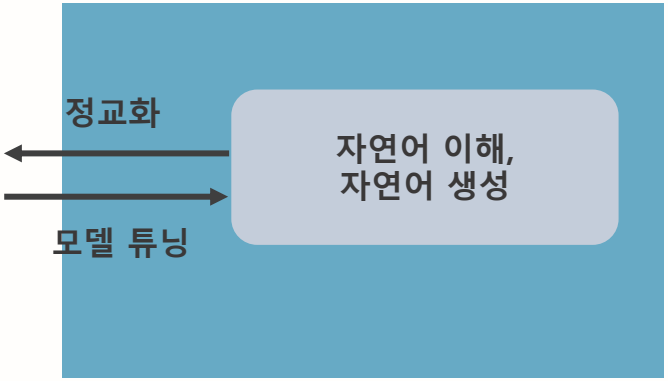
모두의 말뭉치



총 13종 18억 어절 분량의
정교하게 구축된 말뭉치



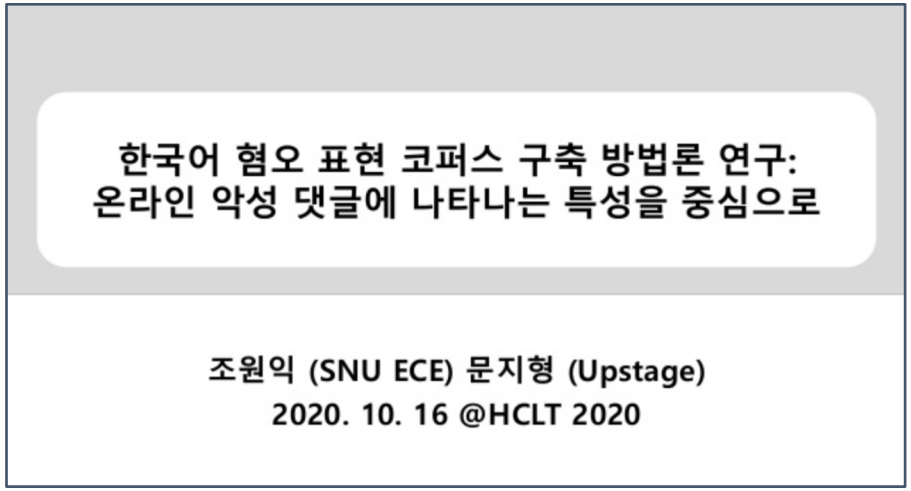


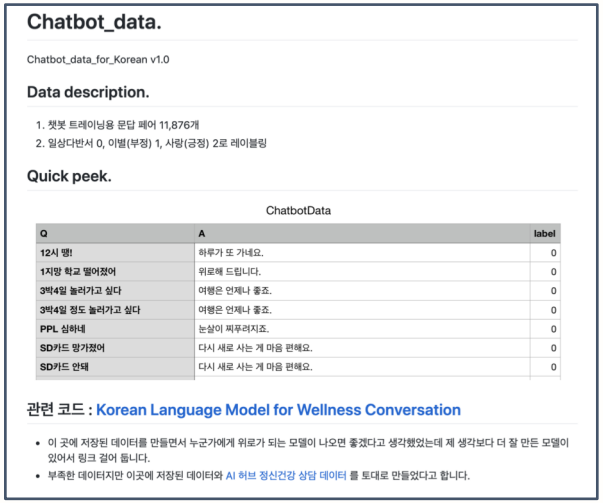


관련기사) https://news.v.daum.net/v/20201010050552824

의미와 화용 정보

Dataset	Typical Usage	Provider	Docu	License	Volume	Goal	Lang	Description
NSMC	Sentiment analysis	Academia	DOC	com/red	150K / 50K (s)	-	ko	댓글을 통한 감성 분석 데이터 세트
BEEP!	Hate speech detection	Academia	INT	com/red	8K / 500 / 1,000 (s)	-	ko	혐오 표현 관련 데이터
3i4K	Speech act classification	Academia	INT	com/red	55K / 6K (s)	-	ko	Intonation-aided intention identification for Korean
KorQuAD1	QA	Industry	INT	com/red (mod-x)	60K / 5K / 4K (p)	-	ko	질의 응답 데이트 세트 KorQuAD 설명 동영상
KorQuAD2	QA	Industry	article	com/red (mod-x)	80K / 10K / 10K (p)	-	ko	-
bab2min corpus	Sentiment analysis	Public Domain	DOC	com/red	200K / 100K (s)	-	ko	네이버 쇼핑과 게임 유통 서비스 Stream 리뷰





- 데이터의 다양화
- 구축 규모가 커지는 추세
- 혐오 표현, 감성 분석 등은 정교화, 세분화

출처) https://github.com/songys/AwesomeKorean_Data
혐오표현 관련 발표 링크 https://www.slideshare.net/WonIkCho/2010-hclt-hate-speech-238890804?fbclid=IwAR1CI4ATV-qTAG2J5WAnqmfsIHq6RjR-HkIGCUAMONbASY4gn4vZjaes3K0

화행, 음성 분석

더 많은 한국어 음성 데이터가 필요한 분야

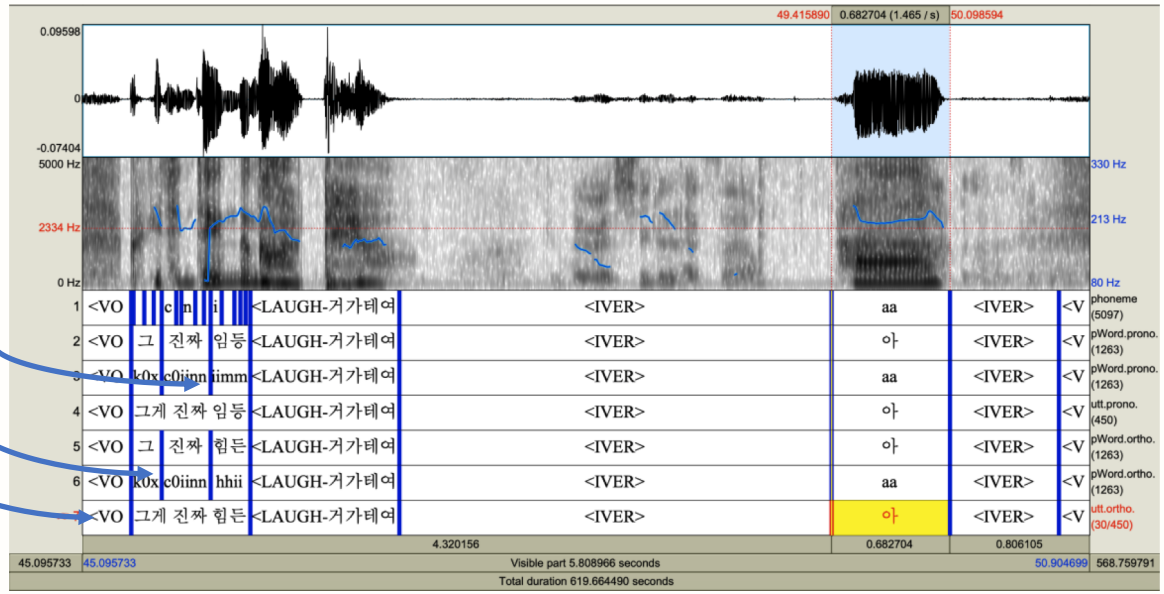
THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	B			r					R		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

IPA전사
발음전사
철자전사



Yun, W.,Yoon, K., Park, S., Lee, J., Cho, S., Kang, D., Byun, K., Hahn,H., & Kim, J. (2015). The Korean Corpus of Spontaneous Speech. Phoneticsand Speech Sciences, 7(2).

IPA : https://www.internationalphoneticassociation.org/
한국어 https://www.aihub.or.kr/aidata/105
소리, 영어 등 https://github.com/jim-schwoebel/voice_datasets

3. 커뮤니티 활동

커뮤니티 활동이란?

- 오래 함께 해 온 커뮤니티 구성원들과 거인의 어깨 위에서 세상 보기!
- 스터디 연합 모임을 통해 다른 전공의 사람들과 만나는 색다른 경험은 덤

NLP_스터디 일정표

파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말 MooSung Kim님이 2016년 4월 14일에 마지막으로 수정

	A	B	C
1		교재	TM - Mining Text Data - http://link.springer.com/book/10.1007/978-1-4614-3223-4/page/1 DNLP - Deep Learning for Natural Language Processing - http://cs224d.stanford.edu/syl 참고강좌) https://www.udacity.com/course/viewer#!/c-ud730/l-6370362152/m-637981181 AI - Knowledge-Based AI: Cognitive Systems - https://www.udacity.com/course/knowledg
2		일정 : 11월 19일~(격주 목요일)	
3		시간 : 7시간 30분~10시 30분(3시간)	
4		발표자 대모집중!!!!	
5	회차	날짜	TM
6	1	11/19/2015	8. Probabilistic Models for Text Mining
7	2	12/3/2015	보강) 중국어 처리
8	3	12/17/2015	9. Mining Text Streams (1)
9			10. Translingual Mining from Text Data
10	4	1/7/2015	9. Mining Text Streams (2)
11	5	1/21/2016	11.Text Mining in Multimedia
12	6	2/4/2016	12. Text Analytics in Social Media
13	7	2/18/2016	13. A Survey of Opinion Mining and Sentiment Analysis
14			감성분석 실습(http://doc.mindscale.kr/blog/2016/1/25/introduction-to-sentiment-ar
15	8	3/3/2016	14. Biomedical Text Mining: A Survey of Recent Progress
16	9	3/17/2016	Udacity DL - L2: Deep Neural Networks
17	10	3/31/2016	Udacity DL - L3: Convolutional Neural Networks
18	11	4/14/2016	Udacity DL - L4: Deep Models for Text and Sequences
19	회차	날짜	DNLP
20	1	11/19/2015	word2vec특강 (이론과 실습)
21	2	12/3/2015	doc2vec특강 (이론과 실습)
22	3	12/17/2015	
23	4	1/7/2015	GRUs and LSTMs - for machine translation
24	5	1/21/2016	Recursive neural networks -- for parsing
25	6	2/4/2016	Recursive neural networks -- for different tasks (e.g. sentiment analysis)
26	7	2/18/2016	
27			(완전기초) ud730 Lesson 1: From Machine Learning to Deep Learning
28	8	3/3/2016	Convolutional neural networks -- for sentence classification
29			Guest Lecture with Jason Weston from Facebook: Neural Models with Memory

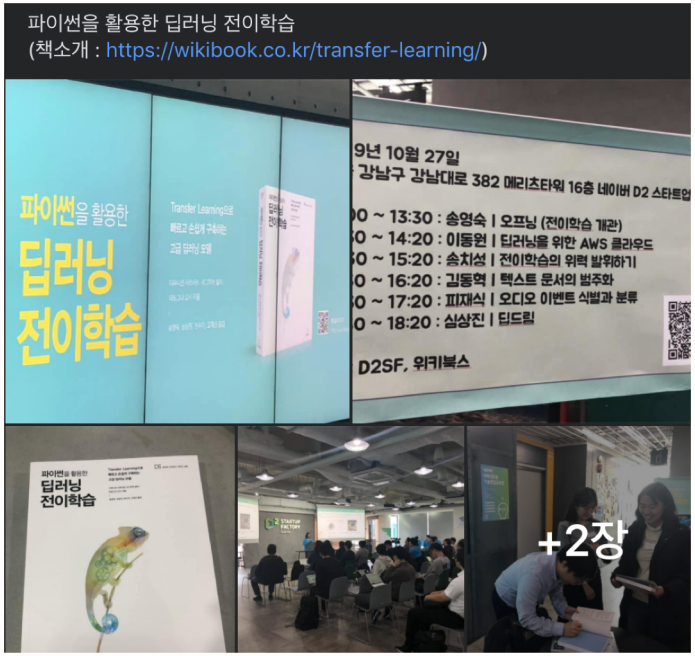
데이터그램 그룹 + 스터디뽀개기 연합 세미터 개최

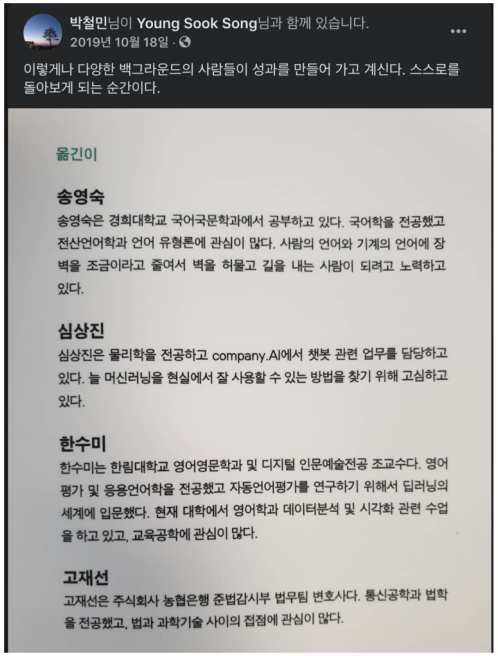
https://www.facebook.com/groups/datergram/

스터디를 리딩하고 내용을 설명하면?

더 많은 것을 배우게 된다!

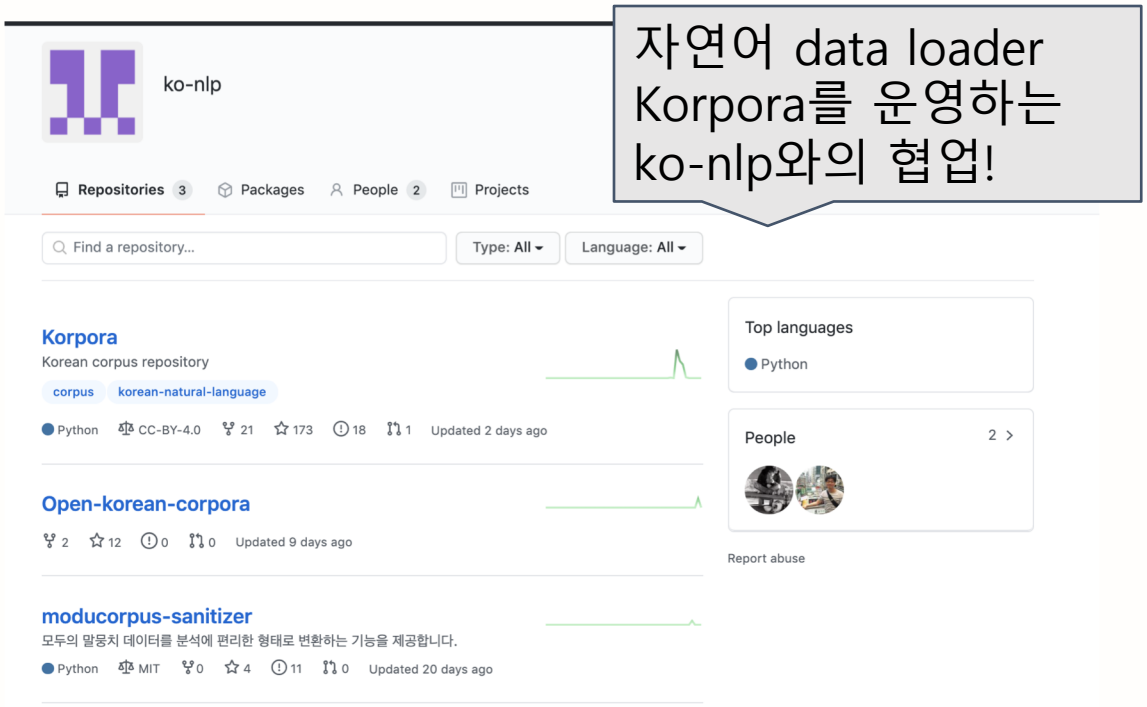
- 더 좋은 모임을 위해 페북 페이지, 그룹뿐 아니라 시간표, 가이드 등을 만들고 논의
- 논의를 통해 과정과 결과를 공유!
- 스터디에서 만난 분들과 번역을 하고 출판 기념회도 개최!





관심사가 비슷한 사람들이 만나면?

시너지 효과가 생긴다!



https://github.com/ko-nlp/Korpora in ko-nlp

스터디 공간등을 지원해 주신 곳들!

여러 단체들로부터의 후원이 있었습니다.
(NIPA, D2SF, MS, Python Korea 등)
이 자리를 빌어서 깊은 감사를 드립니다!
빈 공간에서 뜻 깊은 '만남'이 많았습니다.

후원

정보통신산업진흥원(NIPA)

Open UP

nipa 정보통신산업진흥원 National IT Industry Promotion Agency

위키북스

위키북스

https://github.com/ko-nlp/Korpora in ko-nlp

앞으로 …

OKC as Living Document

- Github repository 유지 관리를 통해 지속적으로 리스트 업데이트
- 한국어 버전과 영어 버전을 동시에 유지해서 한국어를 다루고자 하는 연구자 및 개발자들과 정보 공유
- 정기적으로 arxiv.org에 개정판 업데이트
- 한국어 공개 데이터 공유 장려를 통해 쉽게 협업할 수 있는 분위기 조성
- Data loader와의 협업을 통해 쉽게 다운로드하는 데에 이바지

We need community support!
https://github.com/ko-nlp/Open-korean-corpora/CONTRIBUTING.md

데이터

- 국립국어원 모두의 말뭉치 링크 https://corpus.korean.go.kr/
- 개인이 구축한 데이터 목록 https://github.com/songys/AwesomeKorean_Data

분석기(python 기반)

- PyKoSpacing 한국어 띄어쓰기 패키지
- KoNLPy, Soynlp 한국어 자연어 처리를 위한 형태소 분석기 패키지
- Kiwipiepy 한국어 형태소 분석기인 Kiwi(Korean Intelligent Word Identifier)의 Python 모듈

교재 등

- 자연어처리 교재 : https://wikidocs.net/21667 (attention까지만 집필)
- Transformer 튜토리얼 : https://www.youtube.com/watch?v=wxL3xRvKV9M&t=158s