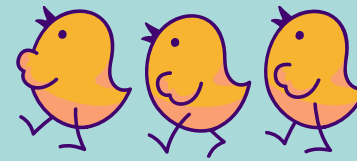


LangCon 2021



밀려오는 자연어 데이터 파도타기

경희대학교 국어국문학과 송영숙

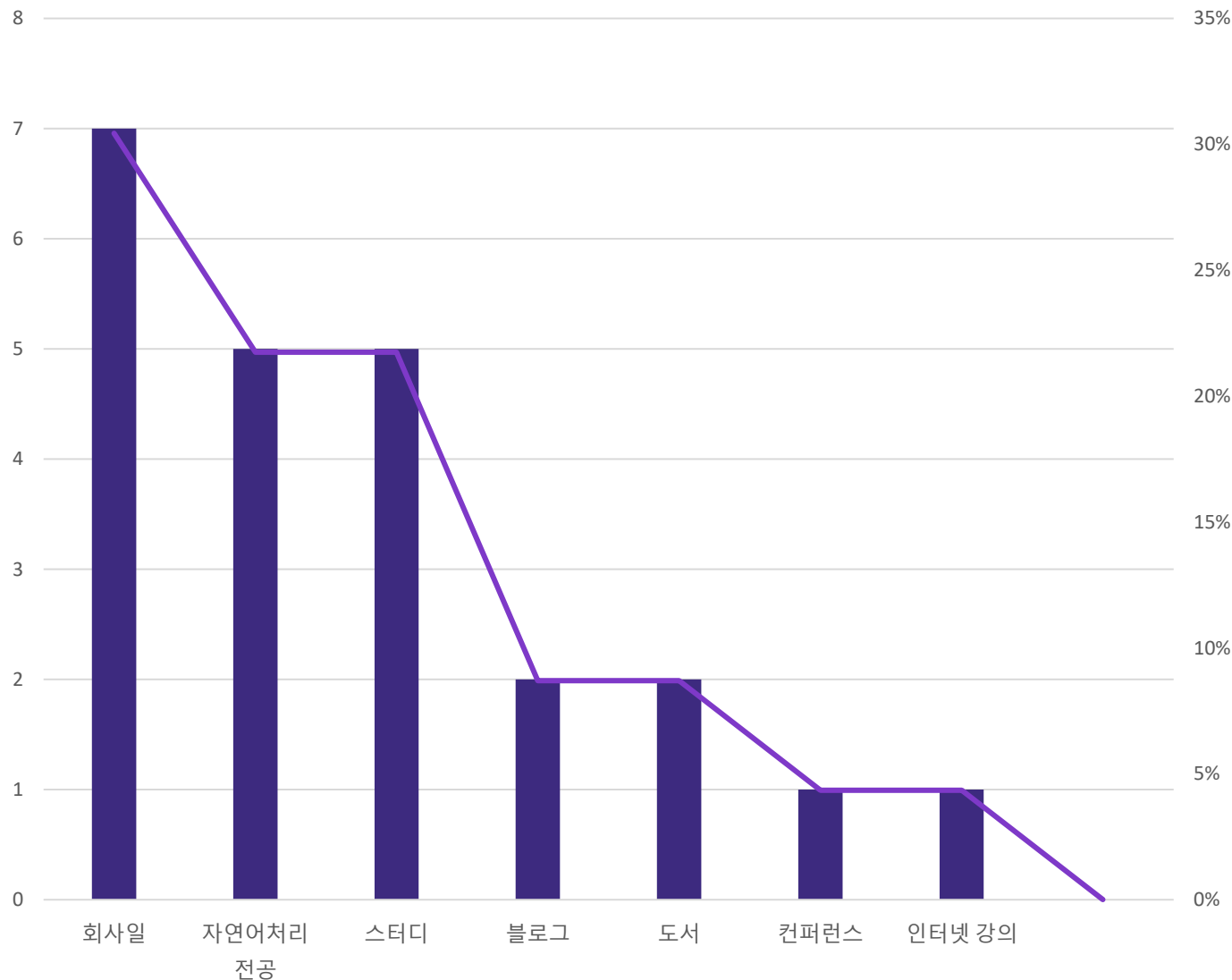
자연어처리에서 가장 중요하게
생각하시는 부분은 무엇인가요?



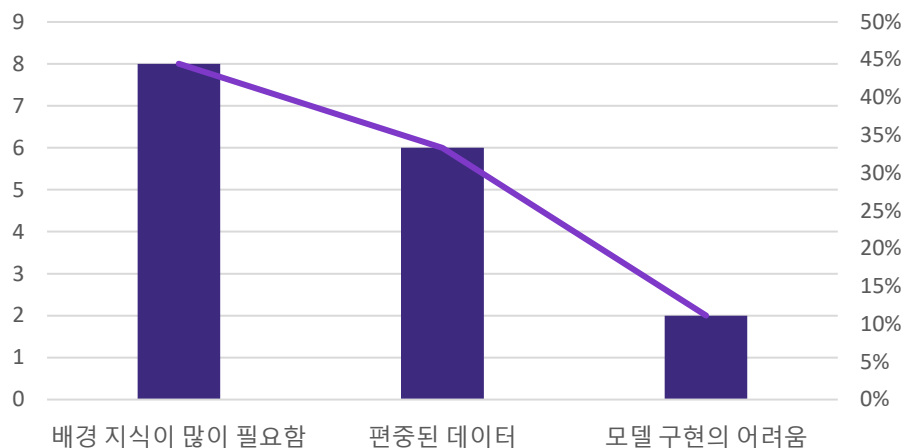


다들 어떻게 시작하셨는지 설문을 해봄

- 이 설문은 예비 질문의 성격을 지니고 있으며 그 결과에 대해서도 경향성을 파악하는 정도에서 이해 부탁드립니다.
- 총 24명이 답해 주셨으나 답하지 않은 항목과 기타 의견으로 답을 주신 분들이 많아서 문항별 참여 인원에는 차이가 있습니다
- 전공자들이 일로 자연어처리를 하고 있는 경우가 많았고, 스터디 등을 하면서 책이나 영상 강의 등의 도움을 받는 경우도 적지 않음을 알 수 있습니다.

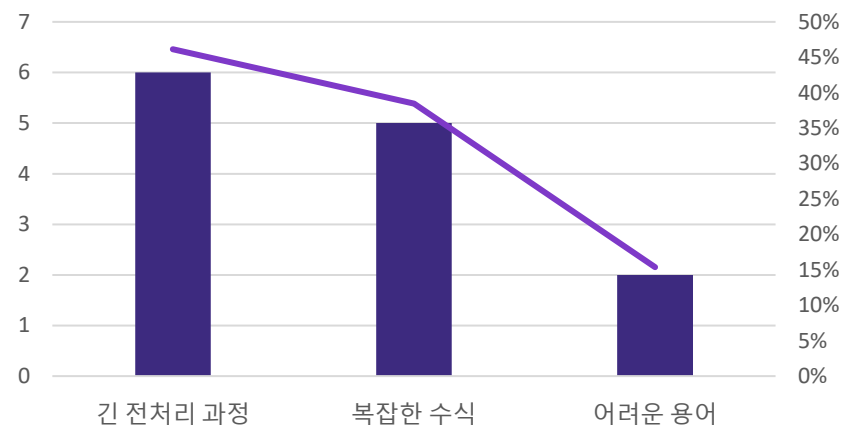


접근을 어렵게 만드는 요소



- 학습자에 대한 교육 및 학습 관련 오픈된 코스웍 및 지속적인 관리/의지
- 파이썬이 주를 이루고 타 언어 배제된 느낌, 그들만의 리그, 돌아가는 예시가 찾기 어려움, 시간이 지나고 나면 코드들이 동작하지 않음, 적절한 예시가 영문으로 되어 있음
- 인간으로써 너무나 자연스러운 자연어 생활. (...초딩때부터 당시에는 몰랐지만, 수업시간 단짓으로 형태소 분석을 시도하고 놀았습니다.)

작업할 때의 어려움



- 데이터셋의 라이선스 정책,
- 필요한 CS 기술들과, 통계 능력 동시에 요구되는 본능적 언어생활을 기계적 task를 분리해야 하는 상황 등
- 어려운 부분에 대한 선행 문제 해결방법 공유(Know-how)
- 언어학 지식배경, 한국어 전처리
- 딥러닝 모델 튜닝

어떤 데이터가 있었으면 좋겠다고 생각하시나요?

공유 정도

- 데이터는 잘 공유되고 있다고 생각합니다
- 공개된 것으로 공부하기엔 충분한 것 같습니다.. 꾸준히 양이 늘어나서 기간에 따른 변화를 알 수 있어도 우리 말에 좋을 것 같네요

최다 요청 감성분석

- 대화마다 감정의 변화 정도가 표현되어 있는 데이터 (나 진짜 기쁘다! (기쁨: 0.7, ...) -> 어..아니라고..? (기쁨: 0.2, ...))
- 정교한 긍부정어 사전?
- 단순 긍부정이 아니라 인간의 감정을 멀티 클래스 레이블링으로 다룬 감정 분석 데이터가 있었으면 좋겠습니다.
- 감정에 따른 음성 데이터

기타 의견

- 동일 의도의 이중 표현 데이터
- 어린이 발화 텍스트 데이터(언어학습 수준별)
- 기계번역 연구자로서 더 다양하고 많은 번역 쌍 데이터가 있으면 좋겠습니다.
- 과목별 교과서와 연습문제 정답지
- 매일 매일 업데이트 되는 신조어사전이 있으면 온라인 상의 텍스트를 분석할 때 유용할 것 같습니다.
- 욕설, 비속어 등 금칙어 및 비문 데이터
- 저작권 소멸 도서 데이터의 public domain화 된 데이터셋

어떤 한국어 데이터가 있나요?

<https://corpus.korean.go.kr/>

The screenshot displays a grid of dataset cards on the Korean National Corpus website. The cards are organized into four columns and five rows. Each card features a small icon, a title, a brief description, and a 'View Details' button. The datasets include:

- 비중반물 일용지 (Basic Vocabulary Daily Use)
- 의미역 기술 모형 (Meaning Role Model)
- 강성 분석 일용지 2020 (Strength Analysis Daily Use 2020)
- 신문 일용지 2020 (Newspaper Daily Use 2020)
- 개체명 분석 일용지 20... (Named Entity Analysis Daily Use 20...)
- 일상 대화 일용지 2020 (Daily Conversation Daily Use 2020)
- 일상 대화 음성 일용지... (Daily Conversation Audio Daily Use...)
- 무형 문화재 복원 일용... (Intangible Cultural Heritage Restoration Daily Use...)
- 상호 참조 해결 일용지... (Cross-reference Resolution Daily Use...)
- 구문 분석 일용지 (Syntax Analysis Daily Use)
- 서울말 남측체 발화 말... (Seoul dialect female style utterance...)
- 구어 일용지 (Spoken Daily Use)
- 신문 일용지 (Newspaper Daily Use)
- 문어 일용지 (Written Daily Use)
- 문서 요약 일용지 (Document Summarization Daily Use)
- 형태 분석 일용지 (Morphological Analysis Daily Use)
- 어휘 의미 분석 일용지 (Vocabulary Meaning Analysis Daily Use)
- 개체명 분석 일용지 (Named Entity Analysis Daily Use)
- 문법성 판단 일용지 (Grammaticality Judgment Daily Use)
- 유사 문장 일용지 (Similar Sentence Daily Use)
- 어휘 관계 자료: NIKLex (Vocabulary Relationship Data: NIKLex)

<https://aihub.or.kr/aihub-data/natural-language/about>

The screenshot displays a grid of task cards on the AI Hub website. The cards are organized into four columns and five rows. Each card features a small icon, a title, a brief description, and a 'View Details' button. The tasks include:

- 음성/자면어 감성 대화 일용지 (Audio/Text Sentiment Conversation Daily Use)
- 음성/자면어 고객 응대 음성 (Audio/Text Customer Service Audio)
- 음성/자면어 공공행정서비스 OCR (Audio/Text Public Administration Service OCR)
- 음성/자면어 기계음 (Audio/Text Machine Sound)
- 음성/자면어 다양한 형태의 한글 문자 OCR (Audio/Text Various forms of Korean text OCR)
- 음성/자면어 도서자료 기계음 (Audio/Text Book material machine sound)
- 음성/자면어 여행어 음성 (여행안내) (Audio/Text Travel audio (Travel Guide))
- 음성/자면어 여행어 음성 (소아, 유아) (Audio/Text Travel audio (Child, Infant))
- 음성/자면어 여행어 음성 (일본어) (Audio/Text Travel audio (Japanese))
- 음성/자면어 문서요약 텍스트 (Audio/Text Document summarization text)
- 음성/자면어 민원(물센터) 질의-응답 (Audio/Text Citizen (Water Center) Inquiry-Response)
- 음성/자면어 법률 자식베이스 (Audio/Text Legal child base)
- 음성/자면어 생활 및 거주환경 기반 VQA (Audio/Text Living and living environment based VQA)
- 음성/자면어 소상공인 고객 주문 질의-응답 텍스트 (Audio/Text Small business customer order inquiry-response text)
- 음성/자면어 수어 영상 (Audio/Text Sign language video)
- 음성/자면어 시각정보 기반 질의응답 (Audio/Text Visual information based inquiry-response)
- 음성/자면어 아외 실제 촬영 한글 이미지 (Audio/Text Outdoor actual shooting Korean image)
- 음성/자면어 일반상식 (Audio/Text General knowledge)
- 음성/자면어 자유대화 음성 (노인/노년) (Audio/Text Free conversation audio (Elderly/Old))
- 음성/자면어 자유대화 음성 (소아, 유아) (Audio/Text Free conversation audio (Child, Infant))
- 음성/자면어 자유대화 음성 (일본어) (Audio/Text Free conversation audio (Japanese))
- 음성/자면어 전문분야 일용지 (Audio/Text Specialized field daily use)
- 음성/자면어 전문분야 일용지 (Audio/Text Specialized field daily use)

• 추가 데이터 안내

https://github.com/songys/AwesomeKorean_Data

<https://klue-benchmark.com/tasks>

The screenshot displays the KLUE benchmark website. The page has a navigation bar with links: Home, Participants, Paper, Task, Leaderboard, and Issue Report. The main content area is titled 'Tasks _ Individual Tasks' and lists several tasks with their descriptions:

- KLUE-TC (a.k.a. YNAT) - Topic Classification**
TC is a task to predict a topic of a news headline. The topic is one of 7 categories: politics, economy, society, culture, world, IT/science, and sports.
- KLUE-STTS - Semantic Textual Similarity**
STTS is a task which aims to predict the semantic similarity of two input sentences as a real value between 0 and 5. Note that we further binarized the prediction scores into two classes with a threshold score 3.0 (paraphrased or not) and evaluated with a classification metric.
- KLUE-NLI - Natural Language Inference**
NLI is a task to infer the relationship between a hypothesis sentence and a premise sentence. Given the premise, the model determines if the hypothesis is true (entailment), false (contradiction), or undetermined (neutral).
- KLUE-NER - Named Entity Recognition**
NER is a task to detect the boundaries of named entities in unstructured text and to classify the types. A named entity can be of one of predefined entity types such as person, location, organization, time expressions, quantities and monetary values.
- KLUE-RE - Relation Extraction**
RE is a task to identify semantic relations between entity pairs in a text. The relation is defined between an entity pair consisting of subject entity and object entity. The goal is then to pick an appropriate relationship between these two entities.
- KLUE-DP - Dependency Parsing**
DP is a task that aims at finding relational information among words. The goal is to predict a graph structure and a dependency label of an input sentence based on the dependency grammar.
- KLUE-MRC - Machine Reading Comprehension**
MRC is a task of evaluating model that can answer a question about a given text passage. Specifically, we formulate the task as a span prediction task, where the answer is a text segment (coined as spans) in the passage.
- KLUE-DST (a.k.a. WoS) - Dialogue State Tracking**
DST is a task to predict slot and value pairs (dialogue states) from a task-oriented dialogue. The potential pairs are predefined by a given task schema and knowledge base (KB).

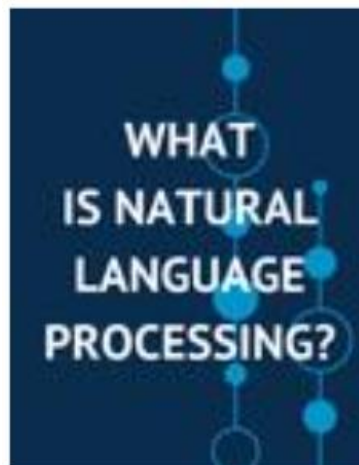
이 기 창 (2019),
한국어 임베딩
, 에이콘출판



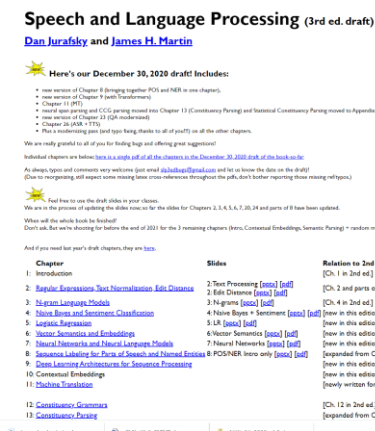
김기현(2019),
김기현의 자연어
처리 딥러닝 캠프,
한빛미디어



유원준(2021),
딥 러닝을 이용한
자연어 처리 입문
<https://wikidocs.net/book/2155>



Dan Jurafsky
and James H.
Martin(2020),
Speech and
Language
Processing (3rd
ed. draft)



전창욱, 최태균,
조중현(2019),
텐서플로와
머신러닝으로
시작하는
자연어 처리



- 조경현 : 딥러닝을 이용한 자연어 처리
- 강좌 링크:
<https://www.boostcourse.org/ai331>

학습 커리큘럼

01 Introduction

02 Basic ML : 지도학습

03 텍스트 분류

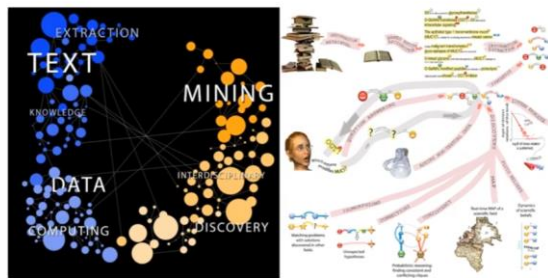
04 뉴럴 랭귀지 모델

05 신경망 기계 번역

06 Case Study

07 강의를 끝내며

- KoreaUniv DSBA
- 강좌링크 :
<https://www.youtube.com/watch?v=Ulnnl60pzkA&list=PLetSIH8YjlfVzHuSXtG4jAC2zbEAErXWm>



Lecture I: Introduction to Text Analytics

Pilsung Kang
School of Industrial Management Engineering
Korea University

- CS224n: Natural Language Processing with Deep Learning
- Winter 2019 강좌 링크:
<https://www.youtube.com/playlist?list=PLoROMvody4rOhcuXMZkNm7j3fVwBBY42z>



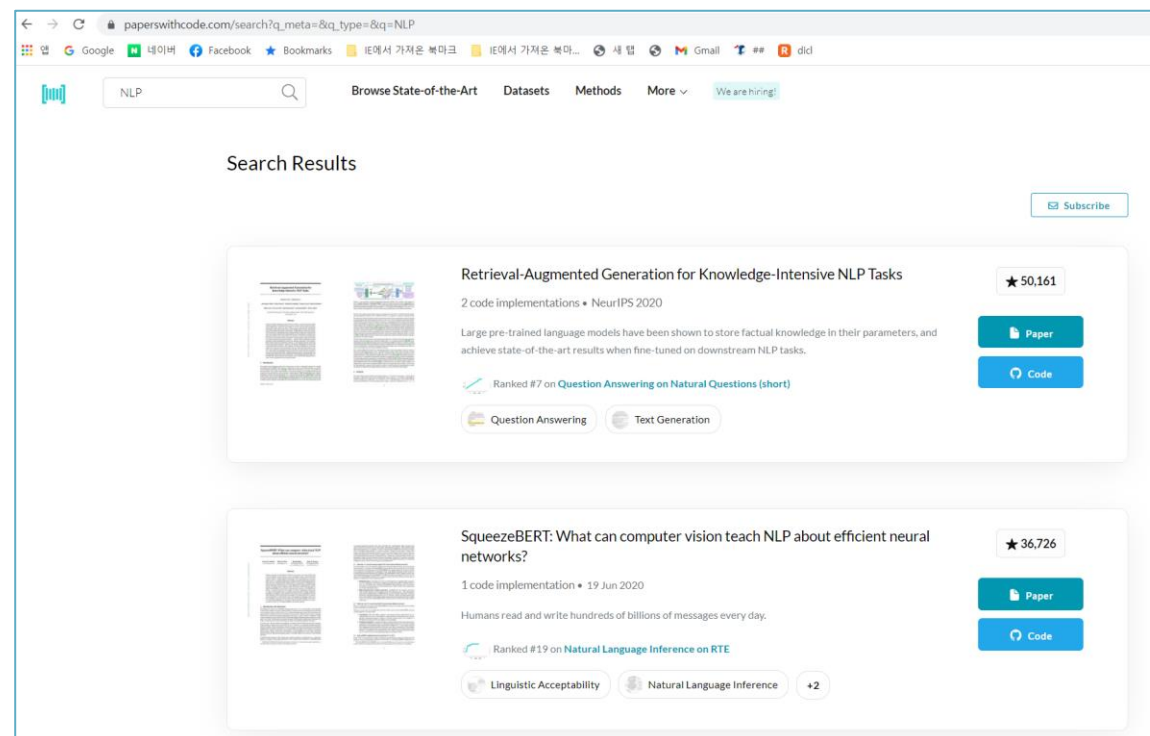
CS224N: Natural Language Processing with Deep Learning

- 김현중 : <https://lovit.github.io/about/>
- ratsgo : <https://ratsgo.github.io/>

- <https://huggingface.co/>
- <https://paperswithcode.com/>



Hugging Face



당신의 처음은 어땠나요?(라떼 한마디)

- 굉장히 신선했고, 우리말을 한층 더 사랑하게 되는 계기였습니다.
- 라떼는 fasttext, bert 하나 이해하려고 라이브러리 코드를 한줄한줄 다 뜯어봤었죠 :-)
- 전처리가 제일 눈물났습니다 $\pi\pi$
- 너무 오래전이라
- 2013 연구실에 들어갔고, 장비에 theano를 세팅하며 시작했습니다.
- 암흑기 - 따라가다가 - 요즘은 따라가기 힘들
- 아직도 저는 한글로 원하는 결과를 얻지 못하였습니다.
- 라떼는 말이야, 파파고 같은 번역기 똑딱 만들 수 있을 줄 알았는데, 공부해보니 아주 많은 기술과 지식이 필요하더라구요
- 스터디하면서 책 한권씩 떼는 즐거움이 있었죠!
- 자료도 거의 없었고 구한 자료도 이해하기 어려웠다
- 자연어처리에 흥미를 가지고 본 지 몇 년이 된 것 같은데 아직도 잘 모릅니다. 계속 공부해야하는 것 같아요..ㅎㅎ
- 여전이 처음이죠
- 텍스트 분석이라고 사장이 우습게 생각했지
- 전 현재 보유한 데이터 양은 충분하며 전처리도 별로 필요없다고 믿는 분과 연구를 했었습니다. 결론적으로 모두 그렇지 않다는 결론을 이끌어냈지만, 그 과정에서 협의나 협업이 안되는 경험을 겪으며 실험정신으로 무장한 동료와 환경이 중요하다는 생각을 하게 되었네요..

- 자연어처리와 관련된 다양한 행사 혹은 컨퍼런스 공유 및 진행
- 온라인 스터디 모임이나 정보 공유가 자주 있었으면 좋겠습니다. 만약 이미 있었다면.. 제가 잘 모르고 있었네요..
- 실제 사람들의 요구사항을 엔트로피를 이용하여 task def에 녹여내는 이야기를 추가해 주셨으면 합니다.
- 홈페이지가 어떻게 됐었죠? 하하하 ㅠㅠ
- 파이썬이 세상의 전부입니까?
- 지속적인 오픈된 세미나
- 좋은 커뮤니티로 성장하길!
- 오래오래 꾸준히 지속되었으면 합니다
- 항상 응원합니다!
- 내년에도 잘 부탁드립니다~
- 흥하시기를 기원합니다 ^^
- 올해도 기대하고 있습니다. 화이팅!!!
- 존재로 감사함

TIOBE Index						PYPL Index (Worldwide)				
Aug 2021 ▲	Aug 2020 ◆	Change ◆	Programming language ◆	Ratings ◆	Change ◆	Aug 2021 ▲	Change ◆	Programming language ◆	Share ◆	Trends ◆
1	1		C	12.57%	-4.41%	1		Python	29.93 %	-2.2 %
2	3	↑	Python	11.86%	+2.17%	2		Java	17.78 %	+1.2 %
3	2	↓	Java	10.43%	-4.00%	3		JavaScript	8.79 %	+0.6 %
4	4		C++	7.36%	+0.52%	4		C#	6.73 %	+0.2 %
5	5		C#	5.14%	+0.46%	5	↑	C/C++	6.45 %	+0.7 %
6	6		Visual Basic	4.67%	+0.01%	6	↓	PHP	5.76 %	-0.0 %
7	7		JavaScript	2.95%	+0.07%	7		R	3.92 %	-0.1 %
8	9	↑	PHP	2.19%	-0.05%	8		Objective-C	2.26 %	-0.3 %
9	14	↑ ↑	Assembly language	2.03%	+0.99%	9	↑	TypeScript	2.11 %	+0.2 %
10	10		SQL	1.47%	+0.02%	10	↓	Swift	1.96 %	-0.3 %
11	18	↑ ↑	Groovy	1.36%	+0.59%	11	↑	Kotlin	1.81 %	+0.3 %
12	17	↑ ↑	Classic Visual Basic	1.23%	+0.41%	12	↓	Matlab	1.48 %	-0.4 %
13	42	↑ ↑	Fortran	1.14%	+0.83%	13		Go	1.29 %	-0.2 %
14	8	↓ ↓	R	1.05%	-1.75%	14	↑ ↑	Rust	1.21 %	+0.2 %
15	15		Ruby	1.01%	-0.03%	15	↓	VBA	1.16 %	-0.1 %
16	12	↓ ↓	Swift	0.98%	-0.44%	16	↓	Ruby	1.02 %	-0.1 %
17	16	↓	MATLAB	0.98%	+0.11%	17		Scala	0.79 %	-0.1 %
18	11	↓ ↓	Go	0.90%	-0.52%	18	↑	Ada	0.77 %	+0.2 %
19	36	↑ ↑	Prolog	0.80%	+0.41%	19	↓	Visual Basic	0.75 %	+0.0 %
20	13	↓ ↓	Perl	0.78%	-0.33%	20		Dart	0.68 %	+0.2 %
						21		Lua	0.58 %	+0.1 %

<https://statisticstimes.com/tech/top-computer-languages.php>



Q&A

- 감성 인식 데이터에 대한 요구가 많았는데 감정은 객관적으로 측정할 수 있는 방법이 있을까요?
- 금칙어 사전 필요한가요? 오히려 윤리적 문제가 생기지 않을까요?
- 그들만의 리그라는 평이 있었는데요. 프로그래머가 아닌 분들이 자연어처리(텍스트 및 음성) 분야에서 할 수 있는 일들은 어떤 것이 있을까요?
- 데이터는 많아졌는데 편중되어 있다고 느끼는 분들도 많은 것 같습니다. 처음 텍스트 데이터를 접하는 분들에게 할 수 있는 조언은 무엇이 있을까요?
- 처음으로 돌아가서 자연어처리에서 가장 중요하게 생각하시는 부분은 무엇인가요

여러분의 질문을 남겨주세요. 답해 드립니다.

- shorturl.at/dEW45



감사합니다