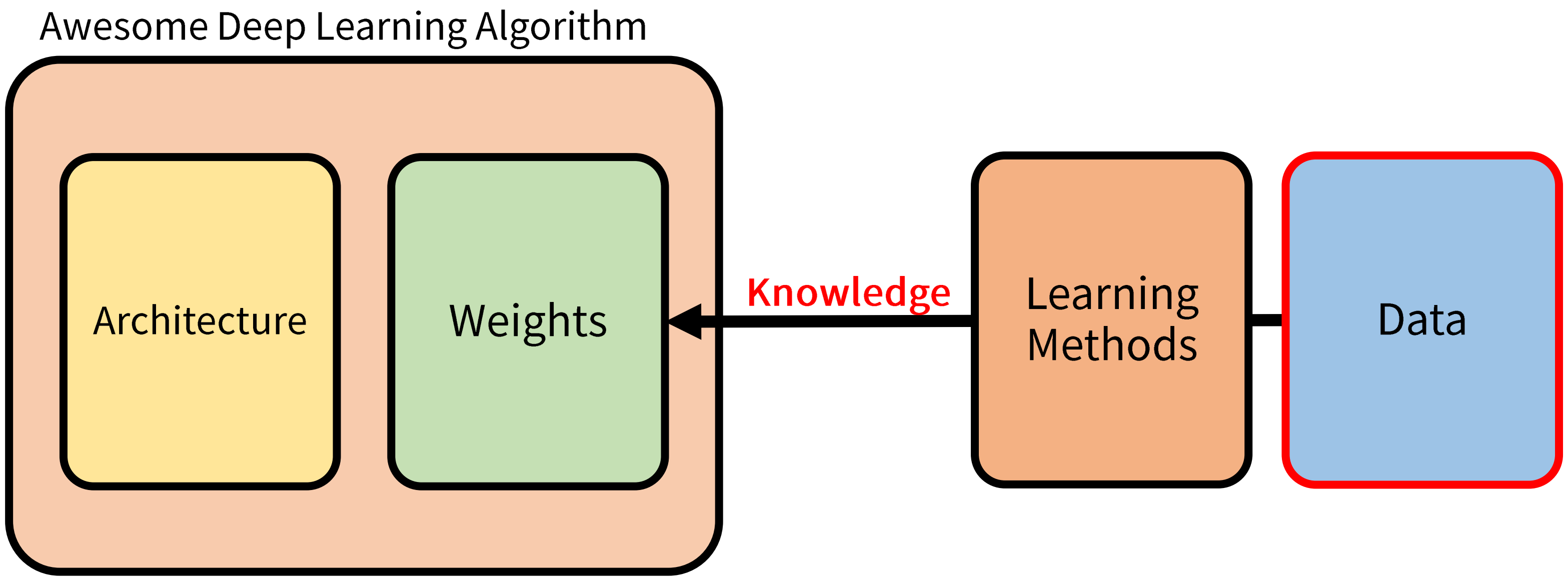
Chapter 04. 자연어처리 (Natural Language Processing)

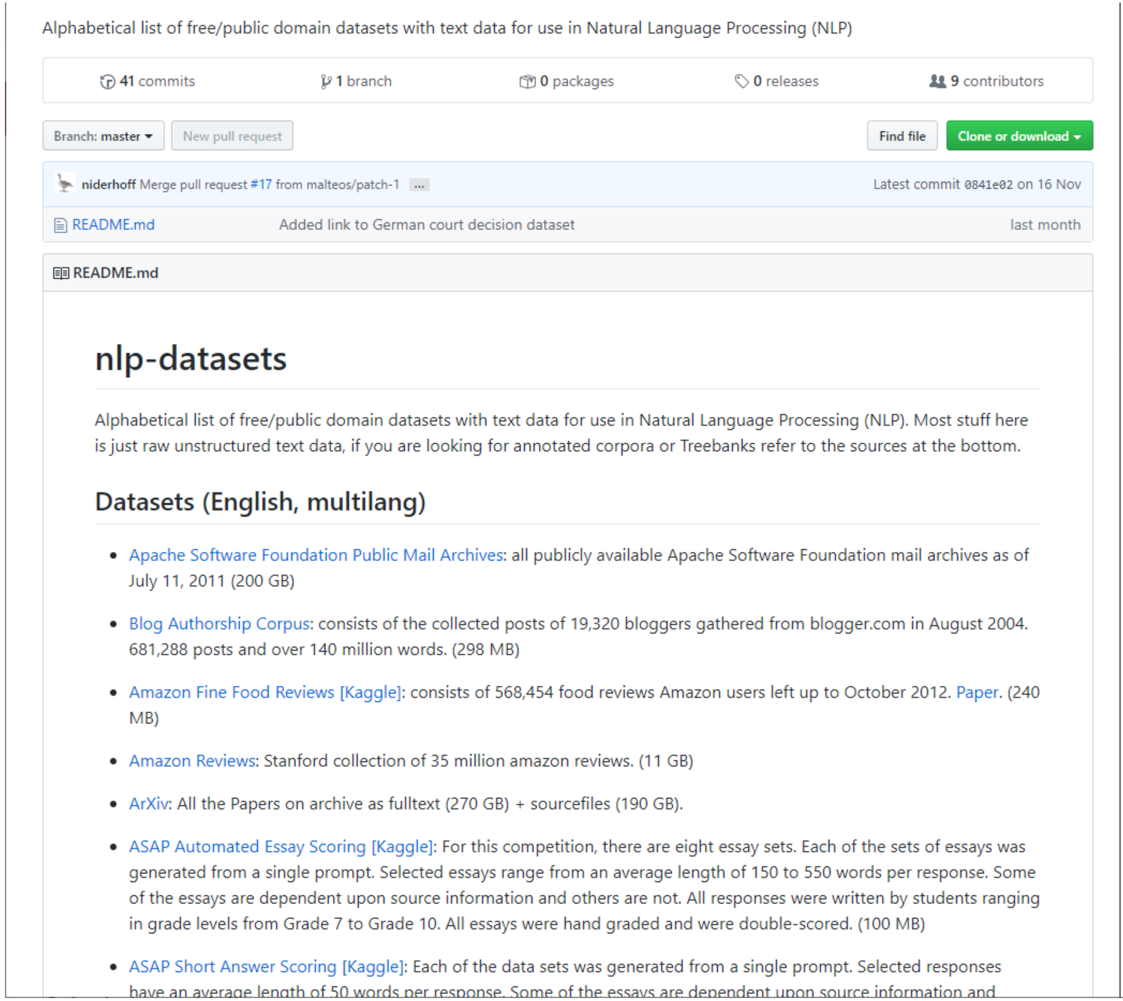
응용하기 좋은 데이터셋 소개

데이터셋의 중요성



아무리 잘 짜여진 딥러닝 알고리즘도, '지식'이 없으면 무용지물이다. 그 '지식'은 데이터셋으로 부터 나온다.

데이터셋 참고



Alphabetical list of free/public domain datasets with text data for use in Natural Language Processing (NLP)

niderhoff Merge pull request #17 from malteos/patch-1 … Latest commit 0841e02 on 16 Nov

README.md Added link to German court decision dataset last month

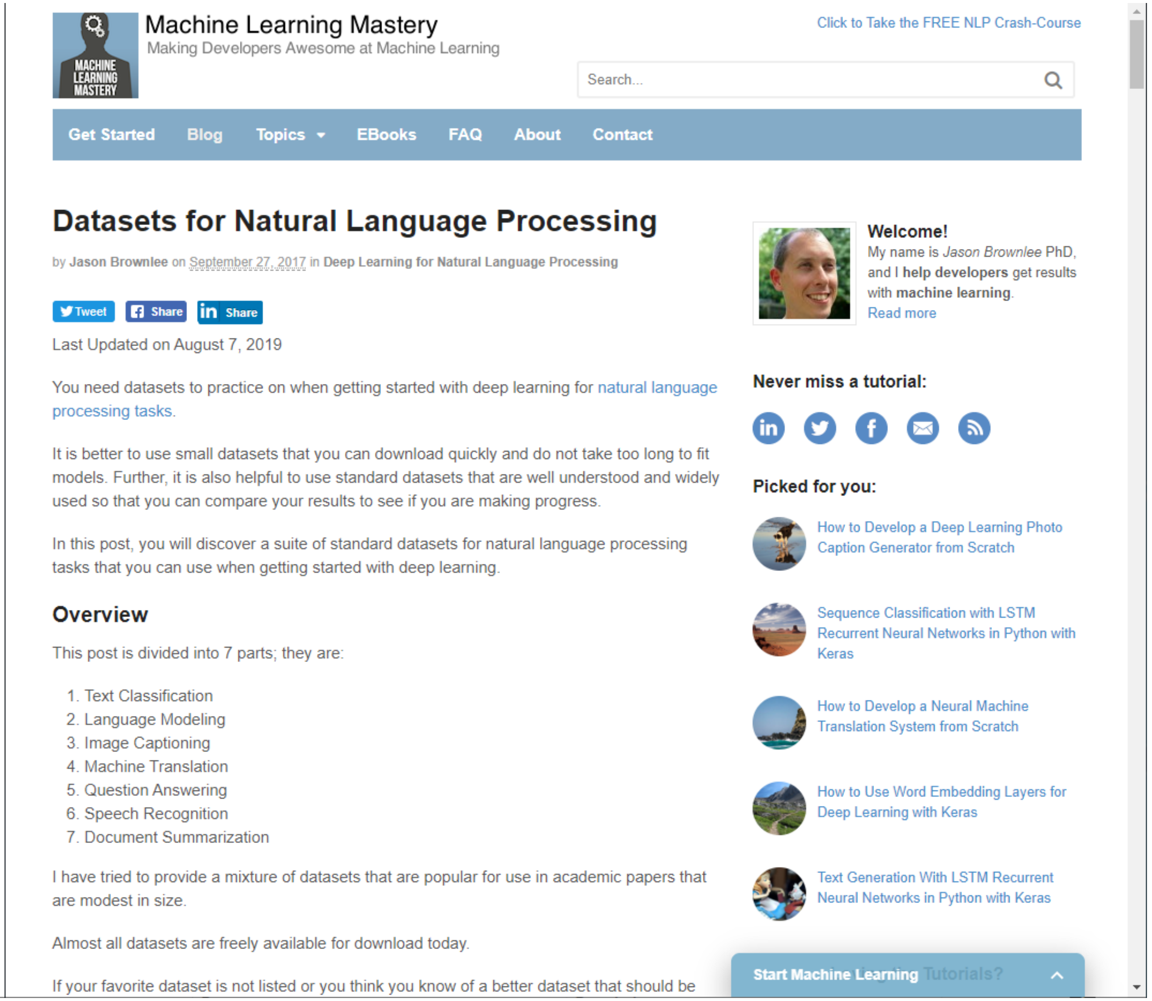
README.md

nlp-datasets

Alphabetical list of free/public domain datasets with text data for use in Natural Language Processing (NLP). Most stuff here is just raw unstructured text data, if you are looking for annotated corpora or Treebanks refer to the sources at the bottom.

Datasets (English, multilang)

- Apache Software Foundation Public Mail Archives: all publicly available Apache Software Foundation mail archives as of July 11, 2011 (200 GB)
- Blog Authorship Corpus: consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. 681,288 posts and over 140 million words. (298 MB)
- Amazon Fine Food Reviews [Kaggle]: consists of 568,454 food reviews Amazon users left up to October 2012. Paper. (240 MB)
- Amazon Reviews: Stanford collection of 35 million amazon reviews. (11 GB)
- ArXiv: All the Papers on archive as fulltext (270 GB) + sourcefiles (190 GB).
- ASAP Automated Essay Scoring [Kaggle]: For this competition, there are eight essay sets. Each of the sets of essays was generated from a single prompt. Selected essays range from an average length of 150 to 550 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students ranging in grade levels from Grade 7 to Grade 10. All essays were hand graded and were double-scored. (100 MB)
- ASAP Short Answer Scoring [Kaggle]: Each of the data sets was generated from a single prompt. Selected responses have an average length of 50 words per response. Some of the essays are dependent upon source information and



Machine Learning Mastery
Making Developers Awesome at Machine Learning

Datasets for Natural Language Processing

by Jason Brownlee on September 27, 2017 in Deep Learning for Natural Language Processing

Last Updated on August 7, 2019

You need datasets to practice on when getting started with deep learning for natural language processing tasks.

It is better to use small datasets that you can download quickly and do not take too long to fit models. Further, it is also helpful to use standard datasets that are well understood and widely used so that you can compare your results to see if you are making progress.

In this post, you will discover a suite of standard datasets for natural language processing tasks that you can use when getting started with deep learning.

Overview

This post is divided into 7 parts; they are:

1. Text Classification
2. Language Modeling
3. Image Captioning
4. Machine Translation
5. Question Answering
6. Speech Recognition
7. Document Summarization

I have tried to provide a mixture of datasets that are popular for use in academic papers that are modest in size.

Almost all datasets are freely available for download today.

If your favorite dataset is not listed or you think you know of a better dataset that should be

Welcome!
My name is *Jason Brownlee* PhD, and I **help developers** get results with **machine learning**.
Read more

Never miss a tutorial:

Picked for you:

How to Develop a Deep Learning Photo Caption Generator from Scratch

Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras

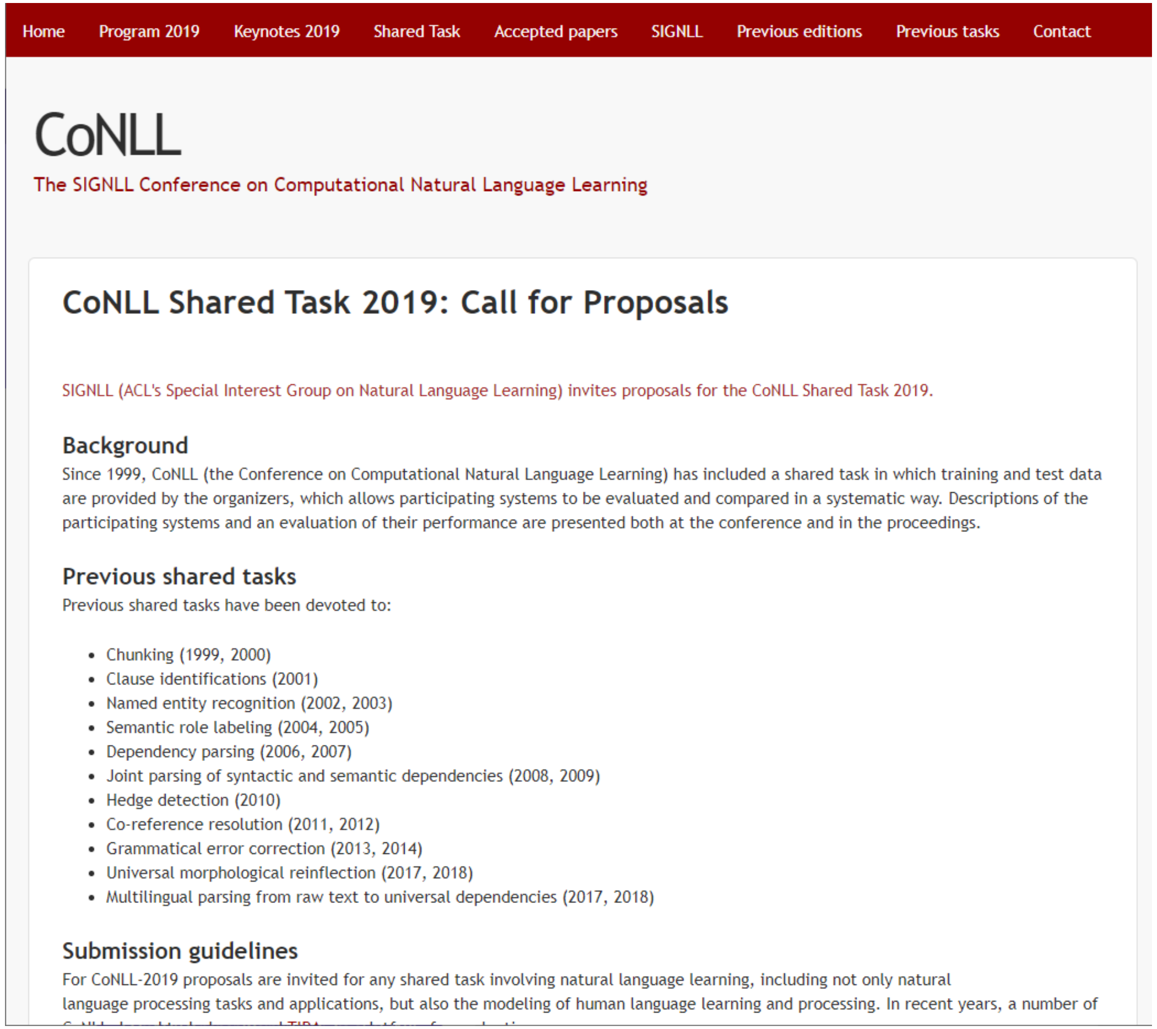
How to Develop a Neural Machine Translation System from Scratch

How to Use Word Embedding Layers for Deep Learning with Keras

Text Generation With LSTM Recurrent Neural Networks in Python with Keras

https://github.com/niderhoff/nlp-datasets https://machinelearningmastery.com/datasets-natural-language-processing/

CoNLL Shared Task



CoNLL

The SIGNLL Conference on Computational Natural Language Learning

CoNLL Shared Task 2019: Call for Proposals

SIGNLL (ACL's Special Interest Group on Natural Language Learning) invites proposals for the CoNLL Shared Task 2019.

Background

Since 1999, CoNLL (the Conference on Computational Natural Language Learning) has included a shared task in which training and test data are provided by the organizers, which allows participating systems to be evaluated and compared in a systematic way. Descriptions of the participating systems and an evaluation of their performance are presented both at the conference and in the proceedings.

Previous shared tasks

Previous shared tasks have been devoted to:

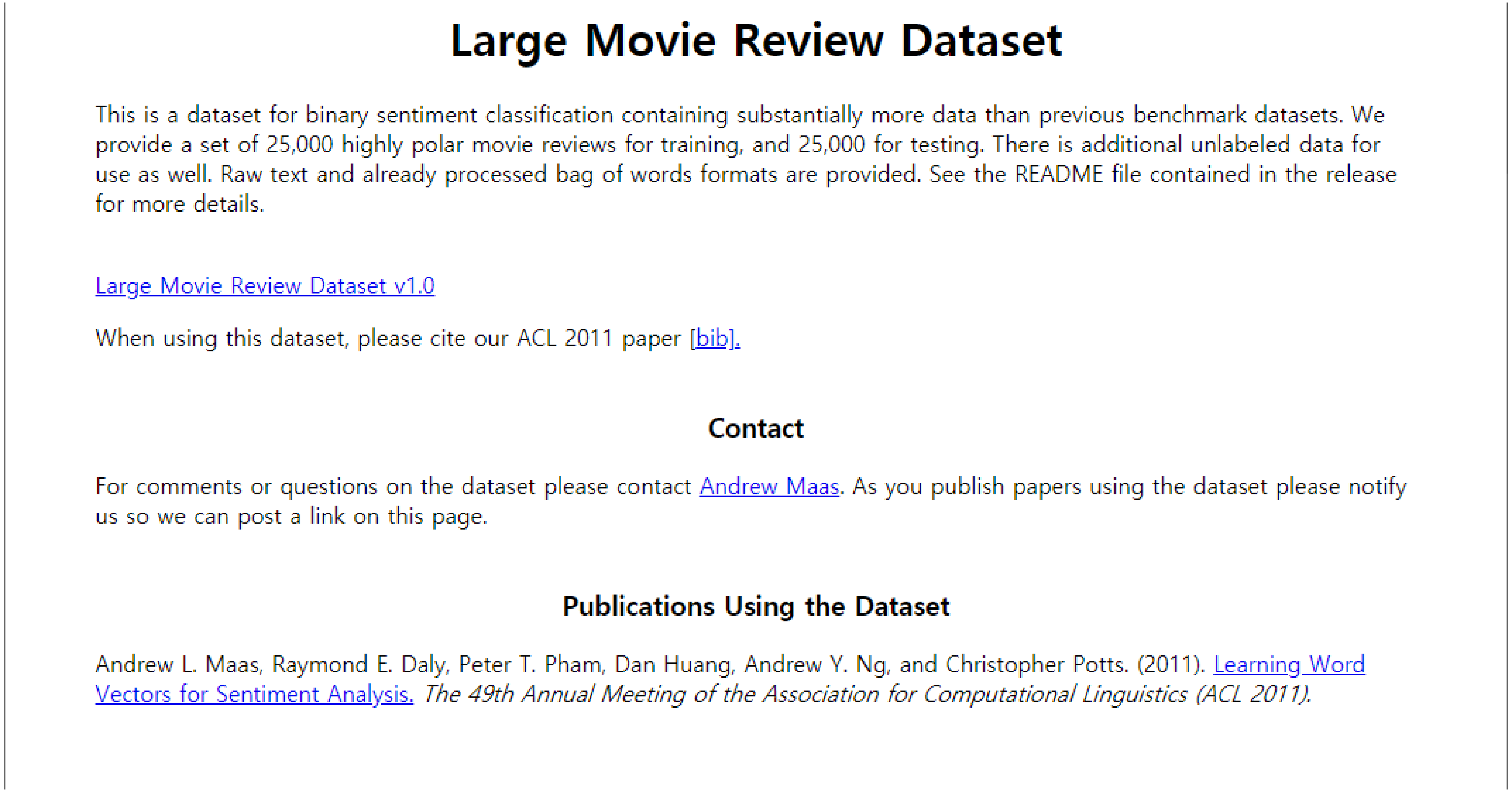
- Chunking (1999, 2000)
- Clause identifications (2001)
- Named entity recognition (2002, 2003)
- Semantic role labeling (2004, 2005)
- Dependency parsing (2006, 2007)
- Joint parsing of syntactic and semantic dependencies (2008, 2009)
- Hedge detection (2010)
- Co-reference resolution (2011, 2012)
- Grammatical error correction (2013, 2014)
- Universal morphological reinflection (2017, 2018)
- Multilingual parsing from raw text to universal dependencies (2017, 2018)

Submission guidelines

For CoNLL-2019 proposals are invited for any shared task involving natural language learning, including not only natural language processing tasks and applications, but also the modeling of human language learning and processing. In recent years, a number of

https://www.conll.org/2019-shared-task

Stanford Large Movie Review Dataset



Large Movie Review Dataset

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided. See the README file contained in the release for more details.

Large Movie Review Dataset v1.0

When using this dataset, please cite our ACL 2011 paper [bib].

Contact

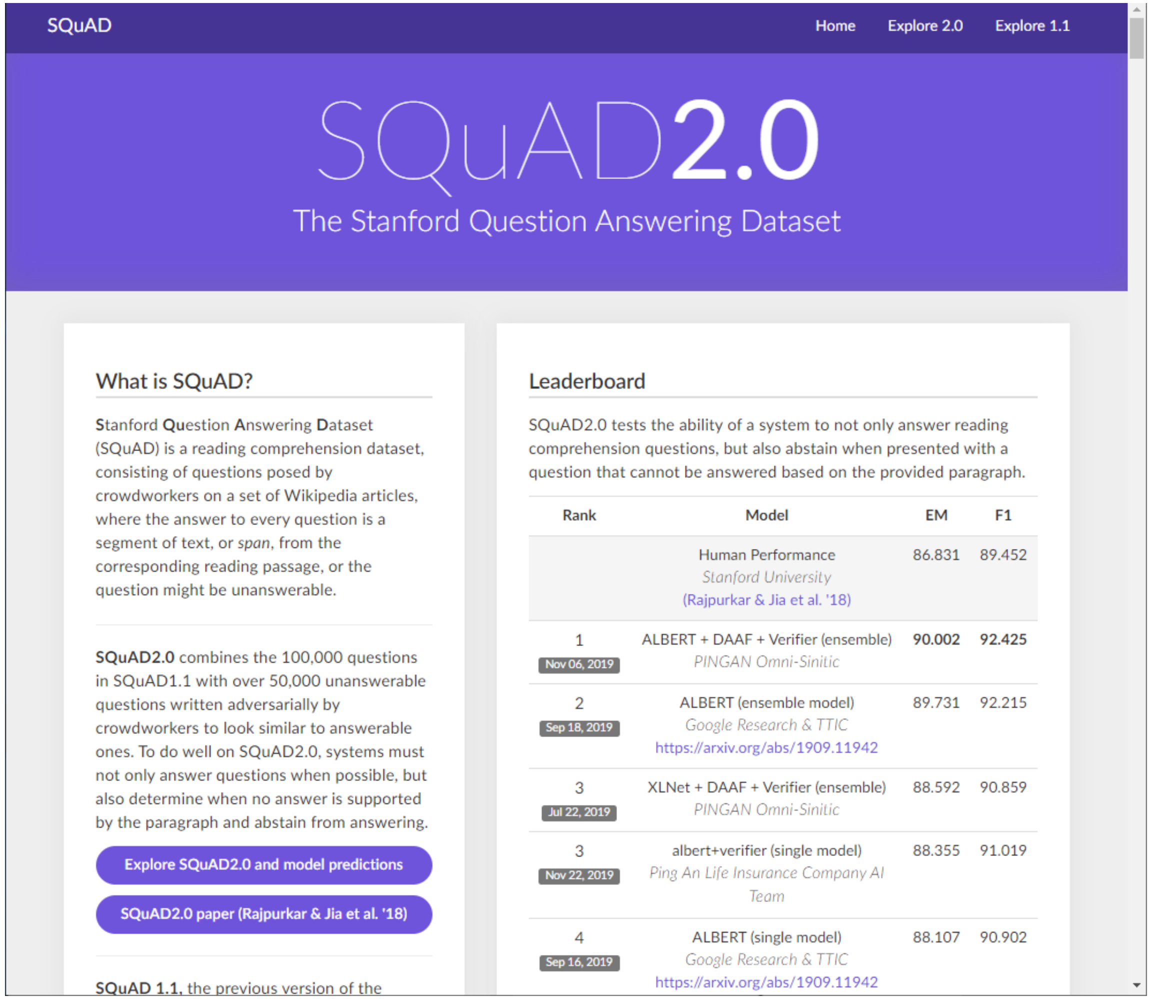
For comments or questions on the dataset please contact Andrew Maas. As you publish papers using the dataset please notify us so we can post a link on this page.

Publications Using the Dataset

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*

http://ai.stanford.edu/~amaas/data/sentiment/

SQuAD2.0



SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

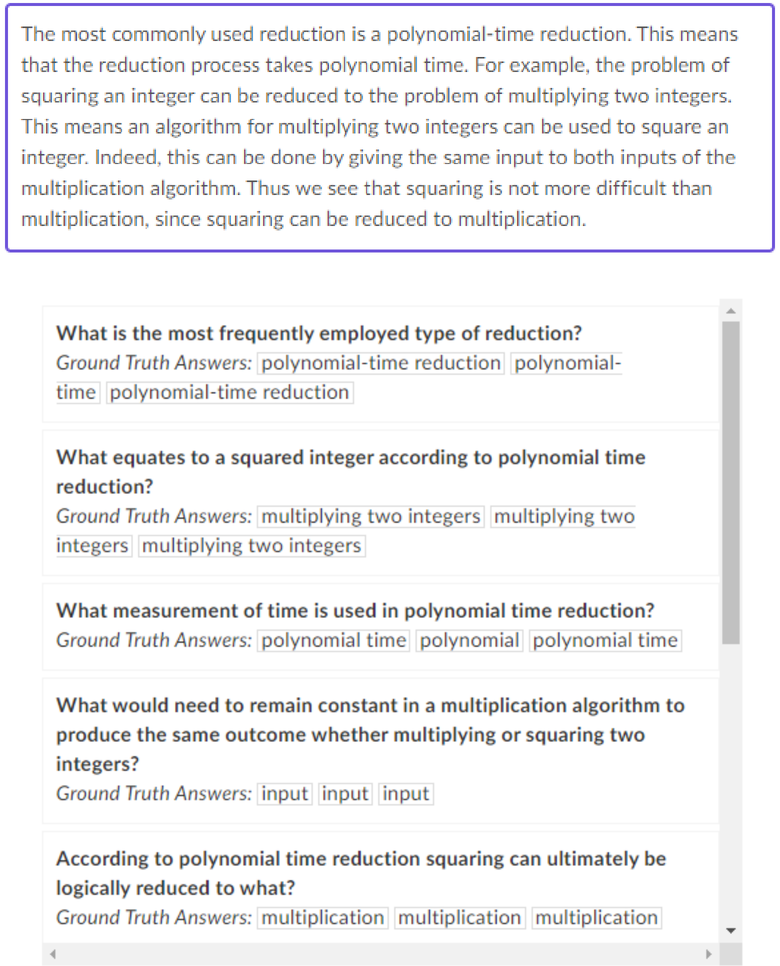
SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**90.002**	**92.425**
2 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
3 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
3 Nov 22, 2019	albert+verifier (single model) *Ping An Life Insurance Company AI Team*	88.355	91.019
4 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902



The most commonly used reduction is a polynomial-time reduction. This means that the reduction process takes polynomial time. For example, the problem of squaring an integer can be reduced to the problem of multiplying two integers. This means an algorithm for multiplying two integers can be used to square an integer. Indeed, this can be done by giving the same input to both inputs of the multiplication algorithm. Thus we see that squaring is not more difficult than multiplication, since squaring can be reduced to multiplication.

What is the most frequently employed type of reduction?
Ground Truth Answers: polynomial-time reduction polynomial-time polynomial-time reduction

What equates to a squared integer according to polynomial time reduction?
Ground Truth Answers: multiplying two integers multiplying two integers multiplying two integers

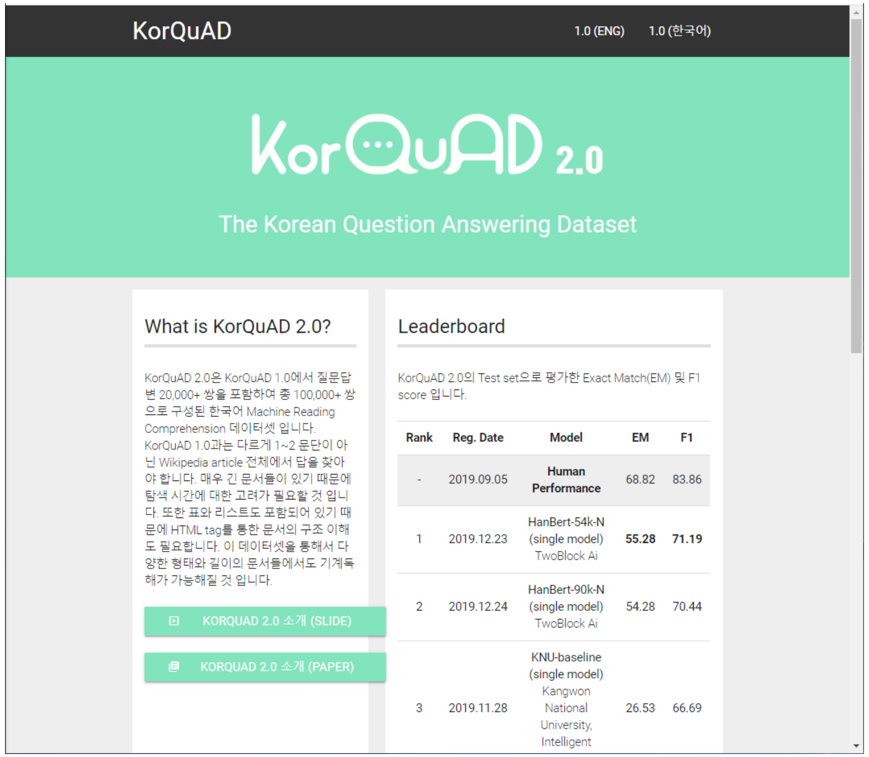
What measurement of time is used in polynomial time reduction?
Ground Truth Answers: polynomial time polynomial polynomial time

What would need to remain constant in a multiplication algorithm to produce the same outcome whether multiplying or squaring two integers?
Ground Truth Answers: input input input

According to polynomial time reduction squaring can ultimately be logically reduced to what?
Ground Truth Answers: multiplication multiplication multiplication

https://rajpurkar.github.io/SQuAD-explorer/

KorQuAD2.0 (2.1)



KorQuAD 1.0 (ENG) 1.0 (한국어)

KorQuAD 2.0

The Korean Question Answering Dataset

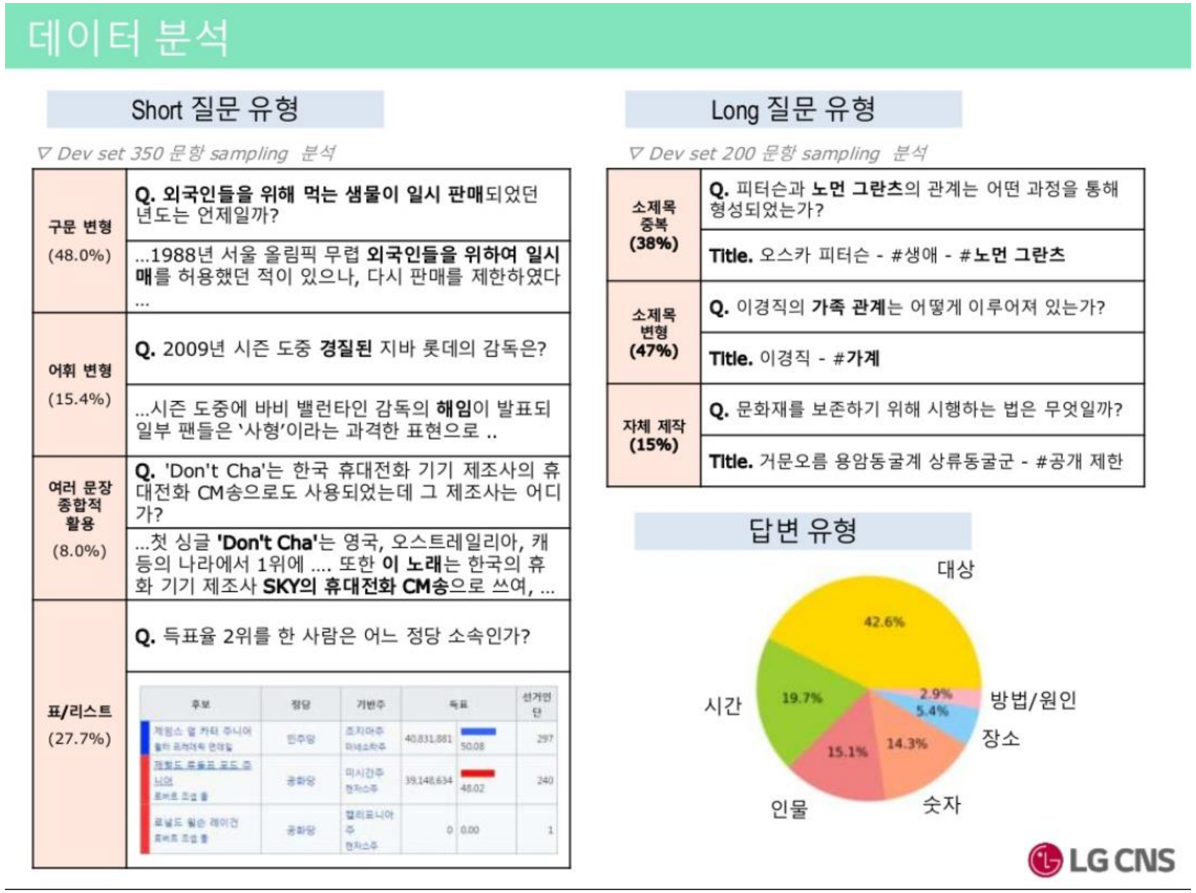
What is KorQuAD 2.0?

KorQuAD 2.0은 KorQuAD 1.0에서 질문답변 20,000+ 쌍을 포함하여 총 100,000+ 쌍으로 구성된 한국어 Machine Reading Comprehension 데이터셋 입니다. KorQuAD 1.0과는 다르게 1~2 문단이 아닌 Wikipedia article 전체에서 답을 찾아야 합니다. 매우 긴 문서들이 있기 때문에 탐색 시간에 대한 고려가 필요할 것 입니다. 또한 표와 리스트도 포함되어 있기 때문에 HTML tag를 통한 문서의 구조 이해도 필요합니다. 이 데이터셋을 통해서 다양한 형태와 길이의 문서들에서도 기계독해가 가능해질 것 입니다.

Leaderboard

KorQuAD 2.0의 Test set으로 평가한 Exact Match(EM) 및 F1 score 입니다.

Rank	Reg. Date	Model	EM	F1
-	2019.09.05	**Human Performance**	68.82	83.86
1	2019.12.23	HanBert-54k-N (single model) TwoBlock Ai	**55.28**	**71.19**
2	2019.12.24	HanBert-90k-N (single model) TwoBlock Ai	54.28	70.44
3	2019.11.28	KNU-baseline (single model) Kangwon National University, Intelligent	26.53	66.69



데이터 분석

Short 질문 유형

▽ Dev set 350 문항 sampling 분석

구문 변형 (48.0%)	**Q. 외국인들을 위해 먹는 샘물이 일시 판매**되었던 년도는 언제일까? ...1988년 서울 올림픽 무렵 **외국인들을 위하여 일시 매**를 허용했던 적이 있으나, 다시 판매를 제한하였다 ...
어휘 변형 (15.4%)	**Q.** 2009년 시즌 도중 **경질된** 지바 롯데의 감독은? ...시즌 도중에 바비 밸런타인 감독의 **해임**이 발표되 일부 팬들은 '사형'이라는 과격한 표현으로 ..
여러 문장 종합적 활용 (8.0%)	**Q.** 'Don't Cha'는 한국 휴대전화 기기 제조사의 휴대전화 CM송으로도 사용되었는데 그 제조사는 어디가? ...첫 싱글 **'Don't Cha'**는 영국, 오스트레일리아, 캐 등의 나라에서 1위에 .... 또한 **이 노래**는 한국의 휴화 기기 제조사 **SKY의 휴대전화 CM송**으로 쓰여, ...
표/리스트 (27.7%)	**Q.** 득표율 2위를 한 사람은 어느 정당 소속인가?

후보	정당	기반주	득표		선거인단
제임스 얼 카터 주니어 월터 프레데릭 먼데일	민주당	조지아주 미네소타주	40,831,881	50.08	297
제럴드 루돌프 포드 주니어 로버트 조셉 돌	공화당	미시간주 캔자스주	39,148,634	48.02	240
로널드 윌슨 레이건 로버트 조셉 돌	공화당	캘리포니아주 캔자스주	0	0.00	1

Long 질문 유형

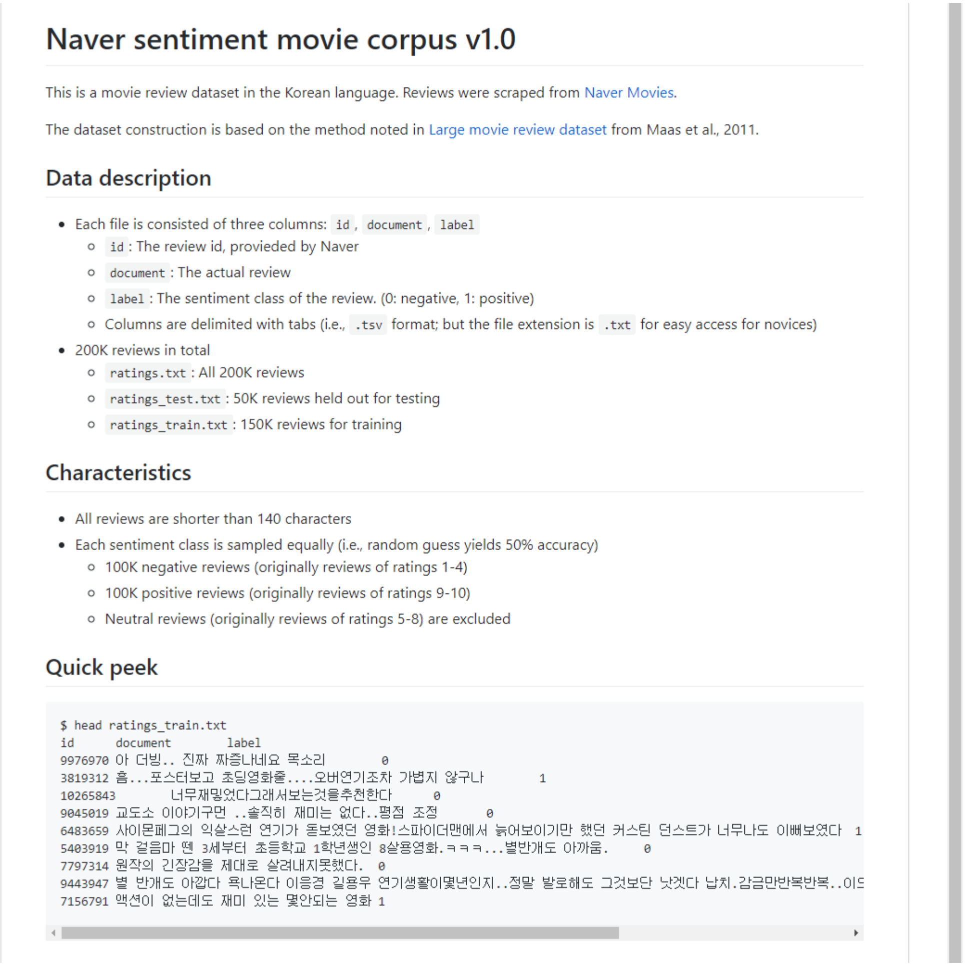
▽ Dev set 200 문항 sampling 분석

소제목 중복 **(38%)**	**Q.** 피터슨과 **노먼 그란츠**의 관계는 어떤 과정을 통해 형성되었는가? **Title.** 오스카 피터슨 - #생애 - **#노먼 그란츠**
소제목 변형 **(47%)**	**Q.** 이경직의 **가족 관계**는 어떻게 이루어져 있는가? **Title.** 이경직 - **#가계**
자체 제작 **(15%)**	**Q.** 문화재를 보존하기 위해 시행하는 법은 무엇일까? **Title.** 거문오름 용암동굴계 상류동굴군 - #공개 제한

답변 유형

https://korquad.github.io/

Naver Sentiment Movie Corpus



Naver sentiment movie corpus v1.0

This is a movie review dataset in the Korean language. Reviews were scraped from Naver Movies.

The dataset construction is based on the method noted in Large movie review dataset from Maas et al., 2011.

Data description

- Each file is consisted of three columns: `id`, `document`, `label`
  - `id`: The review id, provieded by Naver
  - `document`: The actual review
  - `label`: The sentiment class of the review. (0: negative, 1: positive)
  - Columns are delimited with tabs (i.e., `.tsv` format; but the file extension is `.txt` for easy access for novices)
- 200K reviews in total
  - `ratings.txt`: All 200K reviews
  - `ratings_test.txt`: 50K reviews held out for testing
  - `ratings_train.txt`: 150K reviews for training

Characteristics

- All reviews are shorter than 140 characters
- Each sentiment class is sampled equally (i.e., random guess yields 50% accuracy)
  - 100K negative reviews (originally reviews of ratings 1-4)
  - 100K positive reviews (originally reviews of ratings 9-10)
  - Neutral reviews (originally reviews of ratings 5-8) are excluded

Quick peek

```
$ head ratings_train.txt
id      document        label
9976970 아 더빙.. 진짜 짜증나네요 목소리        0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나        1
10265843        너무재밓었다그래서보는것을추천한다      0
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정       0
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다  1
5403919 막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.     0
7797314 원작의 긴장감을 제대로 살려내지못했다.  0
9443947 별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낫겟다 납치.감금만반복반복..이
7156791 액션이 없는데도 재미 있는 몇안되는 영화 1
```

https://github.com/e9t/nsmc