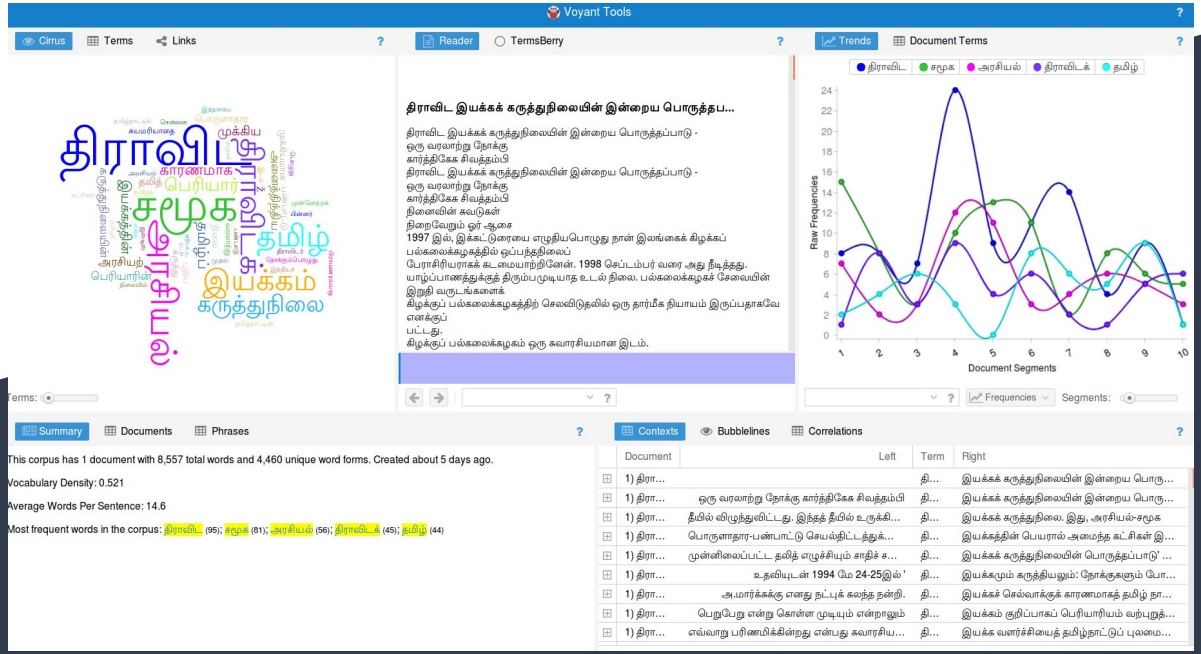
திறந்த தமிழ் தரவுத் தொகுப்புகளின் இன்றைய நிலை
(Current State of Open Tamil Datasets)

சனவரி 18, 2020



திறந்த/கட்டற்ற அணுக்கம்
Open/Free Access

சுதந்திரங்கள் (தகுந்த உரிமைக் குறிப்புகளோடு (with due attribution))

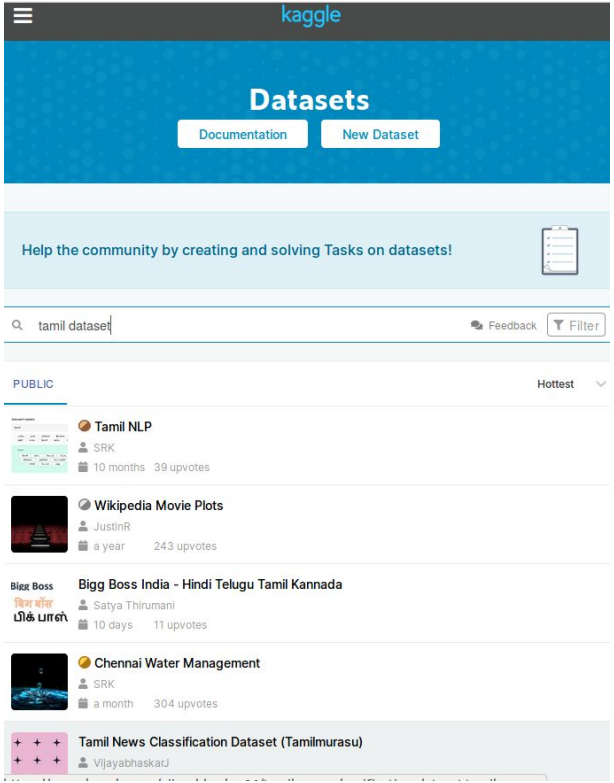
- பயன்படுத்தலாம் - use
- பகிரலாம் - share and redistribute
- ஒன்றுடன் மற்றொன்றைக் கலக்கலாம், மாற்றலாம், மேம்படுத்தலாம் - remix, transform, and build upon the material

திறந்த/கட்டற்ற உரிமங்கள்

- Public Domain = CC0 = பொது உரிமம்
- Creative Commons Share Alike = படைப்பாக்க பொதுமம் - அதே மாதிரிப் பகிர்தல்
- Creative Commons Share Alike - Non Commercial = படைப்பாக்க பொதுமம் - அதே மாதிரிப் பகிர்தல் (இலாப நோக்கமற்ற)

தற்போது கிடைக்கும் தரவு வகைகள்

- எழுத்துப் பிரதிகள்/பனுவல்கள் - Plain Text Data
- கட்டமைக்கப்பட்ட எழுத்துப் பிரதிகள் - Structured Text
- குறியிடப்பட்ட எழுத்துத் தரவுகள் - Annotated Text Data
  - Grammar annotated text
  - Named Entity annotated
- சொல்வலைகள், கட்டமைக்கப்பட்ட அகராதிகள் - Word Net, Word Annotations, Structured Dictionary
- இணையான ஒலி - எழுத்துத் தரவுகள் - Audio with Transcripts
- இணை மொழித் தொகுப்புகள் - Parallel Corpus
- குறியிடப்பட்ட படத் தரவுகள் - Annotated Image Data
- துறை சார்ந்த தரவுத் தொகுப்புகள் - Subject based Data Sets (labelled data)



முழுமையான திறந்த பிரதிகள்/தரவுகள்

பிரதி (Text)

- தமிழ் விக்கியூடகத் தரவுகள் (text, csv)
- மதுரைத் திட்ட பிரதிகள் (html வடிவம்)
- கணியம் அறக்கட்டளை வெளியீடுகள்
- Charles University Tamil Dependency Treebank v0.1
- தமிழ் சொல் வலை (அண்ணாமலைப் பல்கலைக்கழகம், மும்பை இ.தொ. பயிலகம்)

விக்கிப்பீடியா
கட்டற்ற கலைக்களஞ்சியம்

ஒலி (Audio + Text)

- மொட்சிலா பொதுக்குரல் (ஒலி - எழுத்து)

Common Voice
moz://a

காணொளி/படங்கள்/ஒலிப் பதிவுகள் (Video, Image, Aud..,

- விக்கிப் பொதுவகம் (Wikimedia Commons)
- நூலக நிறுவன வெளியீடுகள் (Noolaham Foundations Own Works)

பகிரப்பட முடியாத தரவுகள்

இந்திய மொழிகளின் தரவுகளுக்கான கூட்டமைப்புத் தரவுகள்

- Cannot Redistribute (i.e not include in your applications or teaching material)
- Need to be an Indian resident!

அண்ணாமலைப் பல்கலைக்கழகத் தரவுகள்

- “User shall not publish, retransmit, display, redistribute, reproduce or commercially exploit the Data in any form,”

கீழ் கண்ட தரவுகள் உட்பட

- AUKBC Tamil Part-of-Speech (POS) Corpus
- AUKBC NER Annotated Corpus
- Linguistic Data Consortium for Indian Languages - Speech Data

உரிமை தெளிவற்ற தரவுகள்

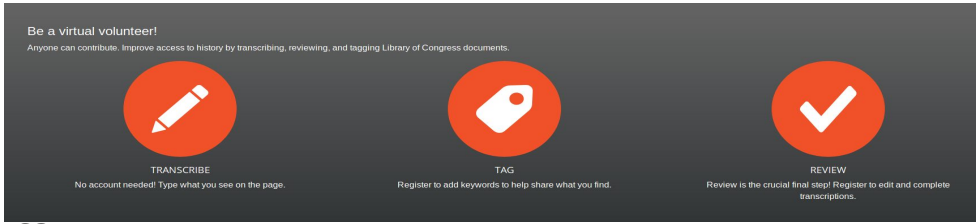
- தினமலர் செய்தித் தரவுத்தொகுதி (1.9 மில்லியன் பதிவுகள்)
- தினமுரசு செய்தித் தரவுத்தொகுதி (127 000 பதிவுகள்)
- தமிழ் இந்து செய்தித் தரவுத்தொகுதி (6500 பதிவுகள்)
- Charles University இணை மொழித் தொகுப்பு
- அரச வலைத்தளங்களின் இணை மொழித் தொகுப்புகள்

தரவுத்தொகுதிகளைத் திறந்த அணுக்கத்தில் தந்துள்ளார்கள், ஆனால் உள்ளடக்க உரிமையாளர்களிடம் உரிமை பெறப்பட்டதா என்பதில் தெளிவில்லை.

சில பயன்பாடுகளுக்கு இவை ஏற்புடையதாக இருக்கலாம்.

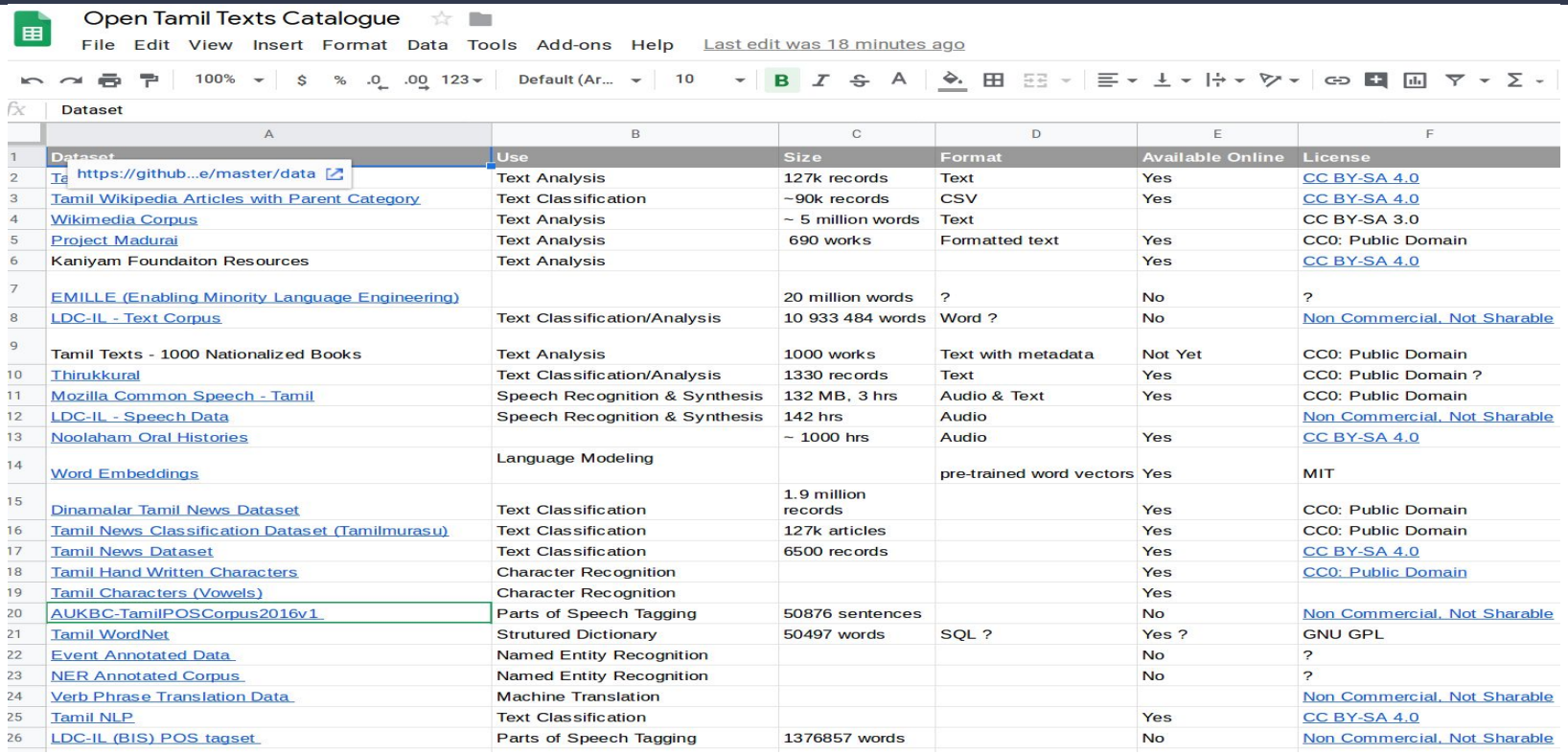
தரவுத்தொகுதிகளை உருவாக்கல்

- பயன்படுத்தக் கூடிய வடிவில் வெளியிடல்
  - Plain text வடிவில் வெளியிடல் (எ.கா freetamilbooks.com, projectmadurai.org)
  - CSV வடிவில் வெளியிடல் (Tamil Word Net)
- கூடிய கட்டமைப்புக் கொண்டதாக மாற்றல் - structuring existing data
  - தமிழ் அகராதிகள் (எ.கா தமிழ் விக்சனரி)
  - தமிழ் இலக்கியத் தொகுப்பு
- கூட்டு உருவாக்கம் (crowdsourcing & partnerships)
  - பொதுக் குரல்
  - விக்கி மூலம்
  - annotations
  - ஒலி நூற்கள்



- parallel corpus crawling
- உரிமங்கள் தொடர்பான விழிப்புணர்வு
- அரச நிறுவனங்களிடம் தரவுகளைக் கட்டற்ற உரிமத்தில் வெளியிடும்படி வேண்டுதல் (advocacy) - Tamil Wiki Community, FOSS Community, Kaniyam Foundation, Openaccessindia.org

திறந்த தமிழ் தரவுகளின் பட்டியல்



Open Tamil Texts Catalogue

File Edit View Insert Format Data Tools Add-ons Help Last edit was 18 minutes ago

Dataset

https://github...e/master/data

	Dataset	Use	Size	Format	Available Online	License
2	Ta	Text Analysis	127k records	Text	Yes	CC BY-SA 4.0
3	Tamil Wikipedia Articles with Parent Category	Text Classification	~90k records	CSV	Yes	CC BY-SA 4.0
4	Wikimedia Corpus	Text Analysis	~ 5 million words	Text		CC BY-SA 3.0
5	Project Madurai	Text Analysis	690 works	Formatted text	Yes	CC0: Public Domain
6	Kaniyam Foundaiton Resources	Text Analysis			Yes	CC BY-SA 4.0
7	EMILLE (Enabling Minority Language Engineering)		20 million words	?	No	?
8	LDC-IL - Text Corpus	Text Classification/Analysis	10 933 484 words	Word ?	No	Non Commercial, Not Sharable
9	Tamil Texts - 1000 Nationalized Books	Text Analysis	1000 works	Text with metadata	Not Yet	CC0: Public Domain
10	Thirukkural	Text Classification/Analysis	1330 records	Text	Yes	CC0: Public Domain ?
11	Mozilla Common Speech - Tamil	Speech Recognition & Synthesis	132 MB, 3 hrs	Audio & Text	Yes	CC0: Public Domain
12	LDC-IL - Speech Data	Speech Recognition & Synthesis	142 hrs	Audio		Non Commercial, Not Sharable
13	Noolaham Oral Histories		~ 1000 hrs	Audio	Yes	CC BY-SA 4.0
14	Word Embeddings	Language Modeling		pre-trained word vectors	Yes	MIT
15	Dinamalar Tamil News Dataset	Text Classification	1.9 million records		Yes	CC0: Public Domain
16	Tamil News Classification Dataset (Tamilmurasu)	Text Classification	127k articles		Yes	CC0: Public Domain
17	Tamil News Dataset	Text Classification	6500 records		Yes	CC BY-SA 4.0
18	Tamil Hand Written Characters	Character Recognition			Yes	CC0: Public Domain
19	Tamil Characters (Vowels)	Character Recognition			Yes	
20	AUKBC-TamilPOSCorpus2016v1	Parts of Speech Tagging	50876 sentences		No	Non Commercial, Not Sharable
21	Tamil WordNet	Strutured Dictionary	50497 words	SQL ?	Yes ?	GNU GPL
22	Event Annotated Data	Named Entity Recognition			No	?
23	NER Annotated Corpus	Named Entity Recognition			No	?
24	Verb Phrase Translation Data	Machine Translation				Non Commercial, Not Sharable
25	Tamil NLP	Text Classification			Yes	CC BY-SA 4.0
26	LDC-IL (BIS) POS tagset	Parts of Speech Tagging	1376857 words		No	Non Commercial, Not Sharable