Hands on advanced machine learning for information extraction from tweets tasks, data, and open source tools

Shubhanshu Mishra

Date: July 24, 2019

Time:: 9:00 am - 1:00 pm

Venue: UIUC EnterpriseWorks Room 130

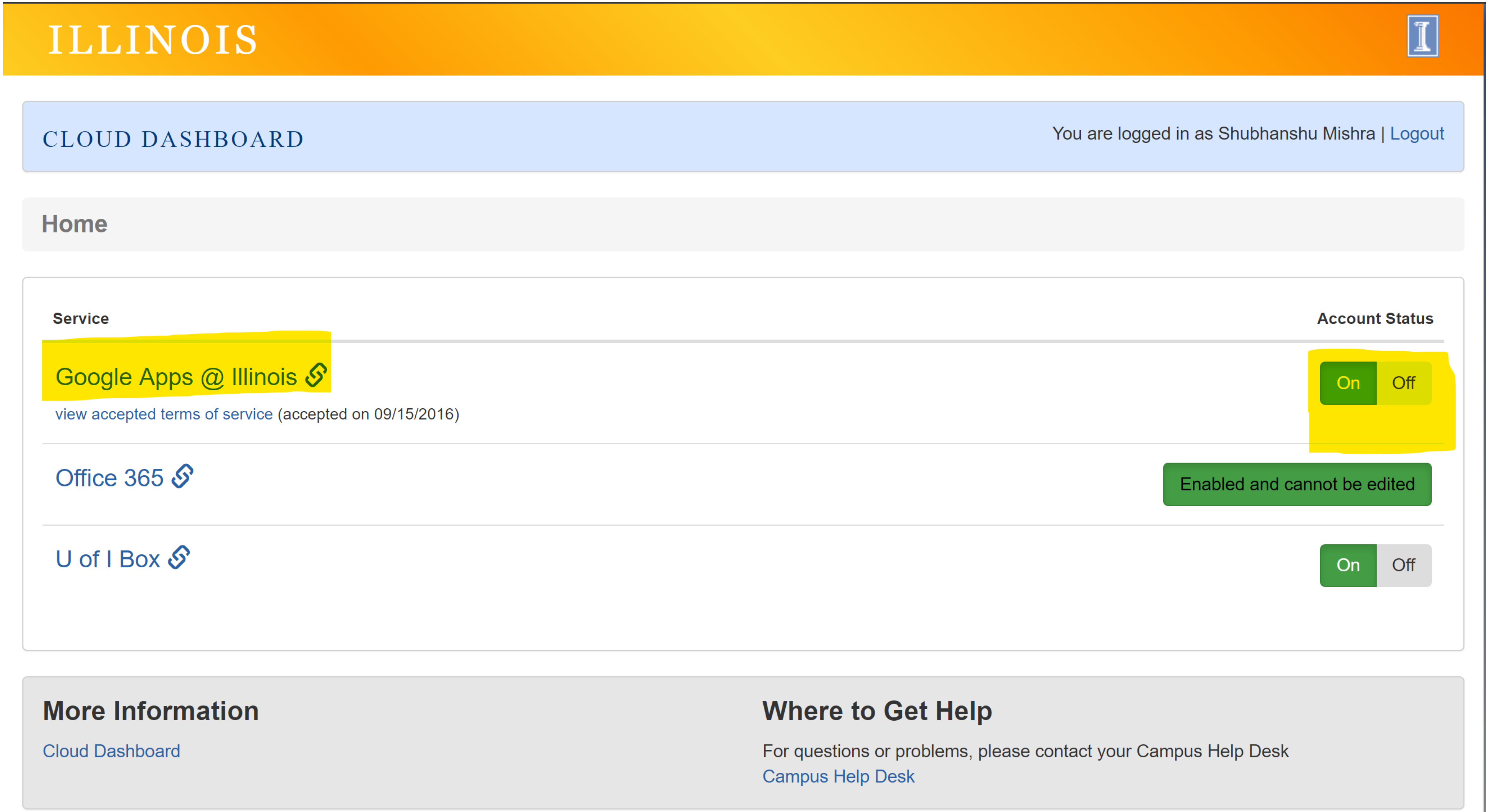
Details: https://socialmediaie.github.io/tutorials/

Overview

- Introduction (15 mins)
- Applications of information extraction (15 mins)
- Responsible and compliant data use of tweets (15 mins)
- Break (15 mins)
- Hands on session (1 hr. 30 mins)
- Conclusion (15 mins)

Setup

- If UIUC student, go to https://cloud-dashboard.illinois.edu/
- Enable google apps
- Install dependencies.
- We will use Google Colab for online hands on.
- Links to install instructions and google colaboratory notebooks at: https://socialmediaie.github.io/tutorials/UIUC2019/



ILLINOIS

CLOUD DASHBOARD

You are logged in as Shubhanshu Mishra | Logout

Home

Service	Account Status
Google Apps @ Illinois view accepted terms of service (accepted on 09/15/2016)	On Off
Office 365	Enabled and cannot be edited
U of I Box	On Off

More Information

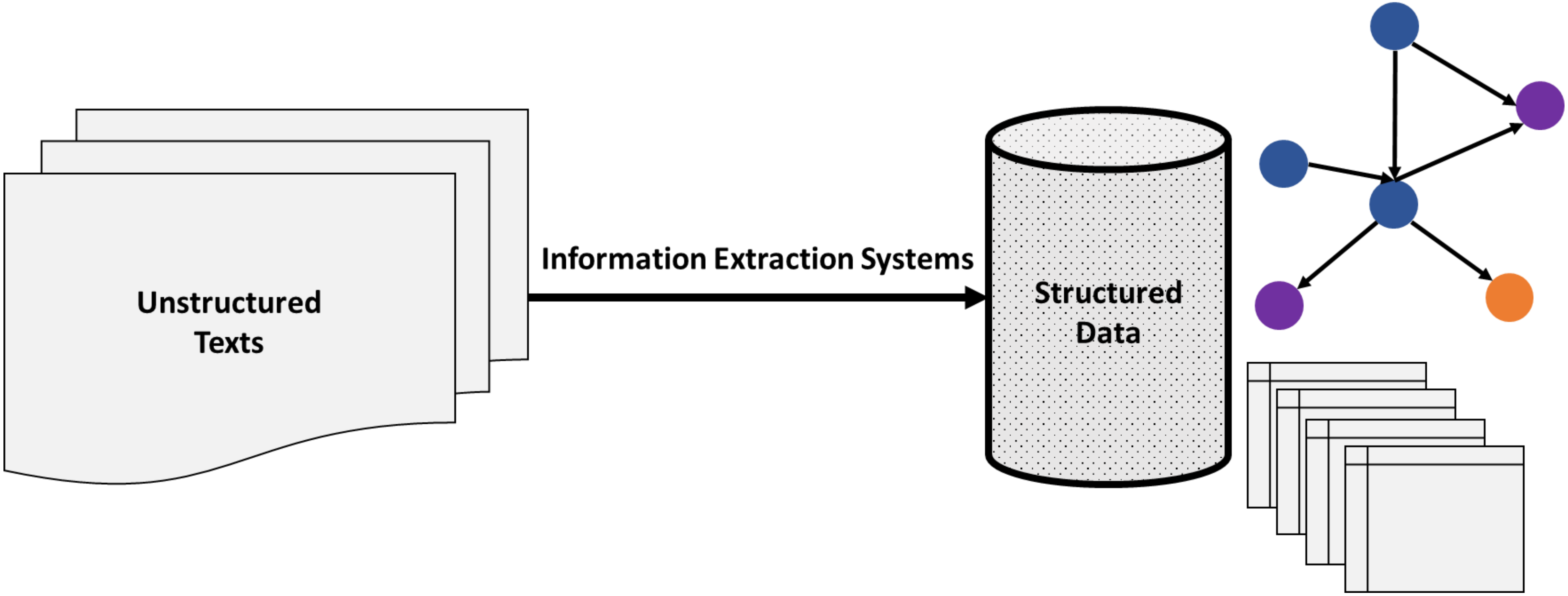
Cloud Dashboard

Where to Get Help

For questions or problems, please contact your Campus Help Desk

Campus Help Desk

Information extraction

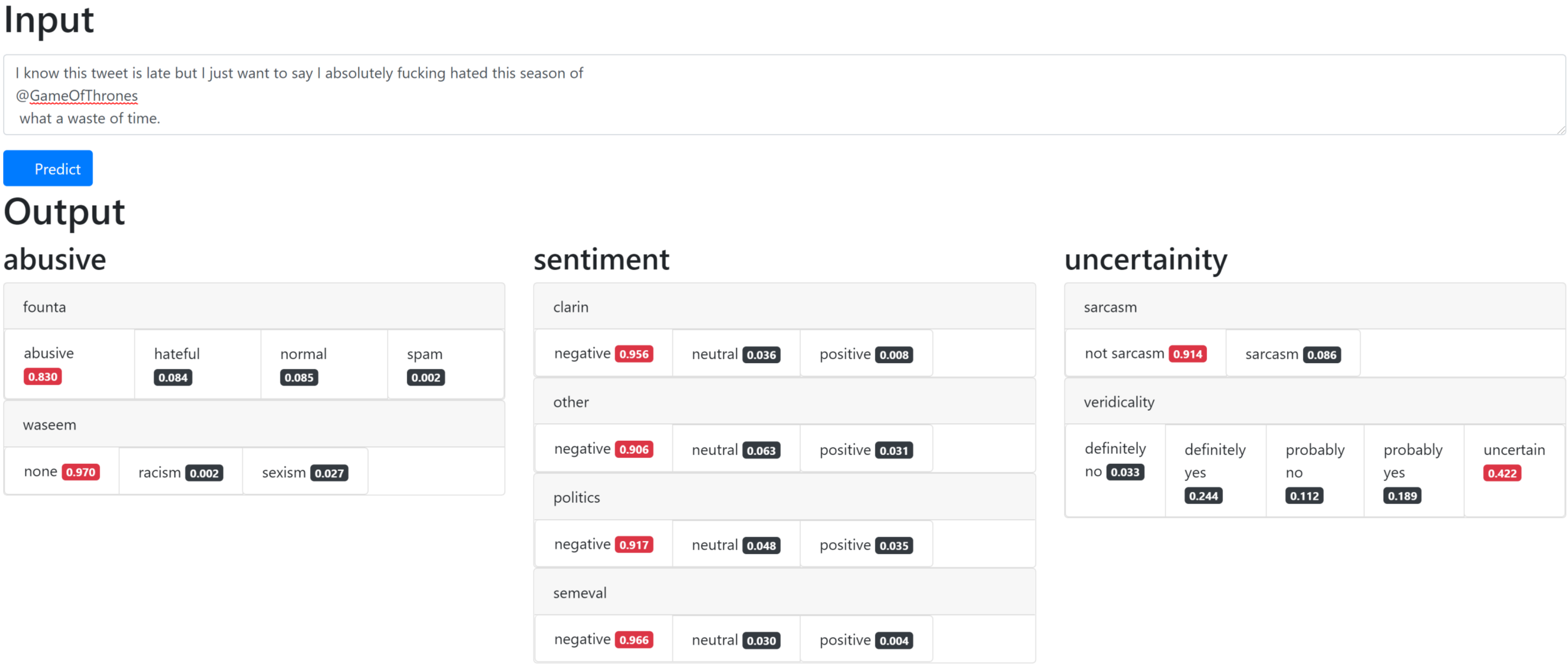


"Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources."
– (Sarawagi, 2008)

Information extraction tasks for text

- **Text classification** : sentiment prediction, sarcasm detection, and abusive content detection.
- **Sequence tagging** : named entity detection and classification, part of speech tagging, chunking, and super-sense tagging.

Text classification



Input

I know this tweet is late but I just want to say I absolutely fucking hated this season of
@GameOfThrones
what a waste of time.

Output

abusive

founta

abusive	hateful	normal	spam
0.830	0.084	0.085	0.002

waseem

none	racism	sexism
0.970	0.002	0.027

sentiment

clarin

negative	neutral	positive
0.956	0.036	0.008

other

negative	neutral	positive
0.906	0.063	0.031

politics

negative	neutral	positive
0.917	0.048	0.035

semeval

negative	neutral	positive
0.966	0.030	0.004

uncertainity

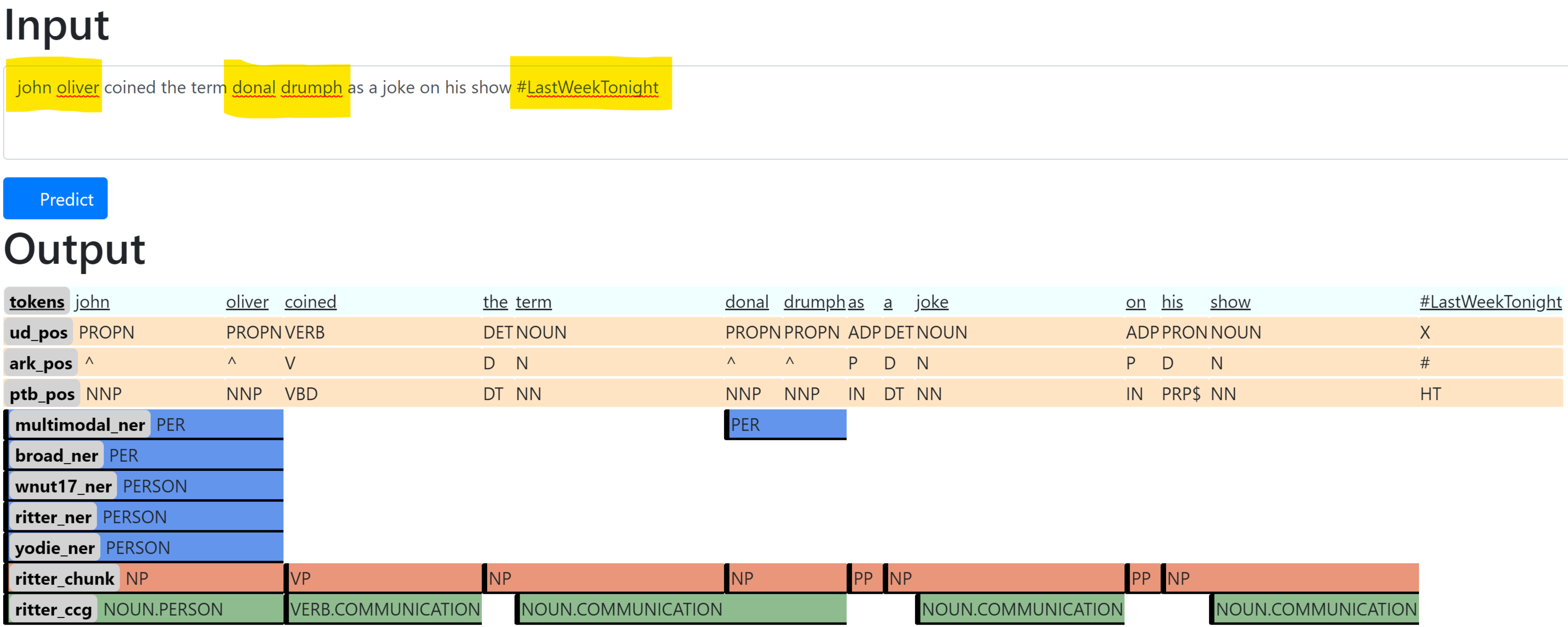
sarcasm

not sarcasm	sarcasm
0.914	0.086

veridicality

definitely no	definitely yes	probably no	probably yes	uncertain
0.033	0.244	0.112	0.189	0.422

Sequence tagging



Input

john oliver coined the term donal drumph as a joke on his show #LastWeekTonight

Output

tokens	john	oliver	coined	the	term	donal	drumph	as	a	joke	on	his	show	#LastWeekTonight
ud_pos	PROPN	PROPN	VERB	DET	NOUN	PROPN	PROPN	ADP	DET	NOUN	ADP	PRON	NOUN	X
ark_pos	^	^	V	D	N	^	^	P	D	N	P	D	N	#
ptb_pos	NNP	NNP	VBD	DT	NN	NNP	NNP	IN	DT	NN	IN	PRP$	NN	HT
multimodal_ner	PER					PER								
broad_ner	PER													
wnut17_ner	PERSON													
ritter_ner	PERSON													
yodie_ner	PERSON													
ritter_chunk	NP		VP	NP		NP		PP	NP		PP	NP		
ritter_ccg	NOUN.PERSON		VERB.COMMUNICATION		NOUN.COMMUNICATION					NOUN.COMMUNICATION			NOUN.COMMUNICATION	

Applications of information extraction

- Index documents by entities
- Entity mention clustering
- Visualizing temporal trends in data: https://shubhanshu.com/social-comm-temporal-graph/

DocID	Entity	Entity type	WikiURL
1	Barack Obama	Person	URL1
2	Facebook	Organization	URL2
3	Katy Perry	Music Artist	URL3

Washington is a great place.
I just visited **Washington**.

Washington was a great president.
Washington made some good changes to constitution.

Responsible and compliant data use of tweets

- Always collect data via Twitter API
- Tweets are often shared via tweetID and the annotation.
- Never publicly share the full text or JSON of the tweet data.
- Some exceptions for academic usage. See: https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.html
- When possible try to respect user privacy.
- When making inference from collected data be responsible. Think what if your data was collect, what all would you be OK with being inferred.

Publicly available Twitter data

- Many researchers make annotated Twitter data publicly available **for academic research**.
- Good place for benchmarking or evaluating your models.
- Many datasets available for text classification.
- Few for information extraction via sequence tagging (but still enough)
- Varied annotation practices and data scope:
- See here: https://socialmediaie.github.io/datasets.html

Hands on session

Links to install instructions and google colaboratory notebooks at: https://socialmediaie.github.io/tutorials/UIUC2019/

Tagging data

Super sense tagging

data	split	labels	sequences	vocab	tokens
Ritter	train	40	551	3174	10652
	dev	37	118	1014	2242
	test	40	118	1011	2291
Johannsen2014	test	37	200	1249	3064

Chunking

data	split	boundaries	labels	labels	sequences	vocab	tokens
Ritter	train	[I, B, O]	[ADJP, PP, INTJ, ADVP, PRT, NP, SBAR, VP, CONJP]	9	551	3158	10584
	dev	[I, B, O]	[ADJP, PP, INTJ, ADVP, PRT, NP, SBAR, VP]	8	118	994	2317
	test	[I, B, O]	[ADJP, PP, INTJ, ADVP, PRT, NP, SBAR, VP]	8	119	988	2310

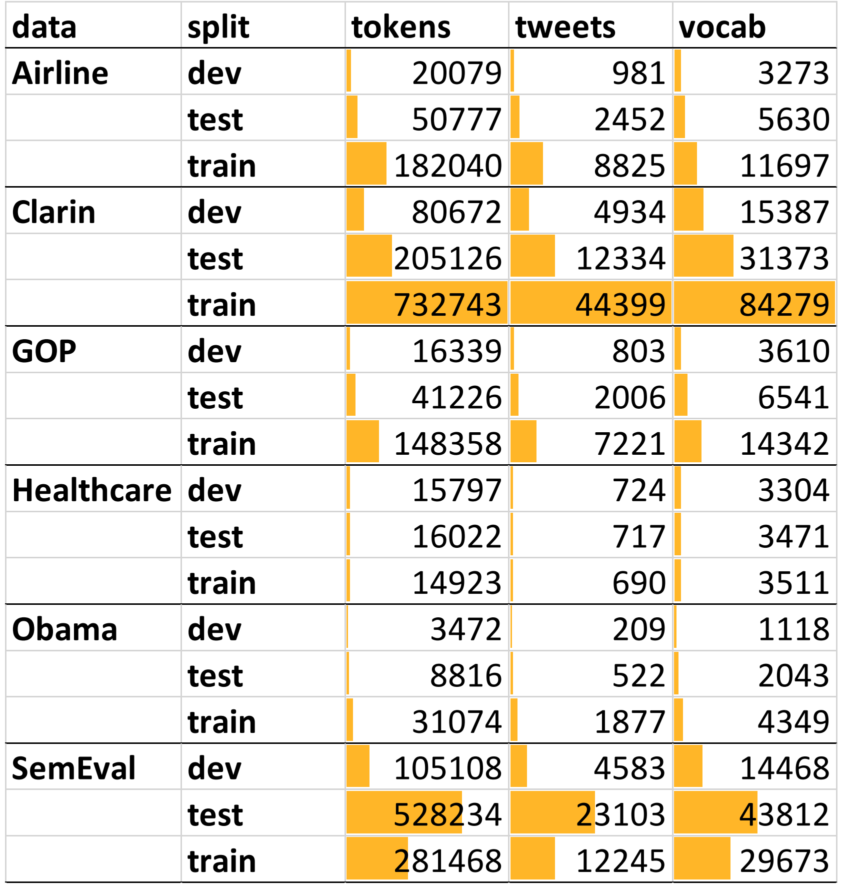
Part of speech tagging

data	split	labels	sequences	vocab	tokens
Owoputi	train	25	1547	6572	22326
	dev	23	327	2036	4823
	test	23	500	2754	7152
TwitIE	dev	43	269	1229	2998
	test	45	632	3539	12196
Ritter	train	45	632	3539	12196
	dev	38	71	695	1362
	test	42	84	735	1627
Tweetbankv2	dev	17	710	3271	11759
	train	17	1639	5632	24753
	test	17	1201	4699	19095
DiMSUM2016	train	17	4799	9113	73826
	test	17	1000	4010	16500
Foster	test	12	250	1068	2841
lowlands	test	12	1318	4805	19794

Named entity recognition

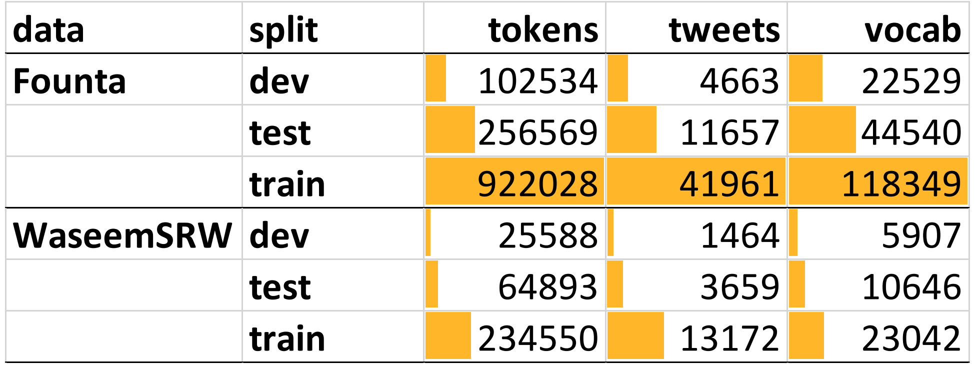
data	split	labels	sequences	vocab	tokens
YODIE	train	13	396	2554	7905
	test	13	397	2578	8032
Ritter	train	10	1900	7695	36936
	dev	10	240	1731	4612
	test	10	254	1776	4921
WNUT2016	train	10	2394	9068	46469
	test	10	3850	16012	61908
	dev	10	1000	5563	16261
WNUT2017	train	6	3394	12840	62730
	dev	6	1009	3538	15733
	test	6	1287	5759	23394
NEEL2016	train	7	2588	9731	51669
	dev	7	88	762	1647
	test	7	2663	9894	47488
Finin	train	3	10000	19663	172188
	test	3	5369	13027	97525
Hege	test	3	1545	4552	20664
BROAD	train	3	5605	19523	90060
	dev	3	933	5312	15169
	test	3	2802	11772	45159
MultiModal	train	4	4000	20221	64439
	dev	4	1000	6832	16178
	test	4	3257	17381	52822
MSM2013	train	4	2815	8514	51521
	test	4	1450	5701	29089

Classification data



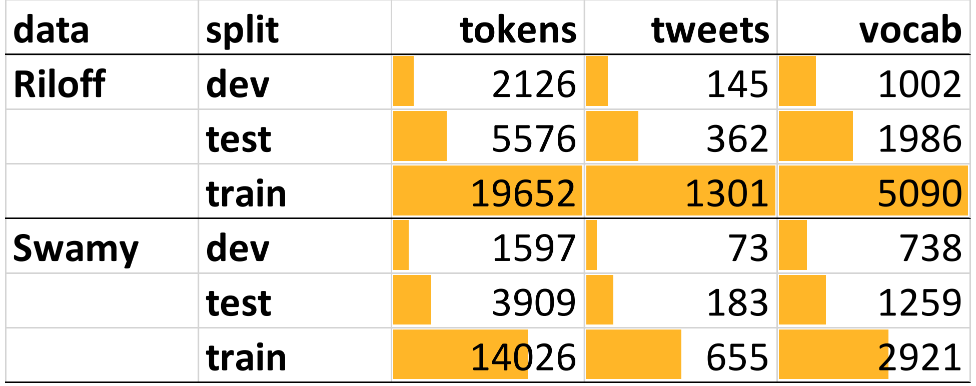
data	split	tokens	tweets	vocab
Airline	dev	20079	981	3273
	test	50777	2452	5630
	train	182040	8825	11697
Clarin	dev	80672	4934	15387
	test	205126	12334	31373
	train	732743	44399	84279
GOP	dev	16339	803	3610
	test	41226	2006	6541
	train	148358	7221	14342
Healthcare	dev	15797	724	3304
	test	16022	717	3471
	train	14923	690	3511
Obama	dev	3472	209	1118
	test	8816	522	2043
	train	31074	1877	4349
SemEval	dev	105108	4583	14468
	test	528234	23103	43812
	train	281468	12245	29673

Sentiment classification



data	split	tokens	tweets	vocab
Founta	dev	102534	4663	22529
	test	256569	11657	44540
	train	922028	41961	118349
WaseemSRW	dev	25588	1464	5907
	test	64893	3659	10646
	train	234550	13172	23042

Abusive content identification



data	split	tokens	tweets	vocab
Riloff	dev	2126	145	1002
	test	5576	362	1986
	train	19652	1301	5090
Swamy	dev	1597	73	738
	test	3909	183	1259
	train	14026	655	2921

Uncertainty indicator classification

Twitter NER

Rank	1	2	3	4	5	6	7	8	9	10	TD	TDT$_E$
10-types	52.4	46.2	44.8	40.1	39.0	37.2	**37.0**	36.2	29.8	19.3	**46.4**	**47.3**
No-types	65.9	63.2	60.2	59.1	55.2	51.4	**47.8**	46.7	44.3	40.7	**57.3**	**59.0**
company	57.2	46.9	43.8	31.3	38.9	34.5	**25.8**	42.6	24.3	10.2	42.1	46.2
facility	42.4	31.6	36.1	36.5	20.3	30.4	**37.0**	40.5	26.3	26.1	37.5	34.8
geo-loc	72.6	68.4	63.3	61.1	61.1	57.0	**64.7**	60.9	47.4	37.0	70.1	71.0
movie	10.9	5.1	4.6	15.8	2.9	0.0	**4.0**	5.0	0.0	5.4	0.0	0.0
musicartist	9.5	8.5	7.0	17.4	5.7	37.2	**1.8**	0.0	2.8	0.0	7.6	5.8
other	31.7	27.1	29.2	26.3	21.1	22.5	**16.2**	13.0	22.6	8.4	31.7	32.4
person	59.0	51.8	52.8	48.8	52.0	42.6	**40.5**	52.3	34.1	20.6	51.3	52.2
product	20.1	11.5	18.3	3.8	10.0	7.3	**5.7**	15.4	6.3	0.8	10.0	9.3
sportsteam	52.4	34.2	38.5	18.5	34.6	15.9	**9.1**	19.7	11.0	0.0	31.3	32.0
tvshow	5.9	0.0	4.7	5.4	7.3	9.8	**4.8**	0.0	5.1	0.0	5.7	5.7
Rank	1	2	3	4	5	6	**7**	8	9	10	~2	~2

Multi-task-multi-dataset learning - tagging

Part of speech tagging

Data	Our best	SOTA	Diff %
DiMSUM2016	86.77	82.49	5%
Owoputi	91.76	88.89	3%
TwitIE	91.62	89.37	3%
Ritter	92.01	90	2%
Tweetbankv2	92.44	93.3	-1%
Foster	69.34	90.4	-23%
lowlands	68.1	89.37	-24%

Super sense tagging

Data	Our best	SOTA	Diff %
Ritter	59.16	57.14	3.5%
Johannsen2014	42.38	42.42	-0.1%

Named entity recognition

Data	Our best	SOTA	Diff %
BROAD	77.40	None	NA
YODIE	65.39	None	NA
Finin	56.42	32.43	74.0%
MSM2013	80.46	58.72	37.0%
Ritter	86.04	82.6	4.2%
MultiModal	73.39	70.69	3.8%
Hege	89.45	86.9	2.9%
WNUT2016	53.16	52.41	1.4%
WNUT2017	49.86	49.49	0.8%

Chunking

Data	Our best	SOTA	Diff %
Ritter	88.92	None	NA

Thanks

- TwitterNER: https://github.com/napsternxg/TwitterNER
- Social Communication Temporal Graph: https://shubhanshu.com/social-comm-temporal-graph/
- SocialMediaIE for multi-task learning: https://socialmediaie.github.io/ (Will be open sourced by August 2019)
- For queries please send a tweet at: @TheShubhanshu