“Strax” — Advancements in Topic Modeling

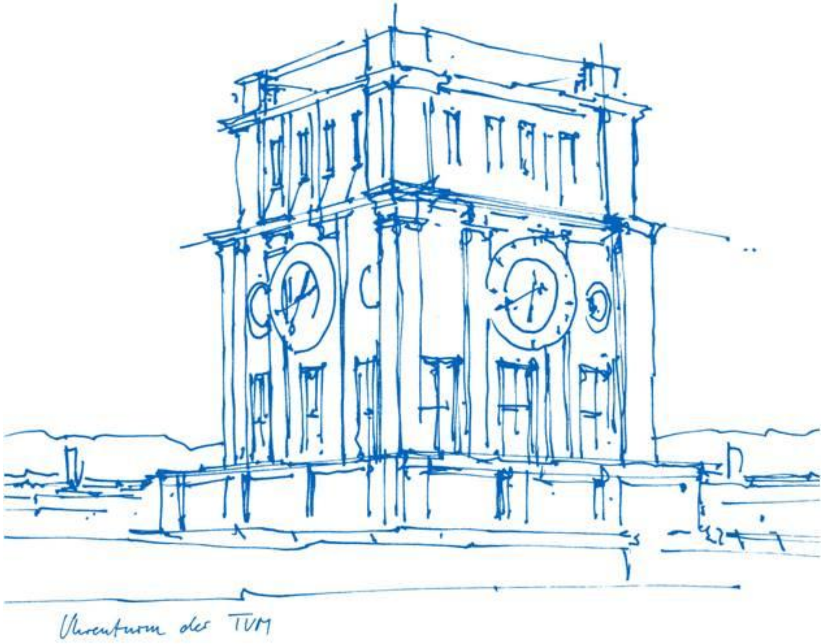
Technische Universität München

Fakultät für Informatik

NLP Lab Course, SS22

14.07.2022

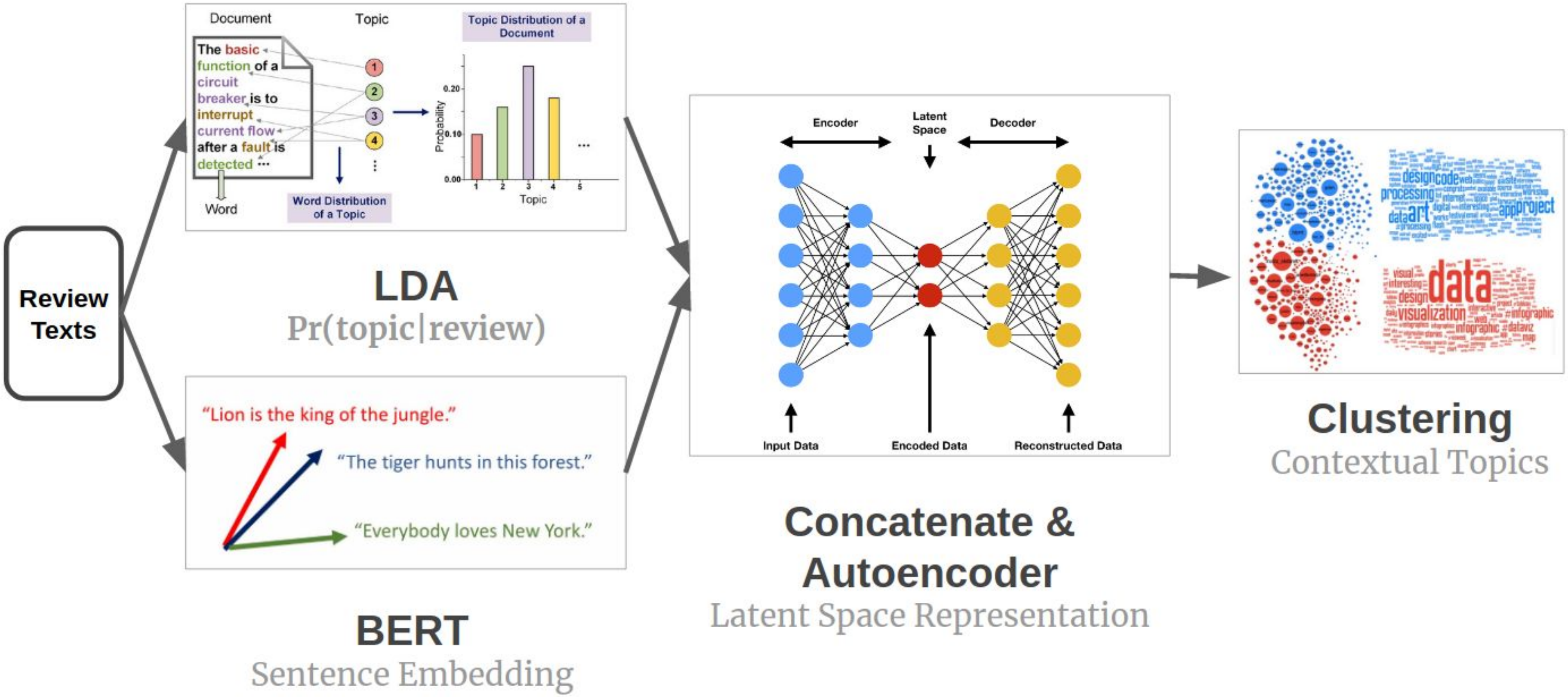
Berk Sudan, Ferdinand Kapl, Yuyin Lang



Overview

- LDA-BERT
- Visualization
- Yahoo Dataset & Visualizer
- Current Status

LDA-BERT


Document
Topic
Topic Distribution of a Document
The basic function of a circuit breaker is to interrupt current flow after a fault is detected ...
1
2
3
4
Word
Word Distribution of a Topic
Probability
0.20
0.10
0.00
1
2
3
4
5
Topic
...
"Lion is the king of the jungle."
"The tiger hunts in this forest."
"Everybody loves New York."
Encoder
Latent Space
Decoder
Input Data
Encoded Data
Reconstructed Data
data
visualization
design
infographic
processing
art
project
code

Review Texts

LDA

Pr(topic|review)

BERT

Sentence Embedding

Concatenate & Autoencoder

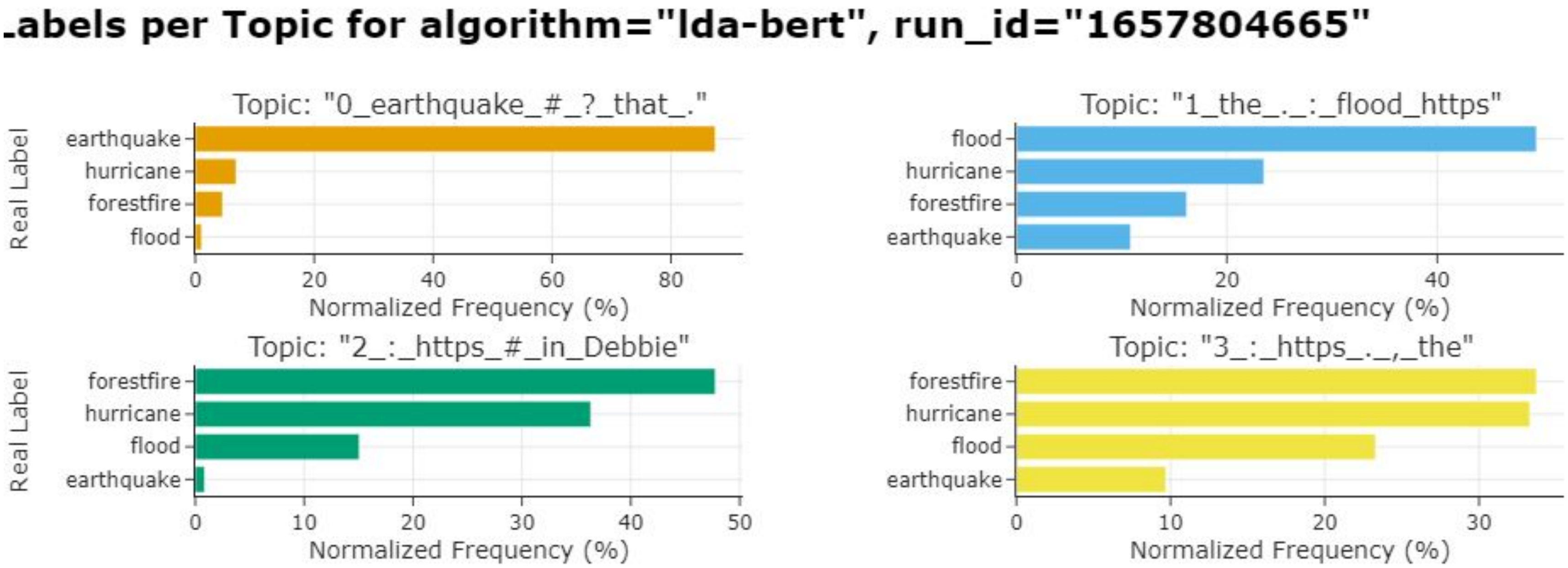
Latent Space Representation

Clustering

Contextual Topics

source: https://blog.insightdatascience.com/contextual-topic-identification-4291d256a032

LDA-BERT: Current Results



.abels per Topic for algorithm="lda-bert", run_id="1657804665"

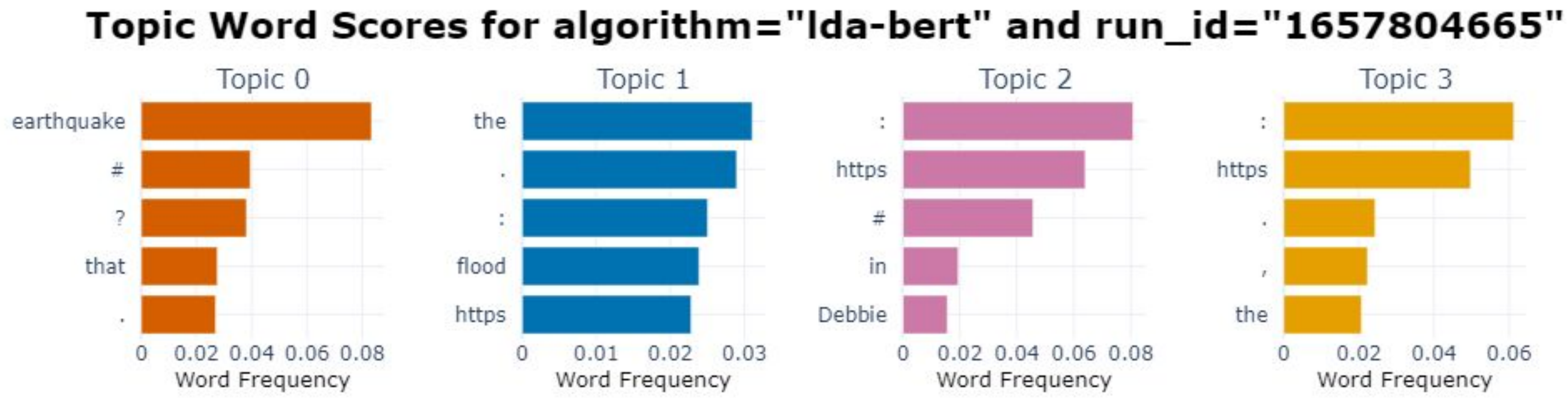
Topic: "0_earthquake_#_?_that_."

Topic: "1_the_._:_flood_https"

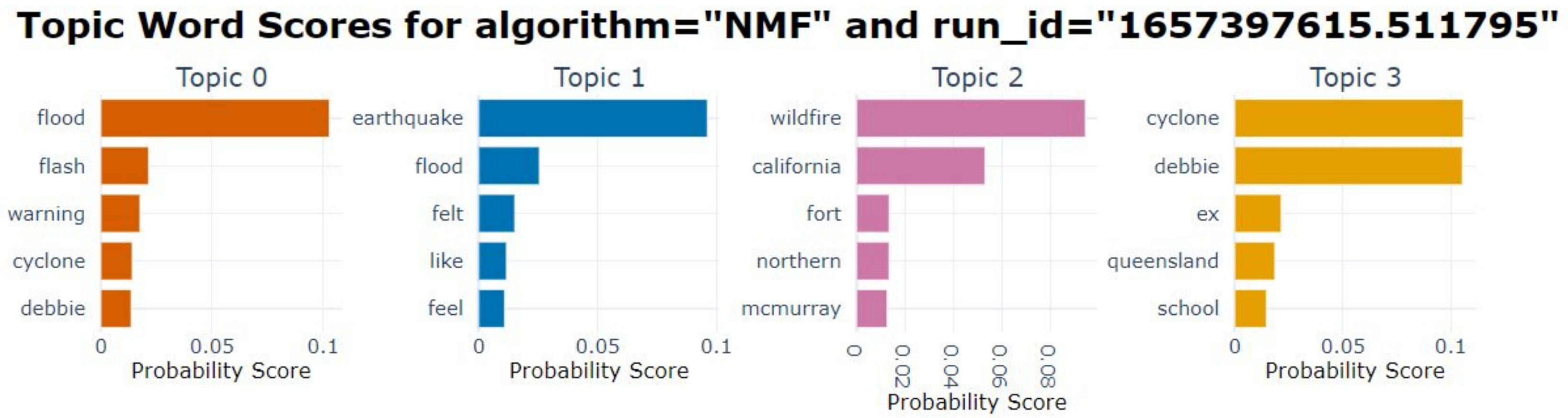
Topic: "2_:_https_#_in_Debbie"

Topic: "3_:_https_._,_the"

LDA-BERT: Current Results

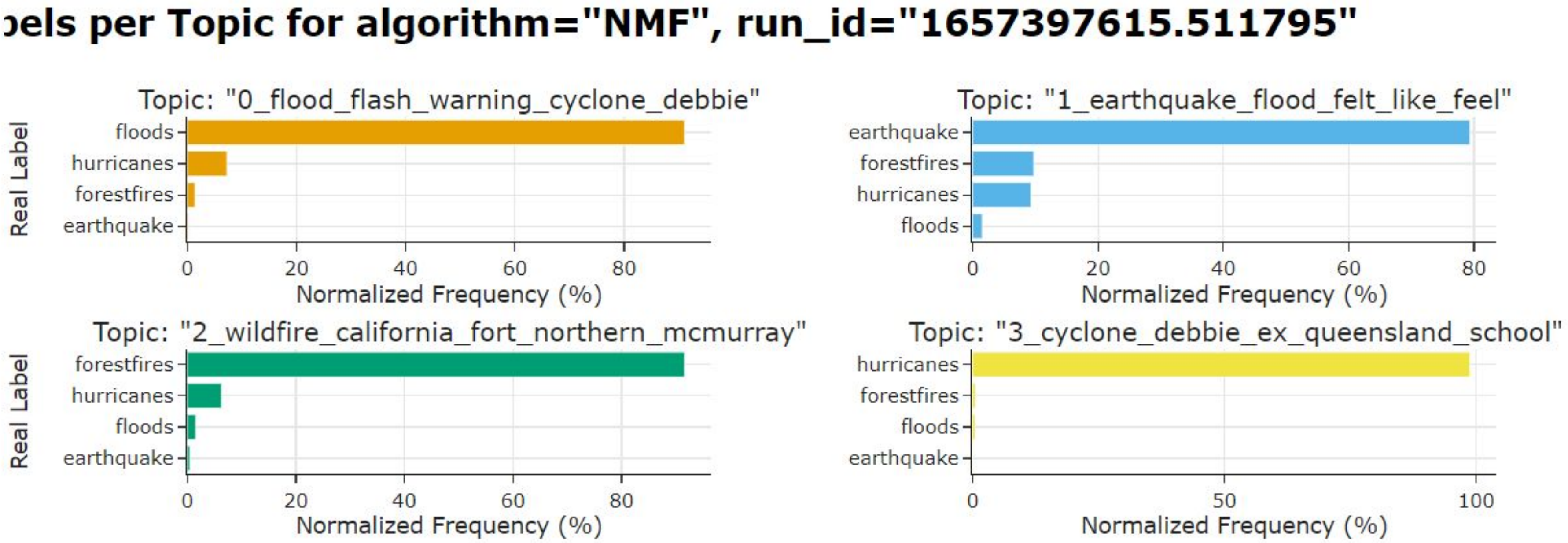


Visualization of LDA and NMF



Topic Word Scores for algorithm="NMF" and run_id="1657397615.511795"

Visualization of LDA and NMF



bels per Topic for algorithm="NMF", run_id="1657397615.511795"

Topic: "0_flood_flash_warning_cyclone_debbie"

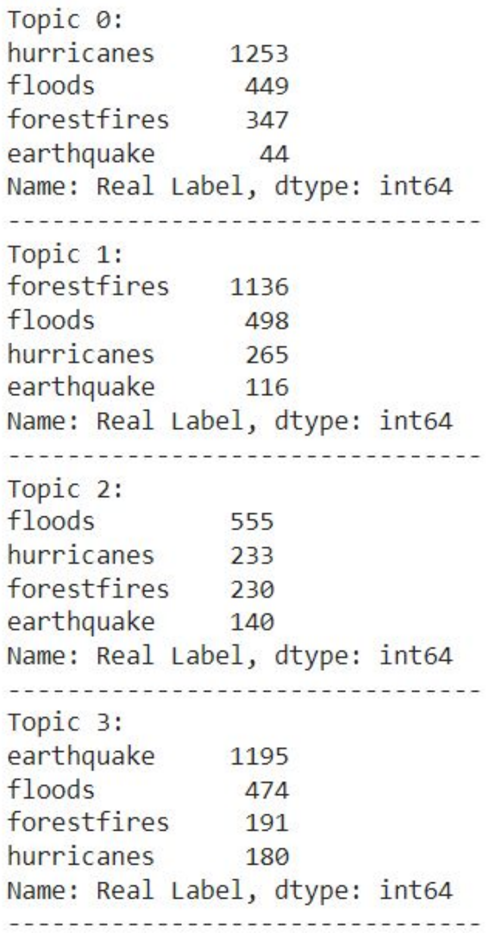
Topic: "1_earthquake_flood_felt_like_feel"

Topic: "2_wildfire_california_fort_northern_mcmurray"

Topic: "3_cyclone_debbie_ex_queensland_school"

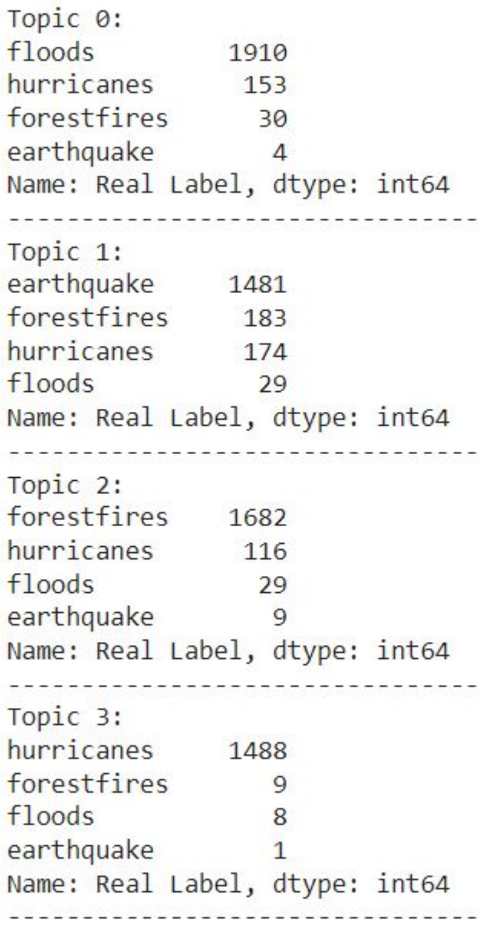
Visualization of LDA and NMF (for comparison)

LDA:



```
Topic 0:
hurricanes     1253
floods          449
forestfires     347
earthquake       44
Name: Real Label, dtype: int64
--------------------------------
Topic 1:
forestfires    1136
floods          498
hurricanes      265
earthquake      116
Name: Real Label, dtype: int64
--------------------------------
Topic 2:
floods         555
hurricanes     233
forestfires    230
earthquake     140
Name: Real Label, dtype: int64
--------------------------------
Topic 3:
earthquake     1195
floods          474
forestfires     191
hurricanes      180
Name: Real Label, dtype: int64
--------------------------------
```

NMF:



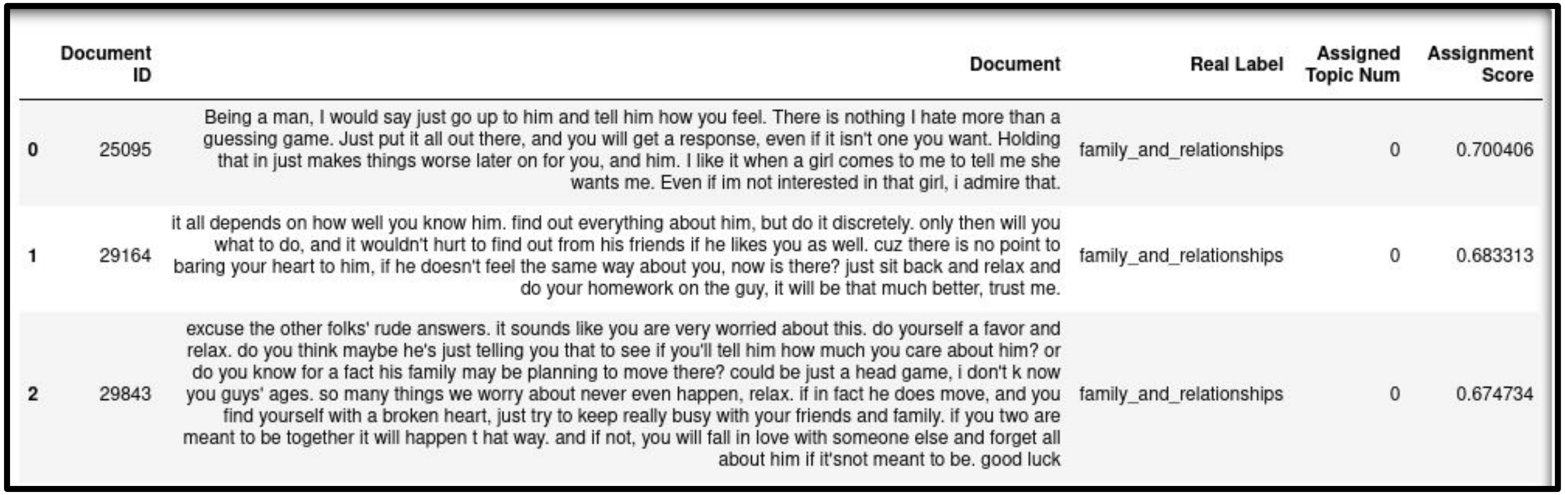
```
Topic 0:
floods         1910
hurricanes      153
forestfires      30
earthquake        4
Name: Real Label, dtype: int64
--------------------------------
Topic 1:
earthquake     1481
forestfires     183
hurricanes      174
floods           29
Name: Real Label, dtype: int64
--------------------------------
Topic 2:
forestfires    1682
hurricanes      116
floods           29
earthquake        9
Name: Real Label, dtype: int64
--------------------------------
Topic 3:
hurricanes     1488
forestfires       9
floods            8
earthquake        1
Name: Real Label, dtype: int64
--------------------------------
```

Glance at optional algorithms

Algorithms	Special for	Code availability
LSA	-	Basic code
PLSA	-	Basic code in Enstop
BTM	Short text	No python
DMM (GSDMM)	Short text	Basic code
WNTM	Short text	No python
CTM	-	OCTIS, need to read source code for more information

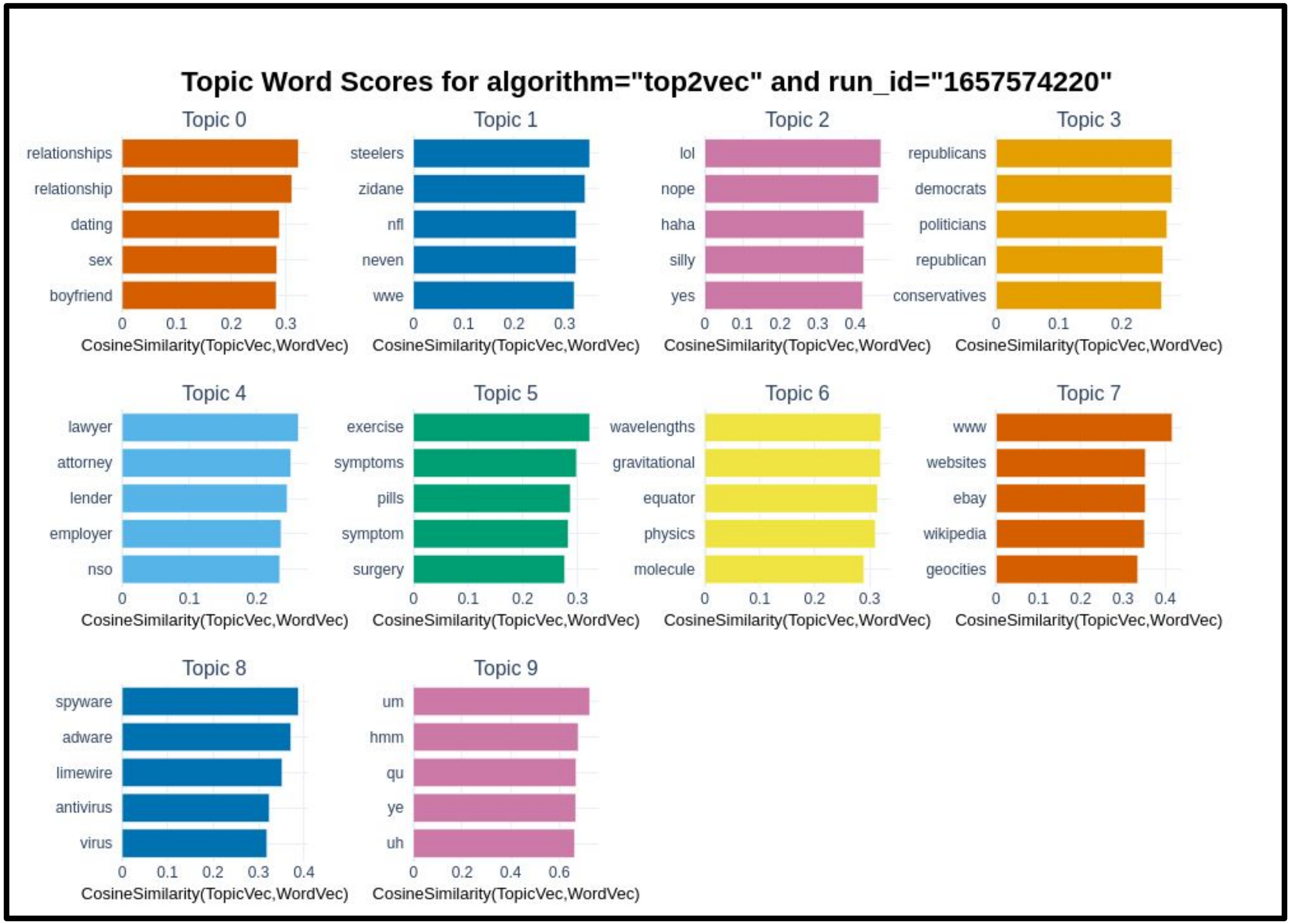
Yahoo Dataset - Characteristics

- A Long Text Dataset
- # of Classes: 10
- # of Instances: 60K (selected randomly 300K+)
- Instance / (Real Label): 6K
- Excerpt from the data:

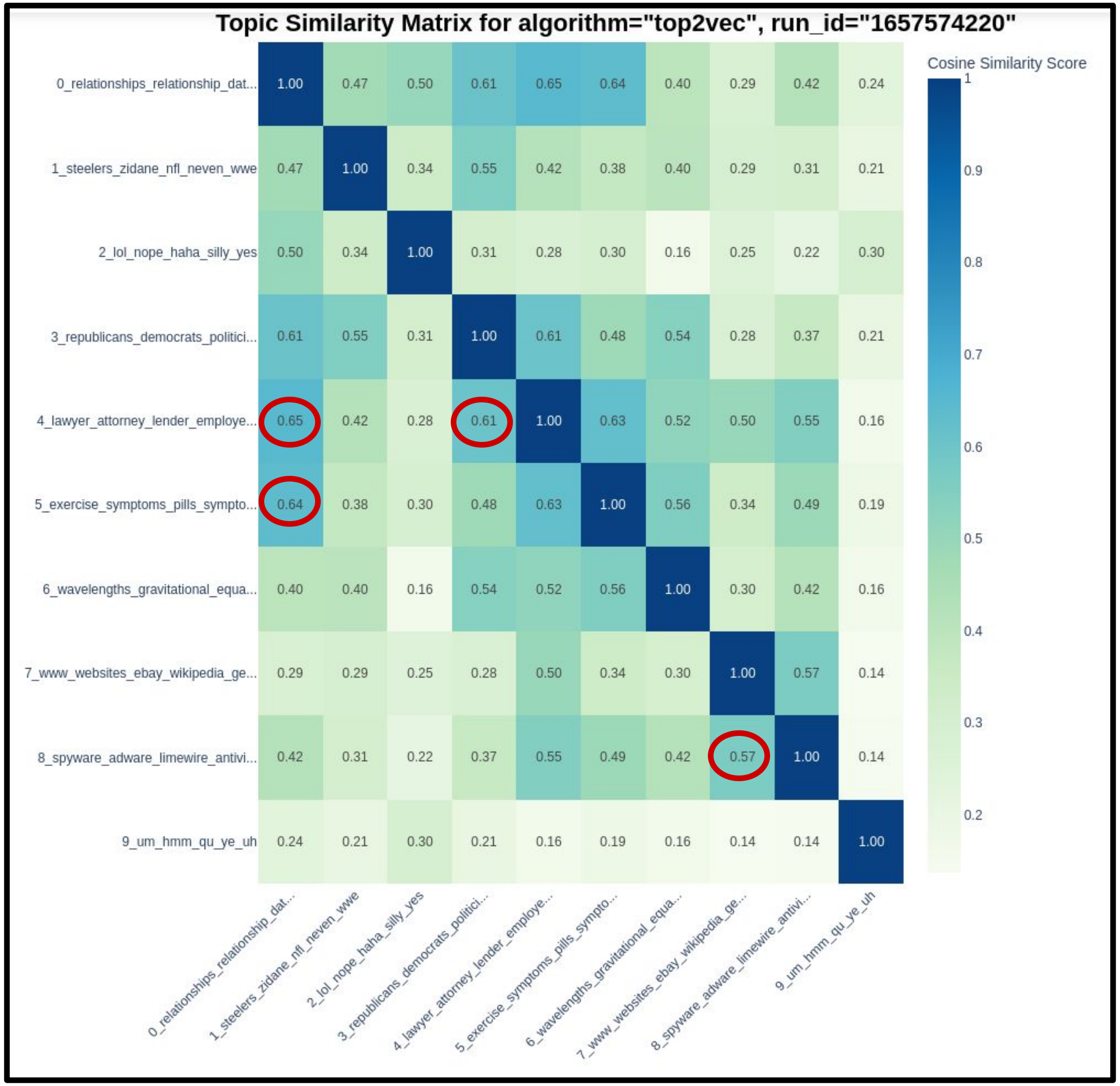


	Document ID	Document	Real Label	Assigned Topic Num	Assignment Score
0	25095	Being a man, I would say just go up to him and tell him how you feel. There is nothing I hate more than a guessing game. Just put it all out there, and you will get a response, even if it isn't one you want. Holding that in just makes things worse later on for you, and him. I like it when a girl comes to me to tell me she wants me. Even if im not interested in that girl, i admire that.	family_and_relationships	0	0.700406
1	29164	it all depends on how well you know him. find out everything about him, but do it discretely. only then will you what to do, and it wouldn't hurt to find out from his friends if he likes you as well. cuz there is no point to baring your heart to him, if he doesn't feel the same way about you, now is there? just sit back and relax and do your homework on the guy, it will be that much better, trust me.	family_and_relationships	0	0.683313
2	29843	excuse the other folks' rude answers. it sounds like you are very worried about this. do yourself a favor and relax. do you think maybe he's just telling you that to see if you'll tell him how much you care about him? or do you know for a fact his family may be planning to move there? could be just a head game, i don't k now you guys' ages. so many things we worry about never even happen, relax. if in fact he does move, and you find yourself with a broken heart, just try to keep really busy with your friends and family. if you two are meant to be together it will happen t hat way. and if not, you will fall in love with someone else and forget all about him if it'snot meant to be. good luck	family_and_relationships	0	0.674734

Yahoo Dataset - Topic Word Scores



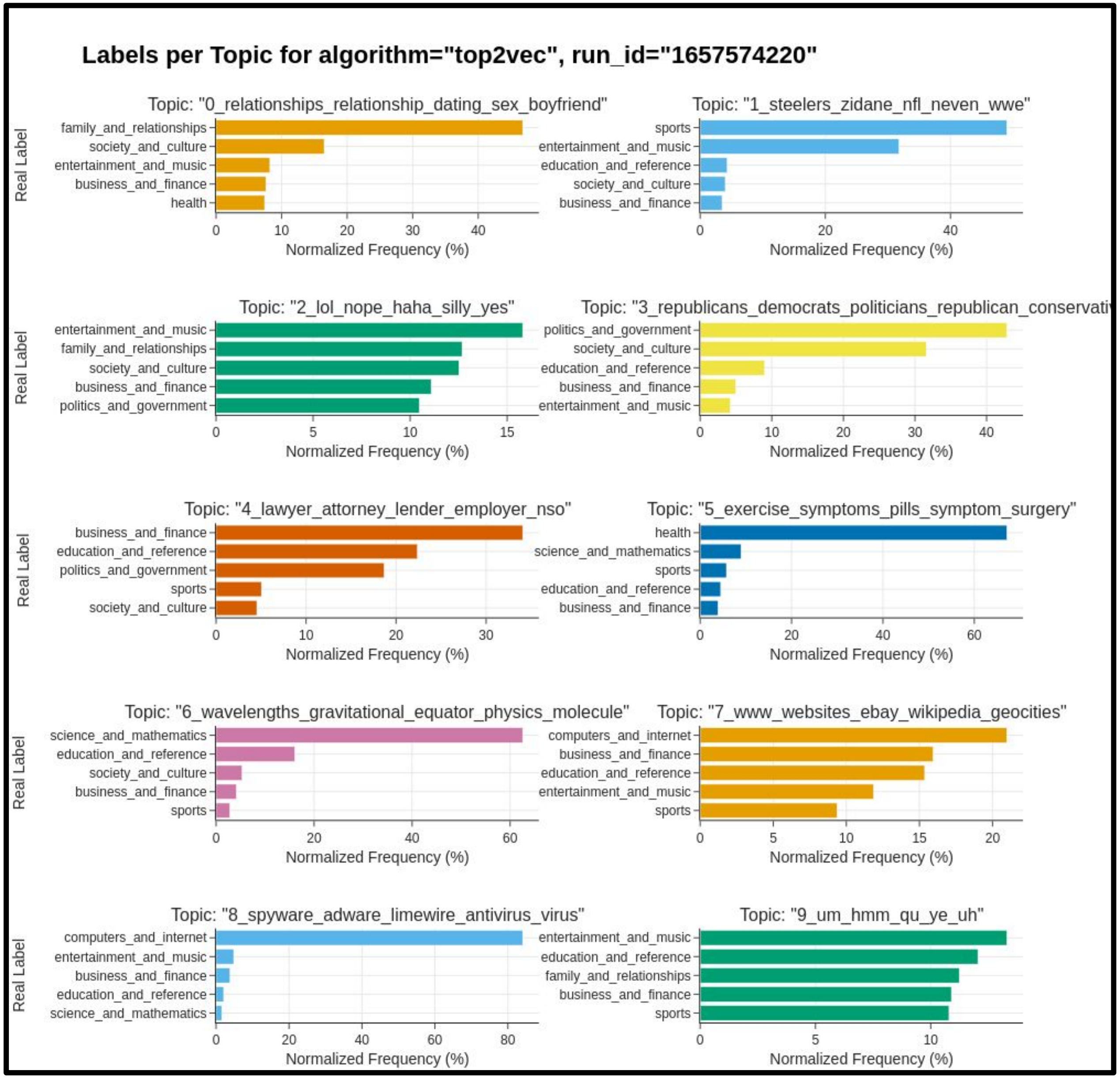
Yahoo Dataset - Topic Similarity Matrix (New)



Topic Similarity Matrix for algorithm="top2vec", run_id="1657574220"

	0_relationships_relationship_dat...	1_steelers_zidane_nfl_neven_wwe	2_lol_nope_haha_silly_yes	3_republicans_democrats_politici...	4_lawyer_attorney_lender_employe...	5_exercise_symptoms_pills_sympto...	6_wavelengths_gravitational_equa...	7_www_websites_ebay_wikipedia_ge...	8_spyware_adware_limewire_antivi...	9_um_hmm_qu_ye_uh
0_relationships_relationship_dat...	1.00	0.47	0.50	0.61	0.65	0.64	0.40	0.29	0.42	0.24
1_steelers_zidane_nfl_neven_wwe	0.47	1.00	0.34	0.55	0.42	0.38	0.40	0.29	0.31	0.21
2_lol_nope_haha_silly_yes	0.50	0.34	1.00	0.31	0.28	0.30	0.16	0.25	0.22	0.30
3_republicans_democrats_politici...	0.61	0.55	0.31	1.00	0.61	0.48	0.54	0.28	0.37	0.21
4_lawyer_attorney_lender_employe...	0.65	0.42	0.28	0.61	1.00	0.63	0.52	0.50	0.55	0.16
5_exercise_symptoms_pills_sympto...	0.64	0.38	0.30	0.48	0.63	1.00	0.56	0.34	0.49	0.19
6_wavelengths_gravitational_equa...	0.40	0.40	0.16	0.54	0.52	0.56	1.00	0.30	0.42	0.16
7_www_websites_ebay_wikipedia_ge...	0.29	0.29	0.25	0.28	0.50	0.34	0.30	1.00	0.57	0.14
8_spyware_adware_limewire_antivi...	0.42	0.31	0.22	0.37	0.55	0.49	0.42	0.57	1.00	0.14
9_um_hmm_qu_ye_uh	0.24	0.21	0.30	0.21	0.16	0.19	0.16	0.14	0.14	1.00

Yahoo Dataset - Labels per Topic (New)



Current Status & Next

DONE:

- ALGs: LDA, NMF, Top2Vec, BERTopic, LDA-BERT
- Visualizations: Topic words barcharts, Real label - Topic distribution, Umap embedding plots, Similarity Heatmap

NEXT:

- Preprocessing
- Evaluation Module:
  - Coherence Metrics: C_V Coherence & NPMI Coherence
  - Diversity Metrics: Unique Words Topic Diversity & Kullback-Liebler Divergence
  - Supervised Metrics: Rand Index

Questions Time!!