Unsupervised Text Classification & Clustering: What are folks doing these days?

Rachael Tatman, Kaggle





						The submit at Don't Get Kicked!, pls help 0 Thursday, June
	and the second se			199	SETTING ICT.	- 1 Commont: [***] [error] late submit at Don't Get Kicked!, pls help - RunningZ commented on the provide the provide the provide the provide the provide the provide the provided the prov
		3		* >	Kaggle 2	[Product Feed] Commenter a subsection of the
	Inbox		-		Veggle	[Getting Star] Comment: Team up with particle - Source-r
*	Starred			立 🕨	Kaggie	Certing Star] Comment: Useful links to Read - Sowbarani Karthikeyan commence
O	Snoozed			☆ >	Kaggle 2	Getting of the second state of the second stat
>	Sent			N D	Kaggle 2	[Questions &] Comment: Close tame
	Drafts	6		-		[Questions &] Comment: NLP - Kuldeep Singn Arya commented
8	Indef			\$ Σ	Kaggle	It earn Comment: What Partial Dependence Plots Would You Be Curious To See Additional Section Section 2010 Thursday, June 6, 10:38 AM UTC I'm not even close to the top but I define
	auto_reply	1		ά Σ	Kaggle 2	[Learn] Commented by Kanada vs Real World ! - Jiwon Park commented on Kaggle vs Real World ! O Thursday, Care
	dataset		-		Kaggle 3	[Questions &] Comment: Kaggie volter and a segmented on why do EDA? 0 Thursday, June 6, 1:59 PM UTC Exploratory Data Analyse
	Kaggle Forums	116	_	м		[Questions &] Comment: why do EDA? - Ravi Maurya commence en
	linkodin snam] ☆]	> Kaggle 3	
	Inkeuit sputt				and the second	



1 44 40



Kaggle 3

Inbox

Starred

Sent

Drafts

Indef

auto_reply dataset

Kaggle Forums

linkedin spam

Snoozed

datetime the forum post was made





Problem: I can't keep reading all the forum posts on Kaggle with my human eyeballs

Problem: I can't keep reading all the forum posts on Kaggle with my

human eyeballs

Solution: Unsupervised clustering to summarize common topics & user

concerns



Some ground rules:

- Needs to be in Python or R
 - I'm livecoding the project in Kernels & those are the only two languages we support
 - I just don't want to use Java or C++ or Matlab whatever
- Needs to be fast to retrain or add new classes
 - New topics emerge very quickly (specific bugs, competition shakeups, ML papers)
 - I'll probably have to re-run it daily or weekly
 - Eventually... streaming?
- Want to avoid large/weird dependencies
 - "Oh, that's just some .jar I downloaded from a random website. The code doesn't run without it and I'm sure it's fine to just stick in our codebase."
- Clusters/topics should be easily interpretable

I asked on Twitter!



Rachael Tatman @rctatman

Follow V

What are y'alls current favorite unsupervised classification/clustering approaches for text? So far I've looked at:

₿ LDA

B Embeddings (doc2vec) + clustering (k-

means

B Unsupervised keyword extraction (YAKE)

Is there something else I should consider?



0 25

11:44 AM - 29 May 2019

1] 22



0 172

Lots of good ideas!

Three main bins:

- End-to-end solutions
- Suggestions for feature engineering
 + clustering
- Misc. tips & tricks (ex: embeddings -> PCA -> remove 1st principle component)

End-to-end solutions

• <u>Gensim</u>

- In Python, no weird dependencies
- \checkmark Old standby that incorporates a looot of differents methods
- Don't need whole corpus in memory (but mine's not that big)
- Under LGPL (probably fine for prototyping, but might need to meet with legal if I'm using it for work stuff)

BigARTM

- ✓ Can incorporate multiple objectives at once (sparsing, smoothing, decorrelation, etc.)
- Weird dependency/install process (it's a C++ library with a Python API)

• <u>TopSBM</u>

- ✓ Came highly recommended: "Scary good"
- Weird dependency (graph-tool, which is C++ with a Python wrapper)

Feature Engineering: Words to numbers

- Traditional Topic Modelling Approaches
 - LDA: Slow, hard to interpret, not my fave
 - **pLSA**: Cheaper version of LSA, tends to overfit
 - **tf-idf**: Hard to interpret, my texts (forums posts) are too short
- Embeddings
 - **GloVe**: considers context, can't handle new words
 - Word2vec: doesn't handle small corpuses very well, very fast to train
 - fasttext: can handle out of vocabulary words (extension of word2vec)
- Contextual embeddings (don't think I have enough data to train my own...)
 - **ELMO, BERT, etc.**: I consider these more of a replacement for language models
 - **USE embeddings**: Not super familiar with this but looks useful for applying to sentence similarity

Feature Engineering: Words to numbers

Traditional Topic Modelling Approaches

- LDA: Slow, hard to interpret, not my fave
- pLSA: Cheaper version of LSA, tends to overfit
- **tf-idf**: Hard to interpret, my texts (forums posts) are too short

• Embeddings

- **GIoVe**: considers context, can't handle new words
- Word2vec: doesn't handle small corpuses very well, very fast to train
- fasttext: can handle out of vocabulary words (extension of word2vec)
- Contextual embeddings (don't think I have enough data to train my own...)
 - ELMO, BERT, etc.: I consider these more of a replacement for language models
 - **USE embeddings**: Not super familiar with this but looks useful for applying to sentence similarity

Feature Engineering: Dimensionality Reduction

• <u>UMAP</u>:

- Recommended to me by, among other people, Leland McInnes, the researcher who developed it limit (he suggested using hellinger distance)
- Similar to t-SNE but can also be used for non-linear dimension reduction
- Something about manifolds? (The math's a little over my head, tbh)
- <u>PCA:</u>
 - OG dimensionality reduction (paper is from 1901!) but on its own maybe not the best
 - Trick: remove first principal component as a way to reduce the weight of "expected" words
 - (from Arora (2018) 'A simple but tough to beat baseline for sentence embeddings')

Feature Engineering: Dimensionality Reduction

• <u>UMAP</u>:

- Recommended to me by, among other people, Leland McInnes, the researcher who developed it limit (he suggested using hellinger distance)
- Similar to t-SNE but can also be used for non-linear dimension reduction
- Something about manifolds? (The math's a little over my head, tbh)
- <u>PCA:</u>
 - OG dimensionality reduction (paper is from 1901!) but on its own maybe not the best
 - Trick: remove first principal component as a way to reduce the weight of "expected" words
 - (from Arora (2018) 'A simple but tough to beat baseline for sentence embeddings')

Wildcard!

- Unsupervised keyword extraction: YAKE
 - Extracts keywords from single texts
 - Could use it as dimensionality reduction
 - Keywords -> embeddings -> clustering?
 - One of their sample texts is about the Kaggle acquisition! ⁽²⁾
 - Haven't played around with it, but came highly recommended
 - o pip install git+https://github.com/LIAAD/yake

Fime spent to run YAKE algorithm 0.37 ms

Annotated text

The top 20 keywords in terms of relevance are annotated in the text.

google is acquiring data science community kaggle

Sources tell us that google is acquiring (aggle), a platform that hosts (data) science and machine learning competitions. Details about the transaction remain somewhat vague, but given that google is hosting its Cloud Next conference in san francisco this week, the official announcement could come as early as tomorrow. Readed by phone, [kaggle] or conumer ce can atthomy goldboom declined to deny that the acquisition is happening. google is tell declined to comment on rumors. [kaggle], which has about thalf a million (data) scientists on its platform), was founded by goldboom and ben hammer in 2010. The service got an early start and even though it has a few competitors like DrivenData, TopCoder and HackerAnk, it has managed to stay well ahead of them by focusing on its specific niche. The service is basically the de facto home for running (data) scientists on its clatform), was founded by goldboom and ben hammer in 2010. The service got an early start and even though it has a few competitors like DrivenData, TopCoder and HackerAnk, it has managed to stay well ahead of them by focusing on its specific niche. The service is basically the de facto home for running (data) scientists - The service is basically the de facto home for running (data) scientists - Earlier this month, google and (laggle) tesmed up to host a \$100.000 machine learning competition and dassifying YouTube videos. That competition had some deep integrations with the google Cloud platform, too. Our understanding is that google will keep the service running likely under its current name. While the acquisition is probably more about (kaggle) scommunity that technology, (kaggle did build some interesting tools for hosting its competition and kernels; too. On Kaggle, kernels are basically the source code for analyzing data is ets and developers can share this code on the platform (the company previously called them 'scripts'). Like similar competition-centric sites (laggle) also runs ajb board, too. It's unclear what google will do wit

Detected language : english

Wildcard!

• <u>Lda2vec</u>

- Embeddings + topic models trained simultaneously
- Developed at StitchFix 3ish years ago
- Still pretty experimental but could be helpful
- Under MIT license
- Has a tutorial notebook
- Might be very slow???



Clustering:

- Brown Clusters
 - Doesn't require feature engineering; can take words directly
 - Hierarchical clusters (could be useful for visualization/exploration)
 - Can be actively updated (wouldn't have to retrain)
- DBSCAN/H(ierarchical)DBSCAN
 - Could take embeddings
 - Clusters assumed to be of similar densities
- Spectral clustering
 - Doesn't make assumptions about spatial distribution of data
 - In sklearn

Clustering:

- Brown Clusters
 - Doesn't require feature engineering; can take words directly
 - Hlerarchical clusters (could be useful for visualization/exploration)
 - Can be actively updated (wouldn't have to retrain)
- DBSCAN/H(ierarchical)DBSCAN
 - Could take embeddings
 - Clusters assumed to be of similar densities
- Spectral clustering
 - Doesn't make assumptions about spatial distribution of data
 - In sklearn

Next stage: Experiments



Next stage: Experiments



Future work

- Slackbot!
 - For now, I'll probably run the code in Kernels
- Other things I want to do as part of this project
 - Identify questions I'm likely to answer
 - $\circ \quad \text{Extend to arbitrary user} \\$
 - Build an alerting system that flags sudden new trends on the forums (competition drama, major bug, etc.)
 - I doooon't want to handle streaming data :weary:

Thanks! I'm very open to feedback/ suggestions :)

@rctatman