PROJECT REPORT
Project: Machine Learning & Deep Learning in Summer term 2022

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Topic Model Evaluation with GPT-3

Saida Yusupova
Technische Universität Kaiserslautern, Department of Computer Science

Topic modeling is a technique for unsupervised analysis of large document corpora. It automatically learns topics, from unlabeled documents, represented as sets of words. In this work, we evaluate three topic models - one classical and two neural - on two datasets. We introduce Generative Pre-trained Transformer 3 (GPT-3) for evaluation of topic model interpretability.

1. Introduction

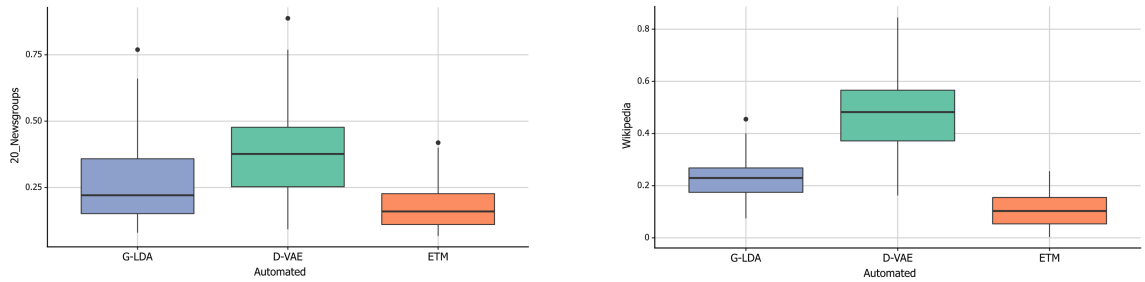
Topic modeling is a set of unsupervised techniques used to analyze text in a collection of documents and identify meaningful group of words, topics [2]. The topic model construction algorithm receives a collection of text documents as input and as an output, for each document, a numerical vector is produced, made up of estimates of the degree of belonging this document to each of the topics. The dimension of this vector, equal to the number of topics, can either be set at the input or determined automatically by the model. It is assumed that each document may relate to one or more topics. Topics differ from each other by varying the frequency of the use of words. The purposes of building a topic models are to help people understand the large corpora, document classification or information retrieval.

Current evaluations of the topic models quality have fluctuated between automated and human assessments, which have shown promising results [1]. As automated coherence metrics, the topic model developers adopted the normalized pointwise mutual information (NPMI) score, which measures word relatedness and correlates with interpretability of the topic [3]. While, obtaining human metrics requires reasonable number of crowdworkers in offline or online mode using survey platforms, so it is time, energy consuming and costly task.

In this work we introduce an approach of topic model evaluation using Generative Pre-trained Transformer 3 (GPT-3). GPT-3 is the third generation autoregressive language model that uses deep learning to process and produce natural language text. The rest of the paper is structured as follows: Section 2 gives an overview of related work and existing evaluation metrics; Section **??** describes the prompts used to get the response from GPT-3; Section 3 shows the results we obtained and their comparison to automated coherence measures; Section 4 explains the main findings of our approach, limitations and future works.

2. Related Work

Before introducing GPT-3 evaluation methods we briefly review the existing work [1], which is the base of our work.



(a) 20 Newsgroups NPMI-values

(b) Wikipedia NPMI-values

Figure 1: Automated evaluations (NPMI) suggest a clear winner between the three models. NPMI declares D-VAE as a winner for topics derived from both datasets, with G-LDA in second place.

2.1. Datasets

Following the original work, we use English articles from Wikipedia with the same settings. However, instead of articles from New York Times, we use The 20 Newsgroups text dataset.

2.2. Models

The model evaluation procedure remains the same. We evaluate one classical model and two neural models:

Gibbs-LDA

3. Results

In our approach, we also test two tasks: intrusion and rating. However, we replace human evaluation with GPT-3 assessment. Appendix A.1 describes the GPT-3 model set up. To calculate the gpt-3 metrics, we first obtain the automated metrics - NPMI values for topics taken from all three models of two datasets[1], Figure 1 shows the models differences. The obtained automated metrics, along with gpt-3 responses used to calculate correlation coefficients.

Intrusion As the first step, we test several versions of a prompt, which are sent to GPT-3, see Appendix A.2 for more details. In Table 1, we show the results of six prompts using the version with high accuracy.
The prompts are:

p_1: Show the least related term

p_2: Select which term is the least related to all other terms

[1] gitlab.rhrk.uni-kl.de/yusupova/topics/-/tree/main/results/readable-format

	20 Newsgroups											
Model →	G-LDA				D-VAE				ETM			
Prompt ↓	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p
1	**0.52**	0.25	0.36	0.27	**0.38**	0.24	**0.27**	0.27	**0.42**	0.24	0.25	0.23
2	**0.52**	0.25	0.35	0.25	0.34	0.22	-0.13	-0.09	0.38	0.24	**0.45**	0.37
3	0.40	0.24	**0.44**	0.33	0.14	0.12	-0.03	-0.002	0.38	0.24	0.15	0.11
4	0.50	0.25	0.39	0.27	0.28	0.20	-0.01	-0.02	**0.42**	0.24	0.42	0.40
5	0.46	0.25	0.40	0.30	0.30	0.21	0.03	0.04	0.38	0.24	0.32	0.29
6	0.42	0.24	0.34	0.35	0.22	0.17	0.12	0.13	0.40	0.24	0.16	0.11

	Wikipedia											
Model →	G-LDA				D-VAE				ETM			
Prompt ↓	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p
1	0.68	0.22	0.23	0.26	**0.86**	0.12	0.17	0.16	0.26	0.19	-0.05	-0.03
2	**0.72**	0.20	0.07	0.14	0.76	0.18	0.19	0.19	**0.28**	0.20	**0.07**	0.09
3	0.48	0.25	0.04	0.05	0.60	0.24	0.20	0.20	0.20	0.16	-0.09	-0.09
4	0.64	0.23	**0.24**	0.27	0.76	0.18	0.19	0.19	0.14	0.12	-0.22	-0.21
5	0.60	0.24	**0.24**	0.28	0.74	0.19	**0.24**	0.24	0.12	0.11	-0.20	-0.19
6	0.56	0.25	0.11	0.07	0.70	0.21	0.23	0.23	0.18	0.15	-0.16	-0.15

Table 1: GPT-3 metrics (intrusion task) for 20 Newsgroups and Wikipedia datasets. **Bold** values have the highest accuracy and correlation coefficient. Underlined values are spearman correlation coefficients between gpt-3 scores and automated metrics, which have p-value < 0.05.

p_3: What is the intruder term in the following terms?

p_4: Which word does not belong?

p_5: Which one of the following words does not belong?

p_6: Find the intruder term

For the intrusion task, we can state that p_1 and p_2, which ask for the least related term, have high accuracies, and the prompt formulation matters.

Rating Following the same procedure we test several versions of a prompt Table 2, more details in Appendix A.3.

	20 Newsgroups											
Model →	G-LDA				D-VAE				ETM			
$p_3 \downarrow$	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p
v_1	2.14	0.80	0.36	0.18	1.44	0.65	**0.27**	0.25	2.02	0.74	**0.14**	0.22
v_2	2.04	0.84	0.38	0.23	1.38	0.60	**0.27**	0.26	2.00	0.76	0.11	0.17
v_3	2.12	0.83	0.35	0.18	1.44	0.69	0.19	0.14	2.12	0.87	0.09	0.15
v_4	1.96	0.80	**0.40**	0.25	1.40	0.64	0.22	0.16	2.04	0.92	0.10	0.18

	Wikipedia											
Model →	G-LDA				D-VAE				ETM			
$p_3 \downarrow$	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p	μ	σ^2	ρ_s	ρ_p
v_1	2.88	0.19	0.25	0.27	2.80	0.32	**0.50**	0.50	nan	nan	nan	nan
v_2	2.78	0.29	**0.35**	0.36	2.68	0.50	0.59	0.58	nan	nan	nan	nan
v_3	2.96	0.08	0.24	0.30	2.80	0.32	0.49	0.49	nan	nan	nan	nan
v_4	2.92	0.15	0.32	0.34	2.70	0.45	0.46	0.43	nan	nan	nan	nan

Table 2: GPT-3 metrics (rating task) for 20 Newsgroups and Wikipedia datasets. **Bold** values have the highest correlation coefficient. Underlined values are spearman correlation coefficients between gpt-3 scores and automated metrics, which have p-value < 0.05.

4. Conclusions

U ALMOST THERE!!!!!!!!!!!!!!!!!!!

References

[1] Denis Peskov Andrew Hian-Cheong Alexander Hoyle, Pravan Goel (2021): *Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. 35th Conference on Neural Information Processing Systems.*

[2] Yuening Hu Jordan Boyd-Graber & David Mimno (2017): *Applications of topic models.*

[3] Karl Grieser David Newman, Jey Han Lau & Timothy Baldwin (2010): *Automatic evaluation of topic coherence. In Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics.*

A. Appendix

A.1. Details on GPT-3 model set up

The set up is delineated in our implementation:[2]

[2] gitlab.rhrk.uni-kl.de/yusupova/topics

- We provide api-key to GPT-3 in file 'gpt3-api-key.txt'.
- We use 'text-davinci-003' engine.
- We set temperature to 0 for deterministic responses.
- Maximum number of response tokens we set 10 for intrusion task, 60 for rating task.
- We set 'frequency_penalty' to 0, so that we don't penalize new tokens based on their existing frequency in the text.
- Also, we set 'presence_penalty' to 0 and don't penalize new tokens based on their appearance in the text.

A.2. Details on prompt variations for intrusion task

The prompt versions, versions with respect to the p_1, are:

v_1: Show the least related term: ['construction', 'locomotives', 'cantata', 'coaster', 'railway', 'trains']

v_2: Show the least related term: 'construction', 'locomotives', 'cantata', 'coaster', 'railway', 'trains'

v_3: Show the least related term: [construction, locomotives, cantata, coaster, railway, trains]

v_4: Show the least related term: construction, locomotives, cantata, coaster, railway, trains

v_5: Show the least related term
Terms: ['construction', 'locomotives', 'cantata', 'coaster', 'railway', 'trains']
Answer:

v_6: Show the least related term
Terms: 'construction', 'locomotives', 'cantata', 'coaster', 'railway', 'trains'
Answer:

v_7: Show the least related term
Terms: [construction, locomotives, cantata, coaster, railway, trains]
Answer:

v_1: Show the least related term
Terms: construction, locomotives, cantata, coaster, railway, trains
Answer:

Versions from v_1 to v_4 represents different set up of square brackets and single quotes, while v_5 to v_8 represent more structured way. We randomly choose two prompts and one prompt, p_2, similar to the question the crowdworkers were asked to determine intruder term[3] in original experiment [1]. Additionally, as the D-VAE fared better on the intrusion task, we perform our initial evaluations on this model for Wikipedia dataset. We test all eight versions and the Table 3 shows the results. We see that v_6 has the highest accuracy in all three versions, so in further testings of the prompts we use v_6.

[3]The topic-intruder files along with the gpt-3 responses can be found at gitlab.rhrk.uni-kl.de/yusupova/topics/-/tree/main/transformer-tests

Version → Prompt ↓	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
1	0.68	0.70	0.66	0.68	0.82	**0.86**	0.82	0.84
2	0.70	0.66	0.64	0.64	0.74	**0.76**	0.72	0.74
3	0.40	0.46	0.42	0.42	0.52	**0.60**	0.56	0.52

Table 3: Accuracy results for three prompts of D-VAE model for topics derived from Wikipedia in intrusion task.

To note, in the initial stage of our experiment, the following prompt were also under consideration, as the transformer were given reasonable responses only if the prompt consisted the responses that we were expecting:

Show the least related term
Terms: image, object, pixel, face, scene, privacy
Answer: privacy

Show the least related term
Terms: graph, edge, message, cell, vertex, propagation
Answer: cell

Show the least related term
Terms: construction, locomotives, cantata, coaster, railway, trains
Answer:

With the new model, 'text-davinci-003', there were no need in further use of this prompt as the responses received using this engine were reasonable and in one-word format for intrusion task.

A.3. Details on prompt variations for rating task

Same as in intrusion task, we test the effect of brackets and single quotes on gpt-3 responses. The prompts are:

p_1: Rate how related the following terms are to each other as 'very related', 'somewhat related'or 'not related': *top 10 words of a topic*

p_2: Rate how related the following terms are to each other in a range from 1 to 3: *top 10 words of a topic*

For above two prompts, the gpt-3 does not give reasonable responses. For a topic terms

['rower', 'hammersmith_bridge', 'rowed', 'mile_post', 'rowing', 'cambridge', 'boat_race', 'chiswick_steps', 'oxford', 'university_of_oxford']

, the following are the examples of such responses (only two terms were considered by gpt-3 or terms were considered pairwise):

Rower and Rowing: Very Related OR *rower: 3hammersmith_*

rower and rowing: very relatedhammersmith bridge and chiswick steps: very relatedmile post and cambridge: not relatedboat race and university of oxford: somewhat related

OR

rower: 3hammersmith_bridge: 1rowed: 3mile_post: 1rowing: 3cambridge: 1boat_race: 2chiswick_steps: 1oxford: 2university_of_oxford

For the following prompt p_3, where we show the gpt-3 the format of the response we are expecting, we also manipulate brackets-quotes and Table shows the results:

p_3: Rate how related the following terms are to each other as '3-very related', '2-somewhat related'or '1-not related': ['file', 'window', 'problem', 'run', 'system', 'program', 'font', 'work', 'win', 'change']
Answer: Very related

Rate how related the following terms are to each other as '3-very related', '2-somewhat related'or '1-not related': ['chip', 'clipper', 'phone', 'key', 'encryption', 'government', 'system', 'write', 'nsa', 'communication']
Answer: somewhat related

Rate how related the following terms are to each other as '3-very related', '2-somewhat related'or '1-not related': *top 10 words of a topic*