


# Generowanie tytułów publikacji naukowych

Przetwarzanie Języków Naturalnych

Szymon M.


<https://github.com/csinva/gpt-paper-title-generator>

 **gpt-paper-title-generator** Public

Watch 4 Fork 28 Star 121

master 2 branches 0 tags

Go to file Add file Code

 **csinva** update date ✓ cb80559 on Oct 20, 2022 66 commits

data	add authors_dict_titles	7 months ago
docs	update date	7 months ago
gpt2	add figs	7 months ago
gpt3	add figs	7 months ago
gptneo	update readme	7 months ago
samples	cache CDNs	7 months ago
scrape	more refactoring	7 months ago
.gitignore	move web to docs folder	7 months ago
readme.md	update readme	7 months ago

readme.md

Forecasting the progress of research is an elusive and important goal. Here, we take a toy step towards this goal by exploring generating new scientific paper titles given past titles on arXiv:

**About**

Generating paper titles (and more!) with GPT trained on data scraped from arXiv.

[csinva.io/gpt-paper-title-generator](https://csinva.io/gpt-paper-title-generator)

machine-learning natural-language-processing ai deep-learning neural-network tensorflow ml text-generation artificial-intelligence transformer tensorflow-experiments nlp-machine-learning generative-models title-generation gpt-2

Readme 121 stars 4 watching 28 forks Report repository



Cornell University



arXiv is a free distribution service and an open-access archive for 2,254,199 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

**Subject search and browse:**

Physics



Search

Form Interface

Catchup

**News**

Read about recent news and updates on [arXiv's blog](#). (View the former "what's new" pages here).

## Showing 1–50 of 155,825 results for all: Machine Learning

Search v0.5.6 released 2020-02-24 [Feedback?](#)Searching in archive **cs**. Search in all archives.

Machine Learning

All fields



Search

 Show abstracts  Hide abstracts[Advanced Search](#)50 results per page. Sort results by Announcement date (newest first) [Go](#)[1](#) [2](#) [3](#) [4](#) [5](#) ...[Next](#)1. [arXiv:2305.07013](#) [[pdf](#), [ps](#), [other](#)] [cs.IT](#)**Computing Unique Information for Poisson and Multinomial Systems****Authors:** Chaitanya Goswami, Amanda Merkle, Pulkit Grover

**Abstract:** ...Y and Z is decomposed into unique, redundant, and synergistic terms. Recently, PID has shown promise as an emerging tool to understand biological systems and biases in **machine learning**. However, computing PID is a challenging problem as it typically involves optimizing over distributions. In this work, we study the pro... [More](#)

**Submitted** 11 May, 2023; **originally announced** May 2023.2. [arXiv:2305.07005](#) [[pdf](#), [other](#)] [cs.CL](#)**Subword Segmental Machine Translation: Unifying Segmentation and Target Sentence Generation****Authors:** Francois Meyer, Jan Buys

**Abstract:** Subword segmenters like BPE operate as a preprocessing step in neural **machine** translation and other (conditional) language models. They are applied to datasets before training, so translation or text generation quality relies on the quality of segmentations. We propose a departure from this paradigm, called subword segmental... [More](#)

**Submitted** 11 May, 2023; **originally announced** May 2023.3. [arXiv:2305.06994](#) [[pdf](#), [other](#)] [cs.LG](#) [cs.CY](#)



# arXiv Dataset

arXiv dataset and metadata of 1.7M+ scholarly papers across STEM



Data Card

Code (86)

Discussion (45)

## About Dataset

### About ArXiv

For nearly 30 years, [ArXiv](#) has served the public and research communities by providing open access to scholarly articles, from the vast branches of physics to the many subdisciplines of computer science to everything in between, including math, statistics, electrical engineering, quantitative biology, and economics. This rich corpus of information offers significant, but sometimes overwhelming depth.

In these times of unique global challenges, efficient extraction of insights from data is essential. To help make the arXiv more accessible, we present a free, open pipeline on Kaggle to the machine-readable arXiv dataset: a repository of 1.7 million articles, with relevant features such as article titles, authors, categories, abstracts, full text PDFs, and more.

Our hope is to empower new use cases that can lead to the exploration of richer machine learning techniques that combine multi-modal features towards applications like trend analysis, paper recommender engines, category prediction, co-citation networks, knowledge graph construction and semantic search interfaces.

The dataset is freely available via Google Cloud Storage buckets ([more info here](#)). Stay tuned for weekly updates to the dataset!

### Usability 🕒

8.75

### License

[CC0: Public Domain](#)

### Expected update frequency

Monthly



```

import json

attributes = ["authors", "title", "update_date", "categories"]

papers = []
count = 0
with open('arxiv-metadata-oai-snapshot.json') as f:
    for line in f:
        count += 1
        if count % 100000 == 0:
            print(f'on paper {count}')
        json_object = json.loads(line)
        paper = {}
        for attr in attributes:
            paper[attr] = json_object[attr]
        papers.append(paper)

import pandas as pd
df = pd.DataFrame(papers)

df

```

✓ 32.8s

	authors	title	update_date	categories
0	C. Bal'azs, E. L. Berger, P. M. Nadolsky, C.-...	Calculation of prompt diphoton production cros...	2008-11-26	hep-ph
1	Ileana Streinu and Louis Theran	Sparsity-certifying Graph Decompositions	2008-12-13	math.CO cs.CG
2	Hongjun Pan	The evolution of the Earth-Moon system based o...	2008-01-13	physics.gen-ph
3	David Callan	A determinant of Stirling cycle numbers counts...	2007-05-23	math.CO
4	Wael Abu-Shammala and Alberto Torchinsky	From dyadic $\lambda_\alpha$ to $\lambda_{\alpha}$	2013-10-15	math.CA math.FA
...	...	...	...	...
2250218	R. Prozorov, M. Konczykowski, B. Schmidt, Y. Y...	On the origin of the irreversibility line in t...	2009-10-30	supr-con cond-mat.supr-con
2250219	Durga P. Choudhury, Balam A. Willemsen, John S...	Nonlinear Response of HTSC Thin Film Microwave...	2016-11-18	supr-con cond-mat.supr-con
2250220	Balam A. Willemsen, J. S. Derov and S.Sridhar ...	Critical State Flux Penetration and Linear Mic...	2009-10-30	supr-con cond-mat.supr-con
2250221	Yasumasa Hasegawa (Himeji Institute of Technol...	Density of States and NMR Relaxation Rate in A...	2009-10-30	supr-con cond-mat.supr-con
2250222	Naoki Enomoto, Masanori Ichioka and Kazushige ...	Ginzburg Landau theory for d-wave pairing and ...	2009-10-30	supr-con cond-mat.supr-con

2250223 rows × 4 columns



```

ai_papers = [paper for paper in papers if ("cs.AI" in paper["categories"] or "stat.ML" in paper["categories"] or "cs.LG" in paper["categories"])
df = pd.DataFrame(ai_papers)

import re
regex = re.compile(f" and |,")

df["authors"] = [[str(a).strip() for a in regex.split(author) if len(str(a).strip()) > 3] for author in df["authors"]]
df = df.sort_values(by="update_date", ascending=False)

```

df  
✓ 3.2s

	authors	title	update_date	categories
168839	[Shuting Shen, Xi Chen, Ethan X. Fang, Junwei Lu]	Combinatorial Inference on the Optimal Assortm...	2023-05-05	stat.ML cs.LG
179649	[Ming-Kun Xie, Jia-Hao Xiao, Gang Niu, Masashi...	Class-Distribution-Aware Pseudo Labeling for S...	2023-05-05	cs.LG
169390	[Alon Albalak, Colin Raffel, William Yang Wang]	Improving Few-Shot Generalization by Exploring...	2023-05-05	cs.LG cs.CL
179618	[UngJin Na, Moonhee Choi, HangJin Jo]	Critical heat flux diagnosis using conditional...	2023-05-05	physics.flu-dyn cs.LG
171498	[Soham Rohit Chitnis, Sidong Liu, Tirtharaj Da...	Domain-Specific Pre-training Improves Confiden...	2023-05-05	cs.CV cs.AI cs.LG
...	...	...	...	...
180173	[Petro M. Gopych]	A Neural Network Assembly Memory Model Based o...	2007-05-23	cs.AI cs.IR cs.NE q-bio.NC q-bio.QM
180172	[Peter D. Turney, Michael L. Littman, Jeffrey ...	Combining Independent Modules to Solve Multipl...	2007-05-23	cs.CL cs.IR cs.LG
180171	[Peter D. Turney (National Research Council of...	Measuring Praise and Criticism: Inference of S...	2007-05-23	cs.CL cs.IR cs.LG
180170	[Wolfgang Mayer, Markus Stumptner]	Model-Based Debugging using Multiple Abstract ...	2007-05-23	cs.SE cs.AI
1	[T. Kosel, I. Grabec]	Intelligent location of simultaneously active ...	2007-05-23	cs.NE cs.AI

181083 rows × 4 columns



```

authors_with_papers = {}

for paper in df.values:
    authors = paper[0]
    title = paper[1]
    for author in authors:
        if author in authors_with_papers:
            authors_with_papers[author].append(title)
        else:
            authors_with_papers[author] = [title]

output = pd.DataFrame([{"author": author, "papers": authors_with_papers[author], "paper_count": len(authors_with_papers[author])} for author in authors_with_papers.keys()])

output.sort_values(by="paper_count", ascending=False)

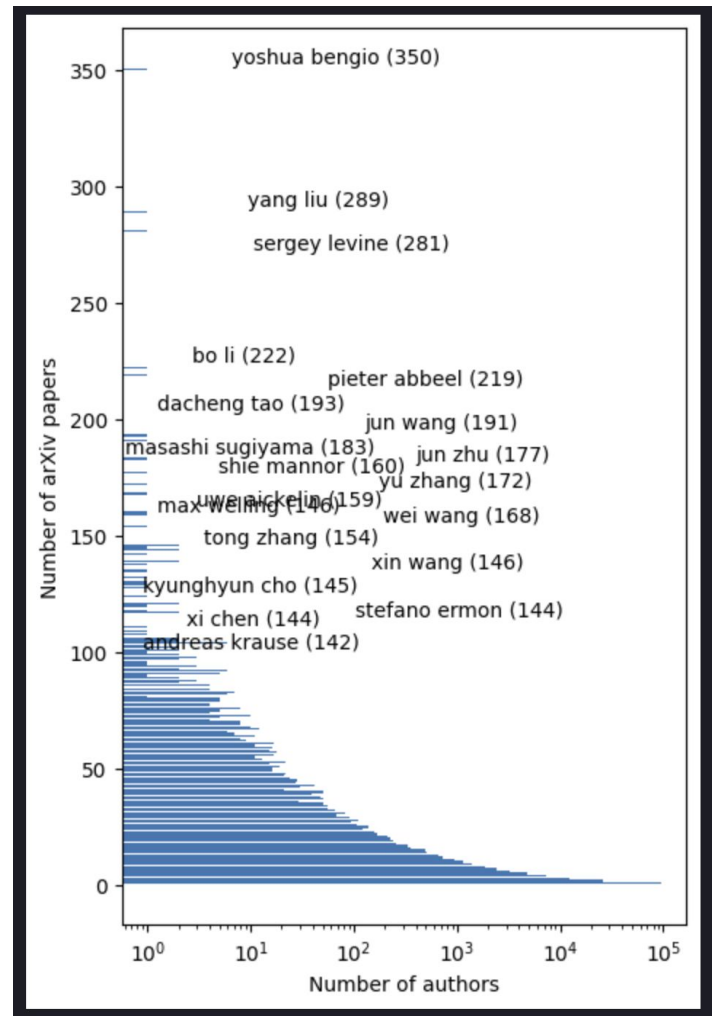
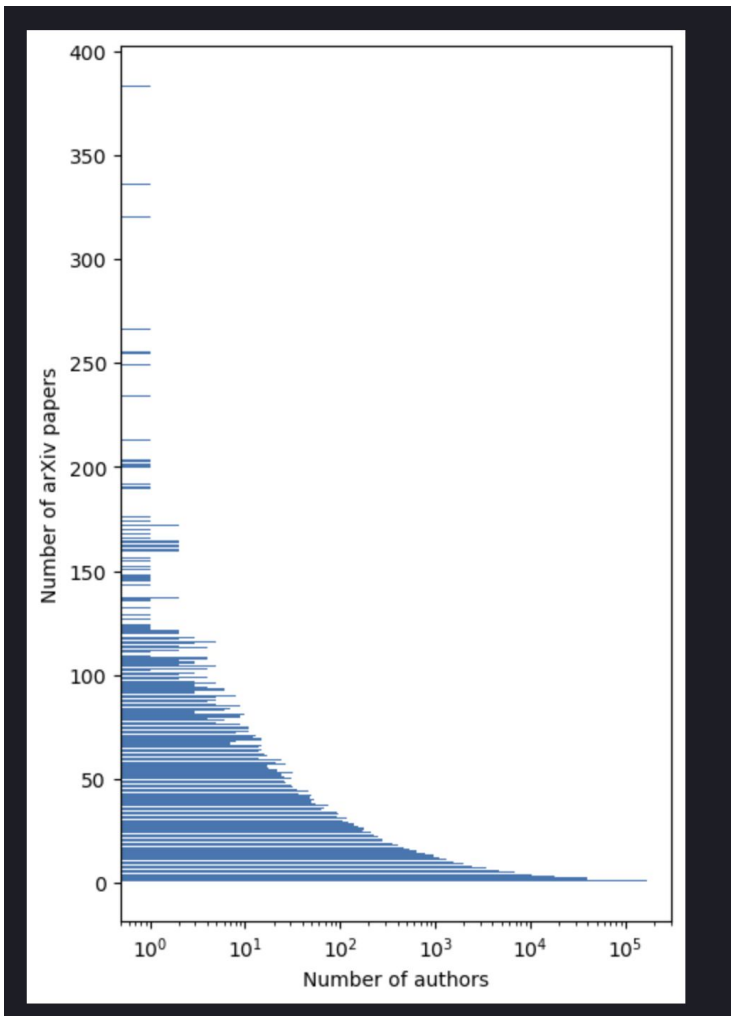
```

✓ 2.9s

Python

	author	papers	paper_count
10152	Yoshua Bengio	[Hyena Hierarchy: Towards Larger Convolutional...	383
2130	Yang Liu	[Performative Prediction with Bandit Feedback:...	336
3403	Sergey Levine	[Offline RL for Natural Language Generation wi...	320
4749	Bo Li	[Model Explainability in Physiological and Hea...	266
15088	Michael I. Jordan	[Online Learning in Stackelberg Games with an ...	255
...	...	...	...
128610	Igor Khokhlov	[Attacks, Defenses, And Tools: A Framework To ...	1
128611	Mehdi Mirakhorli	[Attacks, Defenses, And Tools: A Framework To ...	1
128616	Songqing Yue	[Imbalanced Malware Images Classification: a C...	1
128619	Zhengren Wang	[Listing Maximal k-Plexes in Large Real-World ...	1
267544	Markus Stumptner	[Model-Based Debugging using Multiple Abstract...	1

267545 rows × 3 columns



# 1 Author-specific paper titles (prompting gpt3)

To generate author-specific titles, we take the five most recent titles from each author with at least 3 arXiv AI papers (cs.ML, cs.LG, stat.ML). We then format the papers using the following template and query for a new title using GPT-3:

Here is a list of related machine-learning papers:

```
> [title 1]
> [title 2]
...
> [title 5]
> _____
```

```
for i, author in enumerate(tqdm(authors)):
    ... if not author in authors_save:
        ... query = prompt + '\n>'.join(authors_dict_titles[author][-papers_in_context:]) + '\n>'
        ... completion = openai.Completion.create(
        ...     engine="text-davinci-002", prompt=query,
        ...     n=gens_per_author, stop='>'
        ... )
        ... authors_save[author] = [completion.choices[i].text for i in range(len(completion.choices))]
    ... if i % 200 == 0:
        ... pkl.dump(authors_save, open(f'gen_titles/authors_save_{i}.pkl', 'wb'))
pkl.dump(authors_save, open(f'gen_titles/authors_save_full.pkl', 'wb'))
```

Chandan Singh, 7 months

Here's a concrete example -- when prompting with these 5 recent titles:

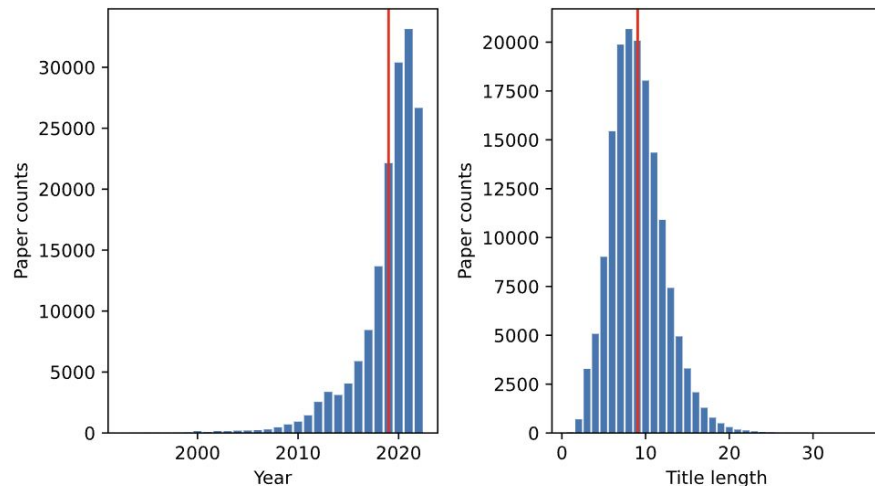
- > Hierarchical Shrinkage: improving the accuracy and interpretability of tree-based methods
- > Fast Interpretable Greedy-Tree Sums (FIGS)
- > Adaptive wavelet distillation from neural networks through interpretations
- > Emb-GAM: an Interpretable and Efficient Predictor using Pre-trained Language Models
- > Explaining Patterns in Data with Language Models via Interpretable Autoprompting
- > \_\_\_\_\_

We get these 5 (independent) random generations for the blank:

1. Towards Interpretable Natural Language Processing: A Survey
2. A Unified Framework for Interpretable Machine Learning
3. Compositional Attention Networks for Machine Reasoning
4. Achieving Open Vocabulary Neural Machine Translation
5. A Deep Understanding of Neural Networks through Deep Visualization

## 2 Finetuned paper title generation (gptneo)


To improve the model's ability to generate cogent titles, we finetune it on a large corpus of titles. We start from the [gpt-neo-2.7B checkpoint](#) (see our [training script](#) for hyperparameters). We finetune on all [paper titles on arXiv](#) in the categories cs.AI, cs.LG, stat.ML up to Oct 13, 2022. We exclude all papers after Apr 1, 2022 (to test the ability to forecast new papers) and an additional random 5% of titles. We also exclude titles with a length of less than 6 words or greater than 20 words. This results in 98,388 papers for finetuning:




<https://huggingface.co/docs/transformers/index>

## Transformers


State-of-the-art Machine Learning for PyTorch, TensorFlow, and JAX.


 Transformers provides APIs and tools to easily download and train state-of-the-art pretrained models. Using pretrained models can reduce your compute costs, carbon footprint, and save you the time and resources required to train a model from scratch. These models support common tasks in different modalities, such as:


 **Natural Language Processing:** text classification, named entity recognition, question answering, language modeling, summarization, translation, multiple choice, and text generation.

 **Computer Vision:** image classification, object detection, and segmentation.

 **Audio:** automatic speech recognition and audio classification.

 **Multimodal:** table question answering, optical character recognition, information extraction from scanned documents, video classification, and visual question answering.

 Transformers support framework interoperability between PyTorch, TensorFlow, and JAX. This provides the flexibility to use a different framework at each stage of a model's life; train a model in three lines of code in one framework, and load it for inference in another. Models can also be exported to a format like ONNX and TorchScript for deployment in production environments.

EleutherAI / **gpt-neo-2.7B**  like 329

 Text Generation  PyTorch  JAX  Rust  Transformers  the\_pile  English  gpt\_neo  text generation  causal-lm  License: mit

**Model card** Files and versions Community  10

 Train  Deploy  Use in Transformers

## GPT-Neo 2.7B

### Model Description

GPT-Neo 2.7B is a transformer model designed using EleutherAI's replication of the GPT-3 architecture. GPT-Neo refers to the class of models, while 2.7B represents the number of parameters of this particular pre-trained model.

### Training data

GPT-Neo 2.7B was trained on the Pile, a large scale curated dataset created by EleutherAI for the purpose of training this model.

### Training procedure

This model was trained for 420 billion tokens over 400,000 steps. It was trained as a masked autoregressive language model, using cross-entropy loss.

 Edit model card

Downloads last month  
**147,358**



### Hosted inference API

 Text Generation

Examples 

My name is Julien and I like to make money with the help of Bitcoin.

To do that, I am currently offering the following services:

I will start by creating Bitcoins for you, then providing you with a secure Bitcoin

Compute 

0,8


Computation time on gpu: cached

 JSON Output

 Maximize





csinva / **gpt-neo-2.7B-titles**  like 0

 Text Generation  PyTorch  Transformers  gpt\_neo  License: apache-2.0

**Model card**  Files and versions  Community

  Train  Deploy  Use in Transformers

 Edit model card

Full code and details at <https://github.com/csinva/gpt-paper-title-generator>

## Model

- finetunes starting from the [gpt-neo-2.7B checkpoint](#)
  - for training details see [the training script](#)
- inference
  - should prepend with a year and two newlines before querying for a title, e.g. 2022\n\n

```
from transformers import AutoModelForCausalLM, pipeline, AutoTokenizer
model = AutoModelForCausalLM.from_pretrained("csinva/gpt-neo-2.7B-titles")
tokenizer = AutoTokenizer.from_pretrained("EleutherAI/gpt-neo-2.7B")
pipe = pipeline('text-generation', model=model, tokenizer=tokenizer)
pipe('2022\n\n')
```

Downloads last month

2



## ⚡ Hosted inference API

 Text Generation

Examples 

2021

Compute

ctrl+Enter

0, 2

This model can be loaded on the Inference API on-demand.

 JSON Output

 Maximize

## 2022

- Diverse Datasets for Learning to Rank via Knowledge Graph Embedding
- Machine learning-driven method for high-throughput single-cell analysis of differentiation and lineage commitment
- On the Sample Complexity of Differentially Private Learning
- Data-Dependent Weight Normalization for Improved Image Resolution
- Adaptive Densely Connected Networks for the Generation and Visualization of Object Deformations
- Exploring the Implicit Bias in Transfer Learning using Imitation Learning

## 2023 (These samples tend to just be similar to 2021/2022 where the majority of the training data lies)

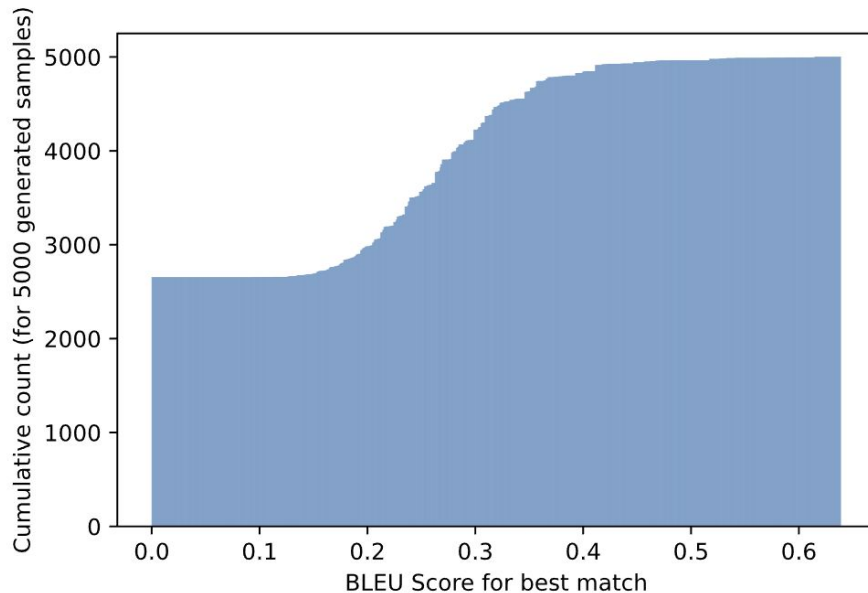
- An Interpretable Dynamic Network for Spatiotemporal Pattern Prediction in High-Dimensional Time Series Data
- Multimodal Deep Learning for Automated Cancer Histopathology Analysis
- Reinforced Learning for Robust and Accurate Object Detection
- A Machine Learning Approach to High Sensitivity Data Processing
- Adversarial Robustness for Graph Neural Networks in Network Intrusion Detection
- Reinforcement Learning via Exploration and Rehearsal for Learning from Demonstrations

## 2010 (Seems to properly generate older titles)

- Learning in a Dynamic, Clustered and Homogeneous Heterogeneous Markov Decision Process
- An Empirical Analysis of the Regularization of the Gaussian Process Regression
- A Scalable Clustering Algorithm under Heterogeneous Data
- A Unified Representation for Probabilistic Time Series Forecasting
- A Hybrid Approach to Automatic Alignment and Localization of Digital Object Platforms
- Bayesian nonparametric modeling of random fields

### 3 Generated paper evaluation

We now evaluate whether the generated titles for 2022 match the real paper titles from the test set (April 1 - Oct 13 2022). Note that the model has never seen any papers from this time period and it's pre-training corpus also only contained text from before 2022. We generate 5,000 titles and find for the closest match for each of them in the test set (which contains ~15,000 titles). The resulting BLEU scores are shown in this figure:





# BLEU [\[edytuj\]](#)

🌐 10 języków ▾

[Artykuł](#) [Dyskusja](#)

[Czytaj](#) [Edytuj](#) [Edytuj kod źródłowy](#) [Wyświetl historię](#) [Narzędzia](#) ▾

**BLEU (Bilingual Evaluation Understudy)** – algorytm do ewaluacji jakości tłumaczenia automatycznego z jednego języka naturalnego na inny. Jakość jest rozumiana jako korelacja między danymi wyjściowymi a tekstem ludzkim: „im bliższe tłumaczenie automatyczne jest profesjonalnemu tłumaczeniu ludzkiemu, tym jest lepsze”<sup>[1]</sup>. BLEU był jedną z pierwszych metryk, która uzyskała wysoką korelację z ludzkim osądem jakości<sup>[1][2]</sup>. Pozostaje także najbardziej popularną z metod.

Punkty liczone są dla pojedynczych przetłumaczonych segmentów – zwykle zdań – przez porównanie ich ze zbiorem tłumaczeń referencyjnych dobrej jakości. Punkty te są następnie uśredniane w obrębie całego korpusu, aby oszacować całkowitą jakość tłumaczenia. Pod uwagę nie są brane zrozumiałość oraz poprawność gramatyczna.

BLEU jest zaprojektowany, aby przybliżyć ludzką ocenę na poziomie dużych korpusów i nie sprawdza się do oceny pojedynczych zdań.

**A**

Understanding the effect of data augmentation in generative adversarial networks

Adversarial attacks on graph neural networks

Differentiable reinforcement learning for continuous control

Multilevel representation learning for time series forecasting

Unsupervised feature learning for medical image segmentation

**B**

Understanding the effect of data augmentation in self-supervised anomaly detection

Sparse vicious attacks on graph neural networks

Normality-guided distributional reinforcement learning for continuous control

Out-of-distribution representation learning for time series classification

Distributed contrastive learning for medical image segmentation