

Kraków, 19.12.2023

Raport

Generowanie tytułów publikacji naukowych w oparciu o zbiór z archiwum arXiv oraz model GPT-2

Autorzy:

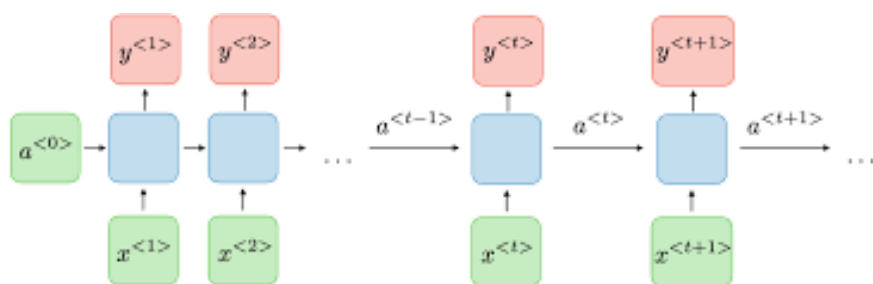
- Bartosz Kniaziewicz
- Krzysztof Kołodziejczyk

1. Opis badanego problemu

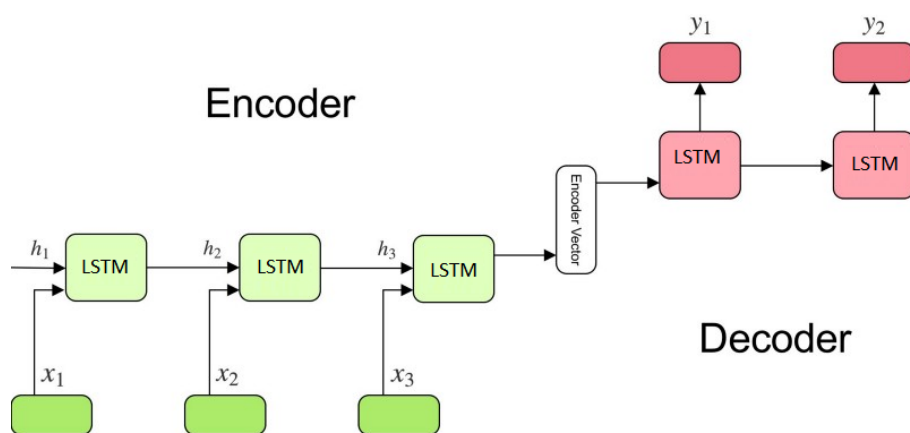
Niniejszy raport skupia się na próbie opracowania rozwiązania umożliwiającego generowanie tytułów publikacji naukowych, wykorzystując dane tytułów publikacji z archiwum ArXiv oraz model GPT-2.

2. Wstęp teoretyczny i możliwe podejścia

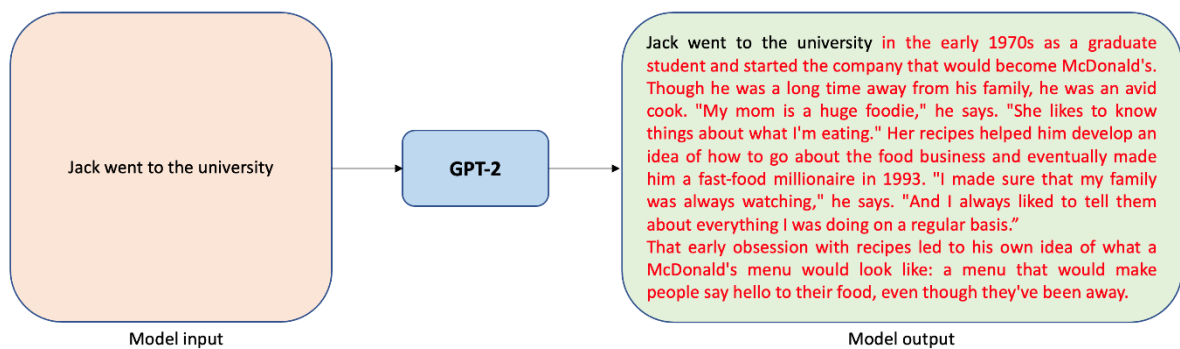
Celem projektu jest prognozowanie postępu badań naukowych w dziedzinie przetwarzania języka naturalnego (NLP). Problem należy do problemów z kategorii generowania tekstu. Najpopularniejszymi modelami używanymi do generowania tekstu są modele RNN (Recurrent Neural Networks), LSTM (Long Short Term Memory) i ostatnio Transformer (GPT-2 jest jednym z nich). Istnieją różne strategie do generowania tekstu, takie jak użycie pełnego zdania jako wejścia dla modelu do przewidzenia następnego słowa, użycie tylko poprzedniego słowa do przewidzenia następnego słowa lub użycie oznakowania początku i końca zdania.



Rys.1 Architektura sieci RNN



Rys.2 Architektura sieci LSTM



Rys.3 Działanie modelu GPT-2 (generowanie tekstu)

3. Opis danych

Dane, z którymi pracujemy, pochodzą z archiwum ArXiv, globalnej bazy danych zawierającej ponad milion artykułów z różnych dziedzin nauki. Każdy rekord w zbiorze danych zawiera tytuł artykułu, abstrakt oraz inne metadane, takie jak autorzy, data publikacji, kategoria itd. Dane zostały pobrane bezpośrednio z datasetu ze strony [Kaggle](https://www.kaggle.com/datasets/arXiv/arXiv):

[20]: `df.head(5)`

[20]:

	id	submitter	authors	title	comments	journal-ref	doi	report-no	categories
0	704.0001	Pavel Nadolsky	C. Bal'azs, E. L. Berger, P. M. Nadolsky, C.-...	Calculation of prompt diphoton production cros...	37 pages, 15 figures; published version	Phys.Rev.D76:013009,2007	10.1103/PhysRevD.76.013009	ANL-HEP-PR-07-12	hep-ph
1	704.0002	Louis Theran	Ileana Streinu and Louis Theran	Sparsity-certifying Graph Decompositions	To appear in Graphs and Combinatorics	None	None	None	math.CO cs.CG
2	704.0003	Hongjun Pan	Hongjun Pan	The evolution of the Earth-Moon system based o...	23 pages, 3 figures	None	None	None	physics.gen-ph
3	704.0004	David Callan	David Callan	A determinant of Stirling cycle numbers counts...	11 pages	None	None	None	math.CO
4	704.0005	Alberto Torchinsky	Wael Abu-Shammala and Alberto Torchinsky	From dyadic Λ_α to $\Lambda_{\alpha,\infty}$	None	Illinois J. Math. 52 (2008) no.2, 681-689	None	None	math.CA math.FA

Rys 4. Fragment danych o publikacjach naukowych z arXiv

	license	abstract	versions	update_date	authors_parsed
	None	A fully differential calculation in perturbation theory	[{'version': 'v1', 'created': 'Mon, 2 Apr 2007...'}]	2008-11-26	[[Balázs, C.,], [Berger, E. L.,], [Nadolsky, ...]]
tp://arxiv.org/licenses/nonexclusive-distrib...		We describe a new algorithm, the (k, ℓ) -...	[{'version': 'v1', 'created': 'Sat, 31 Mar 2007...'}]	2008-12-13	[[Streinu, Ileana,], [Theran, Louis,]]
	None	The evolution of Earth-Moon system is described	[{'version': 'v1', 'created': 'Sun, 1 Apr 2007...'}]	2008-01-13	[[Pan, Hongjun,]]
	None	We show that a determinant of Stirling cycle...	[{'version': 'v1', 'created': 'Sat, 31 Mar 2007...'}]	2007-05-23	[[Callan, David,]]
	None	In this paper we show how to compute the S^1 -...	[{'version': 'v1', 'created': 'Mon, 2 Apr 2007...'}]	2013-10-15	[[Abu-Shammala, Wael,], [Torchinsky, Alberto,]]

4. Przetwarzanie danych

Pierwszym krokiem w przedstawionym procesie jest przygotowanie danych, które mogą być użyte do treningu modelu. W celu skrócenia czasu trenowania oraz lepszej możliwości zweryfikowania wyników wygenerowanych tekstów bierzemy pod uwagę wybraną kategorię (w naszym przypadku wybraliśmy math.MP, co oznacza Mathematical Physics), a następnie przekształcamy wszystkie teksty na małe litery.

```
df_title_gen['categories'].unique()

array(['hep-ph', 'math.CO', 'cs.CG', 'physics.gen-ph', 'math.CA',
       'math.FA', 'cond-mat.mes-hall', 'gr-qc', 'cond-mat.mtrl-sci',
       'astro-ph', 'math.NT', 'math.AG', 'math.AT', 'hep-th', 'math.PR',
       'hep-ex', 'nlin.PS', 'physics.chem-ph', 'q-bio.MN', 'math.NA',
       'cond-mat.str-el', 'cond-mat.stat-mech', 'math.RA',
       'physics.optics', 'physics.comp-ph', 'q-bio.PE', 'q-bio.CB',
       'quant-ph', 'q-bio.QM', 'hep-lat', 'nucl-th', 'math.OA', 'math.QA',
       'math-ph', 'math.MP', 'nlin.CO', 'physics.plasm-ph',
       'physics.space-ph', 'nlin.SI', 'cs.IT', 'math.IT', 'cs.NE',
       'cs.AI', 'physics.ed-ph', 'math.DG', 'cond-mat.soft',
       'physics.pop-ph', 'cs.DS', 'math.CV', 'math.DS', 'physics.soc-ph',
       'nucl-ex', 'math.RT', 'cond-mat.other', 'physics.flu-dyn',
       'physics.data-an', 'cs.CE', 'cs.MS', 'cs.NA', 'math.GR',
       'cond-mat.supr-con', 'math.AC', 'math.SG', 'cs.CC', 'math.KT',
       'math.GT', 'math.AP', 'physics.class-ph', 'q-bio.OT',
       'physics.bio-ph', 'q-bio.BM', 'nlin.CG', 'cs.DM', 'cs.LO',
       'cond-mat.dis-nn', 'math.MG', 'physics.atom-ph', 'math.SP',
       'math.ST', 'stat.TH', 'physics.aos-ph', 'physics.ins-det',
       'q-fin.CP', 'q-fin.PR', 'physics.geo-ph', 'q-bio.NC', 'q-fin.RM',
       'q-bio.SC', 'astro-ph.HE', 'math.OC', 'cs.CR', 'math.CT',
       'math.LO', 'cs.NI', 'q-fin.GN', 'q-fin.ST', 'cs.LG', 'cs.PF',
       'stat.ME', 'stat.AP', 'math.GM', 'physics.atm-clus', 'cs.SE',
       'physics.acc-ph', 'math.GN', 'stat.CO', 'physics.hist-ph', 'cs.AR',
       'cs.SC', 'physics.med-ph', 'stat.ML', 'cs.CY', 'cs.IR', 'q-bio.GN',
       'cs.CV', 'math.HO', 'cs.OH', 'cs.DB', 'cs.DL', 'cs.HC', 'cs.PL',
       'nlin.AO', 'cs.GT', 'cs.DC', 'cond-mat.quant-gas', 'cs.MA',
       'cs.CL', 'q-fin.PM', 'cs.MM', 'astro-ph.EP', 'cs.RO', 'econ.EM',
       'cs.ET', 'q-bio.TO', 'cs.GL', 'astro-ph.SR', 'astro-ph.CO',
       'cs.FL', 'cs.OS', 'q-fin.TR', 'astro-ph.IM', 'cs.SD'], dtype=object)
```

Rys. 5 Lista wszystkich kategorii (w 10k pierwszych artykułach)

```
[35]: titles

[35...] array(['quantum group of isometries in classical and noncommutative geometry',
       'the decomposition method and maple procedure for finding first integrals\n of nonlinear pdes of any order
       with any number of independent variables',
       'quantum deformations of relativistic symmetries',
       'conformal field theory and operator algebras',
       'stringy jacobi fields in morse theory',
       'on the total disconnectedness of the quotient aubry set',
       'a rigorous time-domain analysis of full-wave electromagnetic cloaking\n (invisibility)',
       'dimers on surface graphs and spin structures. ii',
       'mathematics of thermoacoustic tomography',
       'the arctic circle revisited',
       'thermodynamic stability - a note on a footnote in ruelle's book',
       'a variational formulation of electrodynamics with external sources',
       'the veldkamp space of two-qubits',
       'on universality of critical behaviour in the focusing nonlinear\n schr\\\"odinger equation, elliptic umbil
```

Rys. 6 Zbiór danych do procesu trenowania

5. Krótki opis działania modelu GPT-2

GPT-2, znany również jako Generative Pretrained Transformer 2, to model transformerów OpenAI, który został wytrenowany na dużym korpusie tekstu. Model ten jest wysoce skuteczny w generowaniu spójnych i sensownych sekwencji tekstowych, a dzięki transformatorowej architekturze jest w stanie uwzględnić długoterminowe zależności między słowami. W naszej pracy używamy modelu z 117 milionami parametrów. GPT-2 otrzymuje na wejściu sekwencję tokenów (przetworzonych słów), następnie przetwarza

każdy z nich przez serię warstw typu Transformer (Dekoder) oraz iteracyjnie generuje najbardziej prawdopodobny następny token.

6. Opis mechanizmu finetuning

Mechanizm finetuning polega na dostosowaniu nauczonego modelu do nowego zadania. W tym przypadku startujemy z pre-trenowanego modelu GPT-2 i dostosowujemy go do generowania tytułów publikacji naukowych. Proces ten zazwyczaj polega na dodaniu nowej warstwy do modelu i trenowaniu tylko tej nowej warstwy na nowym zbiorze danych. W tym celu posłużyliśmy się biblioteką `gpt_2_simple`, która jest wrapperem nad modelem językowym umożliwiającą szybkie pobranie, ładowanie, trenowanie, finetuning oraz generowanie tytułów za pomocą pojedynczych funkcji. Biblioteka implementuje również tokenizer używany przez autorów GPT2, stąd też nie musimy dbać o podawanie do modelu danych w postaci tokenów, lecz zwykłych ciągów znaków.

```
import os

def is_model_present(directory_name, models_path='models'):
    directory_path = os.path.join(models_path, directory_name)

    if not os.path.exists(models_path):
        os.makedirs(models_path)

    return os.path.isdir(directory_path)

import gpt_2_simple as gpt2

model_name = "117M"
if not is_model_present(model_name):
    gpt2.download_gpt2(model_name=model_name)

sess = gpt2.start_tf_sess()
gpt2.finetune(sess,
               'titles_ref.csv',
               model_name=model_name,
               steps=1000,
               save_every=200,
               sample_every=25)

gpt2.generate(sess)
```

Rys. 7 Użycie interfejsu udostępnianego przez bibliotekę `gpt_2_simple`

100% | 1/1 [00:00<00:00, 499.741 t/s]

dataset has 15583 tokens

Training...

```
[1 | 31.25] loss=2.38 avg=2.38
[2 | 60.55] loss=2.27 avg=2.32
[3 | 88.46] loss=2.19 avg=2.28
[4 | 113.80] loss=2.13 avg=2.24
[5 | 140.05] loss=2.11 avg=2.22
[6 | 165.66] loss=1.97 avg=2.17
[7 | 190.63] loss=1.97 avg=2.14
[8 | 216.91] loss=1.91 avg=2.11
[9 | 242.00] loss=1.83 avg=2.08
[10 | 267.65] loss=1.88 avg=2.06
[11 | 293.21] loss=1.81 avg=2.04
[12 | 320.95] loss=1.75 avg=2.01
[13 | 348.75] loss=1.74 avg=1.99
[14 | 374.97] loss=1.71 avg=1.97
[15 | 400.63] loss=1.65 avg=1.94
[16 | 425.85] loss=1.59 avg=1.92
[17 | 450.58] loss=1.52 avg=1.89
[18 | 476.08] loss=1.48 avg=1.87
[19 | 500.94] loss=1.49 avg=1.85
[20 | 525.93] loss=1.41 avg=1.82
[21 | 550.88] loss=1.37 avg=1.80
[22 | 576.23] loss=1.46 avg=1.78
[23 | 601.98] loss=1.40 avg=1.76
[24 | 626.96] loss=1.27 avg=1.74
[25 | 651.88] loss=1.23 avg=1.72
```

===== SAMPLE 1 =====

```
<|startoftext>[x6+|ä|s|3)|]|(1)|classical topology of bionic fields<|endoftext>|>
<|startoftext>|a canonical approach to waveguides and their derivative: the z/s chain''<|endoftext>|>
<|startoftext>|>self-regulating electromagnetic wave energy in the non-shebang spectrum<|endoftext>|>
<|startoftext>|>influence of potential amplitude and periodic integrator on the yang-mills equation in the baxter mod
el and the <|endoftext>|>
<|startoftext>|> problem for noncommutative integral<|endoftext>|>
<|startoftext>|>integral field for weakly interacting quantum systems: some parameters and new ones<|endoftext>|>
<|startoftext>|>information transport and the theory of control in waveguides with complex interactions<|endoftext>|>
<|startoftext>|>on the potential energy density of non-linear schrödinger equations<|endoftext>|>
<|startoftext>|>on the emergence of the multi-interlacing representation of k-da chains (<|endoftext>|>
<|startoftext>|> for the non-linear schrodinger equation<|endoftext>|>
<|startoftext>|>noninertial coupling time for noncommutative integral measures on waveguide products<|endoftext>|>
<|startoftext>|>a comment on a paper on the random walk approximation of feynman's complete<|endoftext>|>
<|startoftext>|>jacobi theory with periodic and noncommutative determinants: new features<|endoftext>|>
<|startoftext>|>quaternionic random walk with random amplitude and time curvature<|endoftext>|>
<|startoftext>|>multi-displace potential and potential density<|endoftext>|>
<|startoftext>|>self-adjointing approach to three dimensional random fields in matrix theory<|endoftext>|>
<|startoftext>|>non-commutative equation of the weakly coupled einstein equation<|endoftext>|>
<|startoftext>|>generalized euclidean path and its applications to nonlinear<|endoftext>|>
<|startoftext>|> hamiltonians<|endoftext>|>
<|startoftext>|>a note on an integral and its application to multi-layered and<|endoftext>|>
<|startoftext>|> coupled<|endoftext>|>
<|startoftext>|>constant-spin coupling without phase-invariant chain<|endoftext>|>
<|startoftext>|>nonlocality of spin fields<|endoftext>|>
<|startoftext>|>nonlinear physical matrix solutions for a class of coupled<|endoftext>|>
<|startoftext>|> stochastic quantum systems. a comparison with the classical approach<|endoftext>|>
<|startoftext>|>on the failure of the joule de la langue?," ``symmetry and the problem of quantum gravity<|endoftext>|>
>
<|startoftext>|> and the ``eigenvalue of the quantum spectrum''<|endoftext>|>
<|startoftext>|>nonlocality of spin fields<|endoftext>|>
<|startoftext>|>metaclassical construction of some weakly compact group on the local spectrum<|endoftext>|>
<|startoftext>|>on top of the noncommutative and periodic lie theories<|endoftext>|>
<|startoftext>|>on the physical properties and potential of random fluctuations<|endoftext>|>
<|startoftext>|>nonlinear system theory and spin transfer systems<|endoftext>|>
<|startoftext>|>on correlation function for the eigenvalue of a random equation with the limit of<|endoftext>|>
<|startoftext>|> q-rotational quantum fisz systems<|endoftext>|>
<|startoftext>|>critical transformations of waveguide correlations<|endoftext>|>
<|startoftext>|>noncommutative deformation of an euclidean matrix under deformation<|endoftext>|>
<|startoftext>|>anomaly elimination for some noncommutative schrodinger equations and the
```

Rys. 8 Przebieg procesu finetuningu

Proces przebiega prawidłowo, co widać po zmniejszającym się lossie.

7. Przykład z wygenerowanym tytułem publikacji naukowej

```
def generate_title(sess=None):
    text = gpt2.generate(sess,
        length=60,
        temperature=0.7,
        nsamples=1,
        batch_size=1,
        return_as_list=True
    )

    title = text[0]
    for i in ["\n", "<", ">", "|", "startoftext", "endoftext"]:
        title = title.replace(i, "")
    title = title.title()

    return title
```

```
generate_title(sess)
```

```
'"Accelerator-Like Deformities In The Eigenvalue Space For The Eigenvalue Of A Covariant Matrix From An Adjacent Sym  
metric Plane"' (2) The Eigenvalue Of The Eigenvalue Of An Elliptic Ring Abelian Position In The Eigenvalue Space'
```

8. Podsumowanie

W powyższym projekcie skupiliśmy się na zastosowaniu technik uczenia maszynowego do generowania tytułów publikacji naukowych z użyciem małego modelu GPT-2. Daje on zrozumiałe (czytelne) wyniki, jednakże wygenerowany tekst zazwyczaj będzie posiadał błędy logiczne lub merytoryczne, tj. tytuły mogą nie odzwierciedlać tytułów, które rzeczywiście mogłyby zostać opublikowane. Często w tytułach pojawiają się również powtórzenia, przez które wygenerowany tekst traci na wiarygodności i sensie.

9. Bibliografia

- "Language models - GPT and GPT-2", Towards Data Science,
(<https://towardsdatascience.com/language-models-gpt-and-gpt-2-8bdb9867c50a>)
- "Fine-tuning the GPT-2 Large Language Model: Unlocking its Full Potential", Medium,
(<https://212digital.medium.com/fine-tuning-the-gpt-2-large-language-model-unlocking-its-full-potential-66e3a082ab9c>)
- "gpt-paper-title-generator", Github, 2019
(<https://github.com/csinva/gpt-paper-title-generator>)