SYNTHESIS AND EVALUATION OF A DOMAIN-SPECIFIC LARGE DATA SET FOR DUNGEONS & DRAGONS

Akila Peiris and Nisansa de Silva

CONTENT

Dungeons and Dragons

FRW Dataset

GPT Models

Text Generation

Entity Classification

Dungeons and Dragons

DUNGEONS AND DRAGONS (D&D, DnD)

- Open-ended
- Pen and paper
- Tabletop
- Role Playing Game (RPG)
- Since 1974
- Predefined rules
- Setting
  - lore, species, artifacts, statistics, and rules

DUNGEON MASTER (DM)

- Conducts gameplay
- Plays Non-Playable Characters (NPC)
- Regulate player actions

Figure 1: A group of people playing D&D. The Dungeon Master (middle) describes a scenario. Taken from the Critical Role web series.

ADVENTURE

- A guide
- Self-contained game narrative
- Quest

- States:
  - Initial state
  - Intermediate states + transition conditions
  - End state (goal)
- NPCs, rewards, objects + Narrative

FORGOTTEN REALMS

- De-facto default setting for D&D 5e
- High fantasy
- Medieval to Middle Eastern and Asian themes
- Has the most resources

- Fandom Wikia
  - https://forgottenrealms.fandom.com/
- 42000 articles

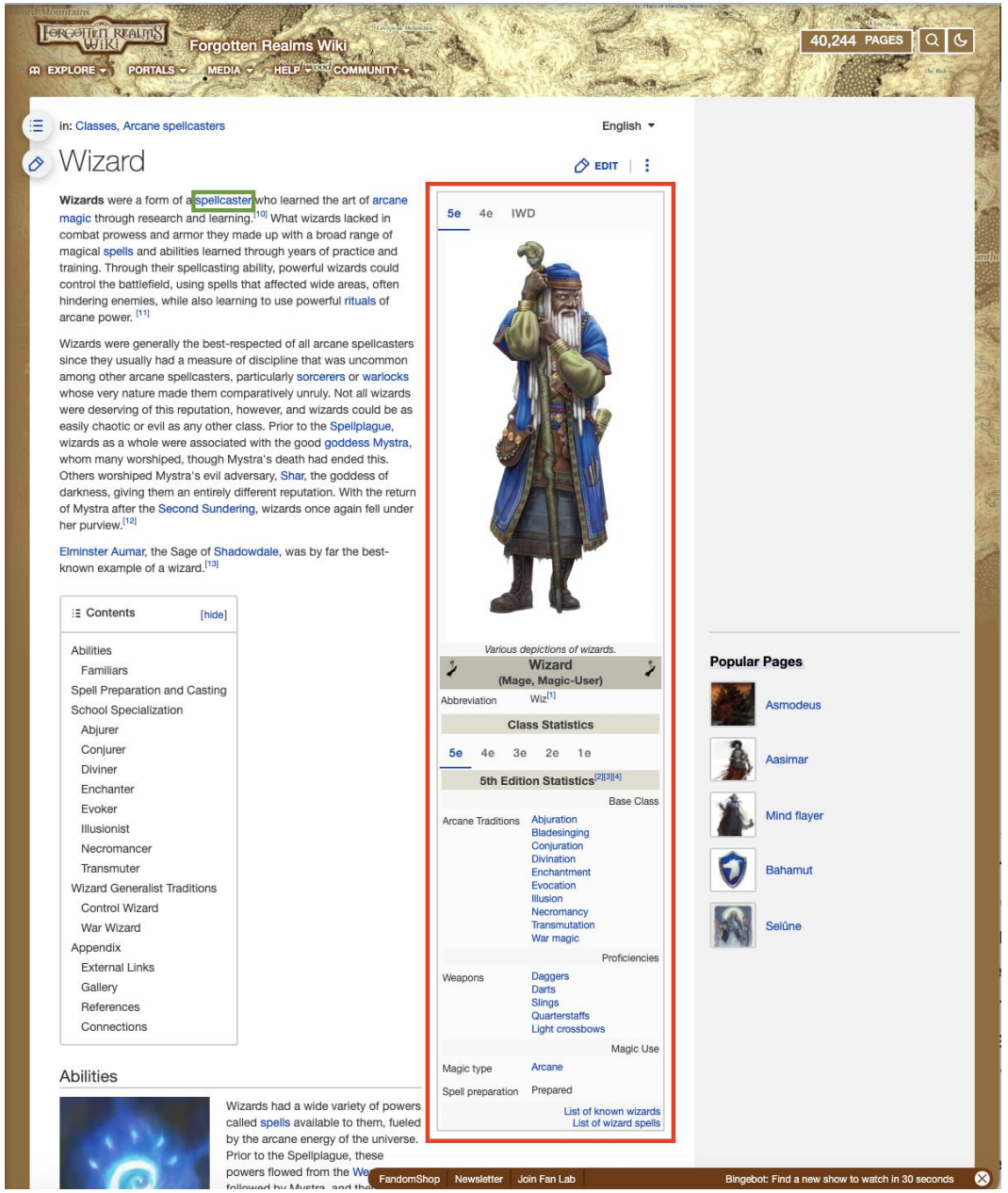
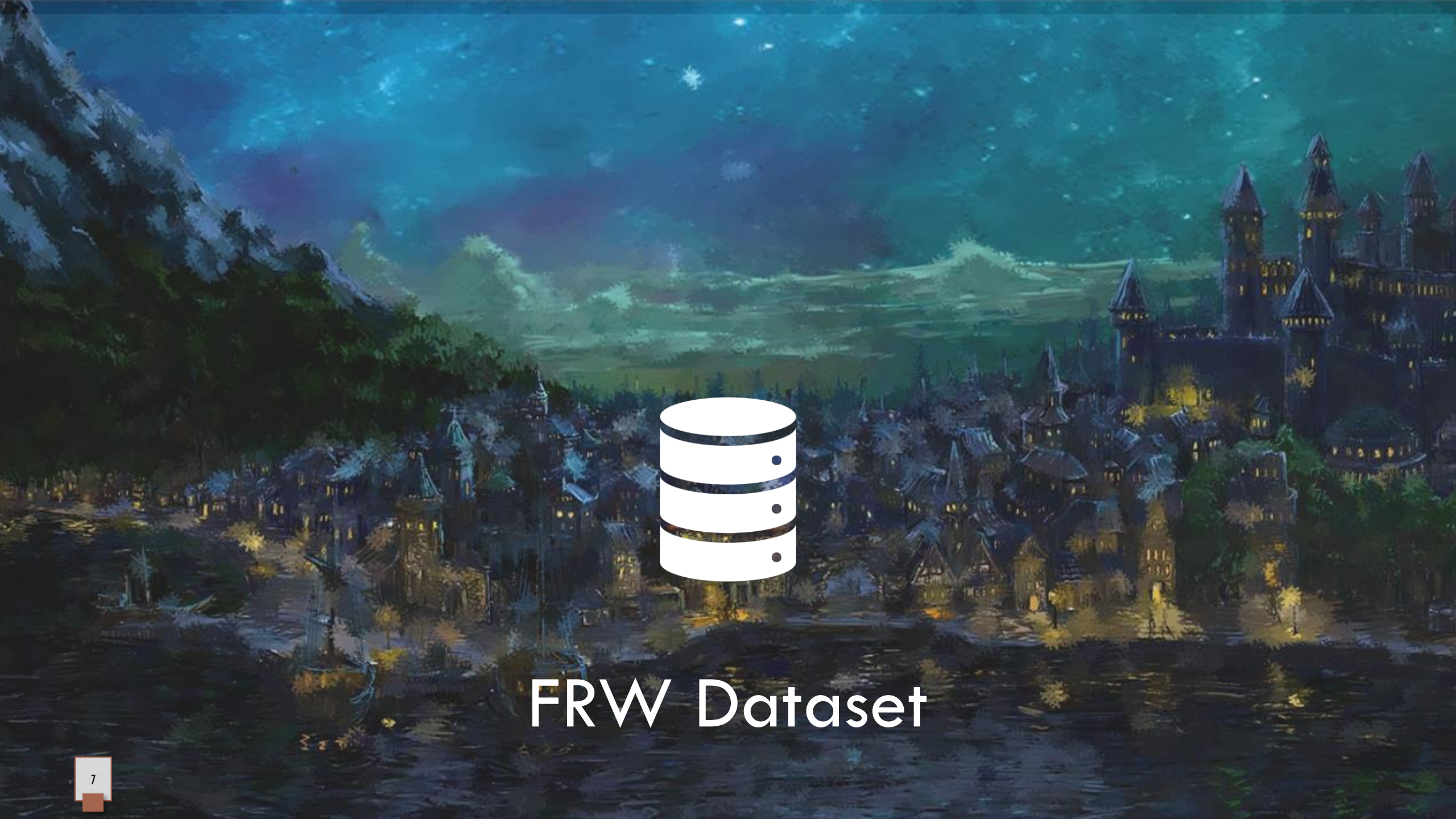


Figure 2: Forgotten Realms Wikia article for *Wizard*.
(Green – first link, Red – Wikipedia Infobox)



FRW Dataset

FORGOTTEN REALMS WIKI (FRW) DATASET

Dataset	Description
FRW-P	Raw plain text
FRW-J	A JSON structure with plain text indexed by article title
FRW-FJ	A JSON structure with only the first paragraph of articles indexed by article title
FRW-L	A directional graph indicating all the references in the articles to other articles
FRW-FL	A directional graph indicating the first references in the articles to other articles
FRW-CL	A directional graph indicating the category references in the articles to category articles
FRW-I	A JSON structure for the Wikipedia infobox substructures indexed by article title
FRW-PE	Poincaré embedding for the first links
FRW-W	2 Word2Vec models for full text (CBOW and Skip-gram)
FRW-D	2 Doc2Vec models for full text (PV-DBOW and PV-DM)
FRW-FD	2 Doc2Vec models for first paragraph text (PV-DBOW and PV-DM)

DATASET : FRW-P

- No markdown
- No templates

Statistic	Value
Total number of words (excluding article titles)	9,189,536
Total number of words (including article titles)	9,287,670
Total number of unique words	145,624
Total number of sentences	517,248

DATASET : FRW-J

- JSON structure
- Raw corpus but indexed

Statistic	Value
Total number of articles	48,892
Average number of words per sentence	17.77
Average number of words per article	187.96
Average number of sentences per article	10.58

DATASET : FRW-FJ

- First paragraph only
- *Wikipedia lead section**
- Self contained summary
- 10% compression

Statistic	Value
Total number of articles	41,204
Total number of words	980,047
Total number of sentences	98,244
Average number of words per sentence	9.98
Average number of words per article	23.78
Average number of sentences per article	2.38

* https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

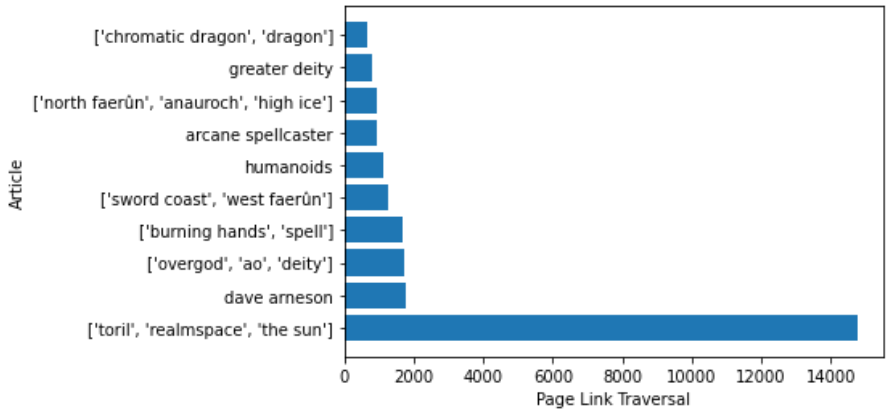
DATASET : FRW-L

- Directed graph
- Edges – References from 1 article to another

Statistic	Value
Total number of nodes	46,910
Total number of edges	570,857
Average number of edges per node	12.16

DATASET : FRW-FL

- First link of the article
- Places the article in context*



Statistic	Value
Total number of nodes	43,329
Total number of edges	41,213
Number of nodes not referenced by others	34,881
Number of nodes with no references	2151

* https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

DATASET : FRW-CL

- Special references in a page
- Virtual pages
- Flat hierarchy
- No ordering

Statistic	Value
Total number of nodes	52,022
Total number of edges	290,292
Average number of edges per node	5.58

DATASET : FRW-I

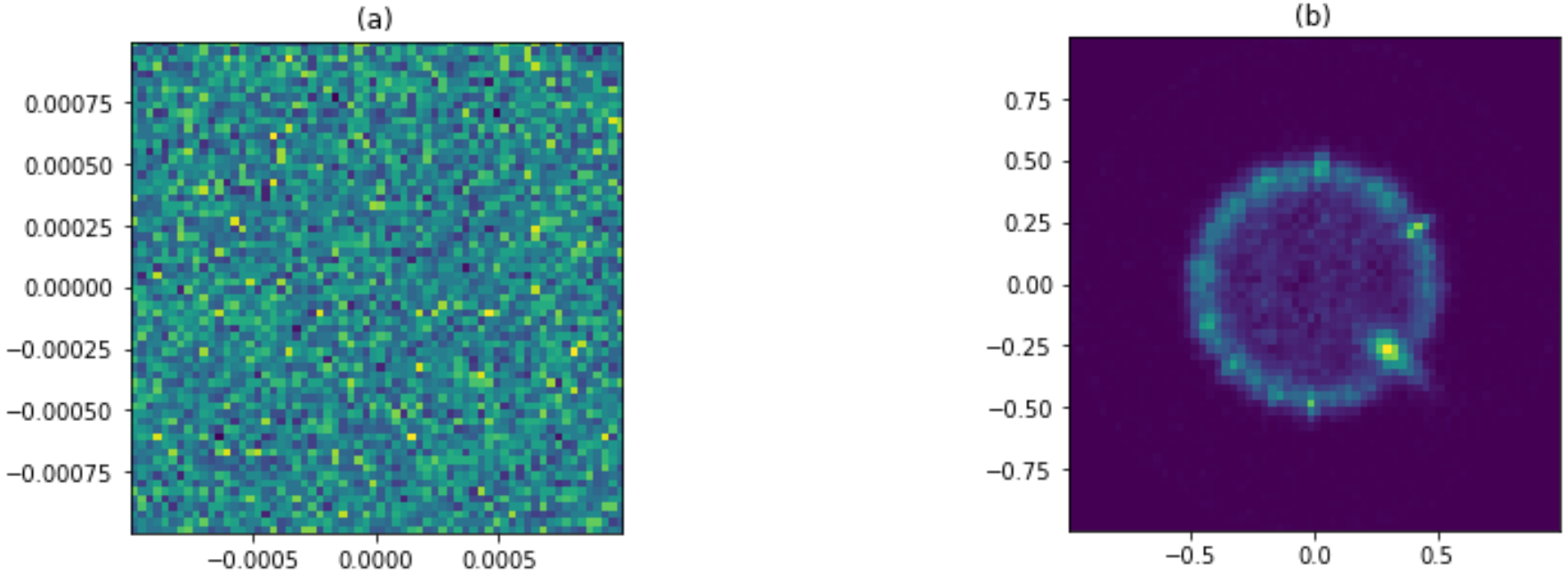
- Key-value format
- JSON structure
- 73% of pages contain infoboxes
  - 1:3 in Wikipedia[1]

Statistic	Value
Average number of attributes per infobox	40.54
Average number of completed (filled) attributes per infobox	10.40
Total number of articles containing infoboxes	35,923

[1] Lange, D., Bo¨hm, C., and Naumann, F. (2010). In *Extracting structured information from wikipedia articles to populate infoboxes.*

DATASET : FRW-PE

- Poincaré[2] embeddings for first links
- Convergence: initial state and after 50 epochs

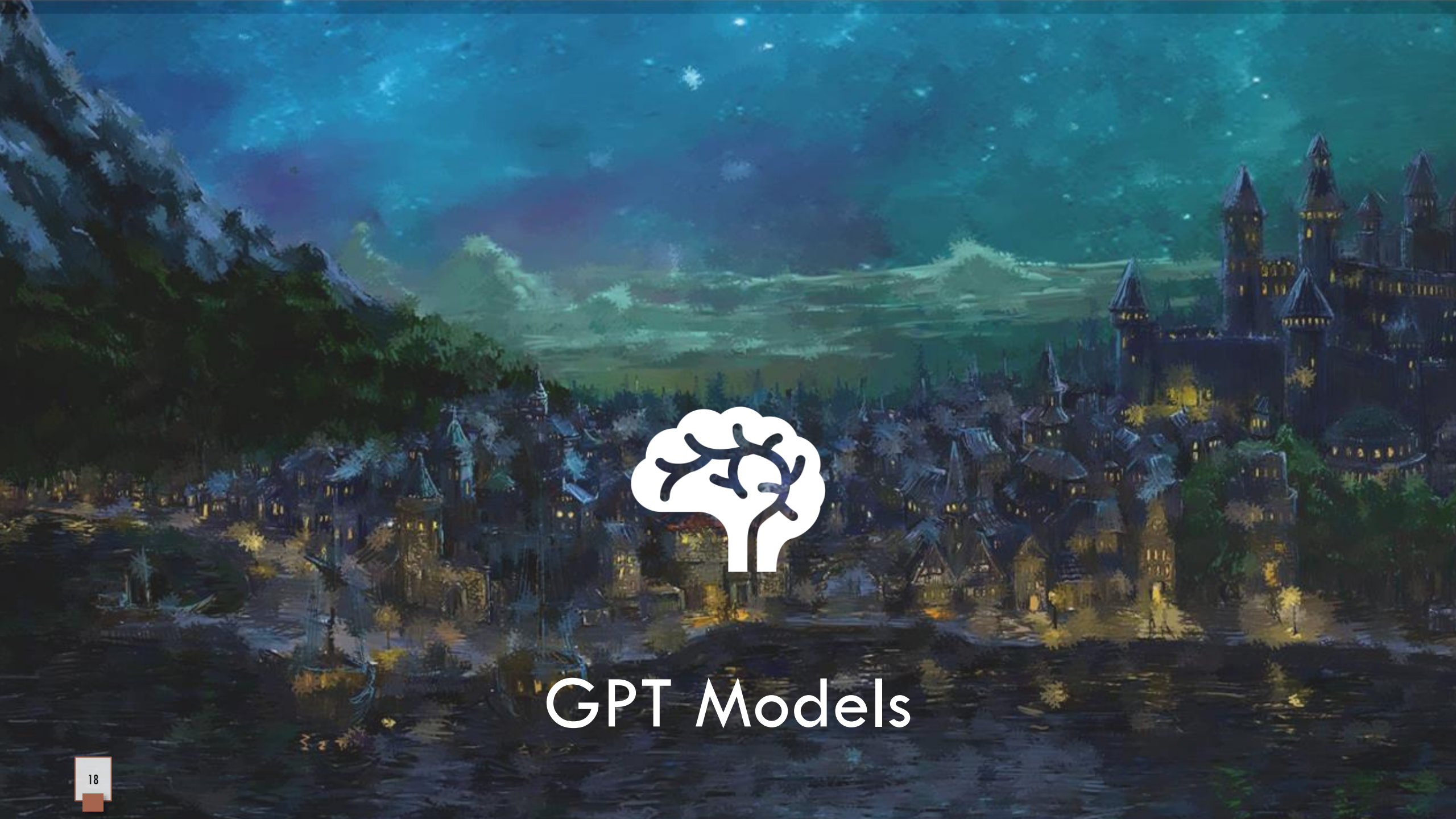


[2] Nickel, M. and Kiela, D. (2017). Poincare´ embed- dings for learning hierarchical representations. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 6341–6350. Curran Associates, Inc.

DATASETS : WORD AND DOCUMENT EMBEDDING

Embedding	Dataset	Uses data from	Techniques
Word embedding (Word2Vec)[3]	FRW-W	FRW-P	CBOW, Skip-gram
Document embedding (Doc2Vec)[4]	FRW-D	FRW-J	PV-DBOW, PV-DM
First paragraph document embedding (Doc2Vec)	FRW-FD	FRW-FJ	PV-DBOW, PV-DM

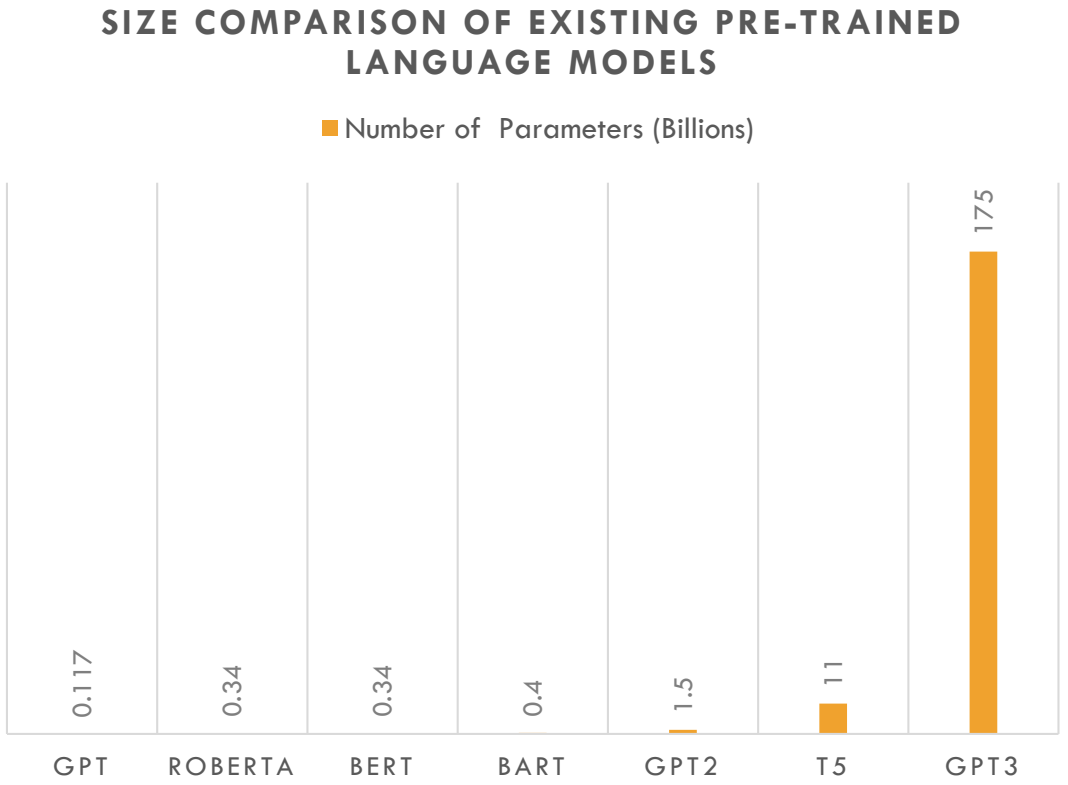
[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S.,and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems, pages 3111–3119
[4] Le, Q. and Mikolov, T. (2014). Distributed representa- tions of sentences and documents. In International conference on machine learning, pages 1188–1196. PMLR.



GPT Models

GENERATIVE PRE-TRAINED TRANSFORMERS (GPT)

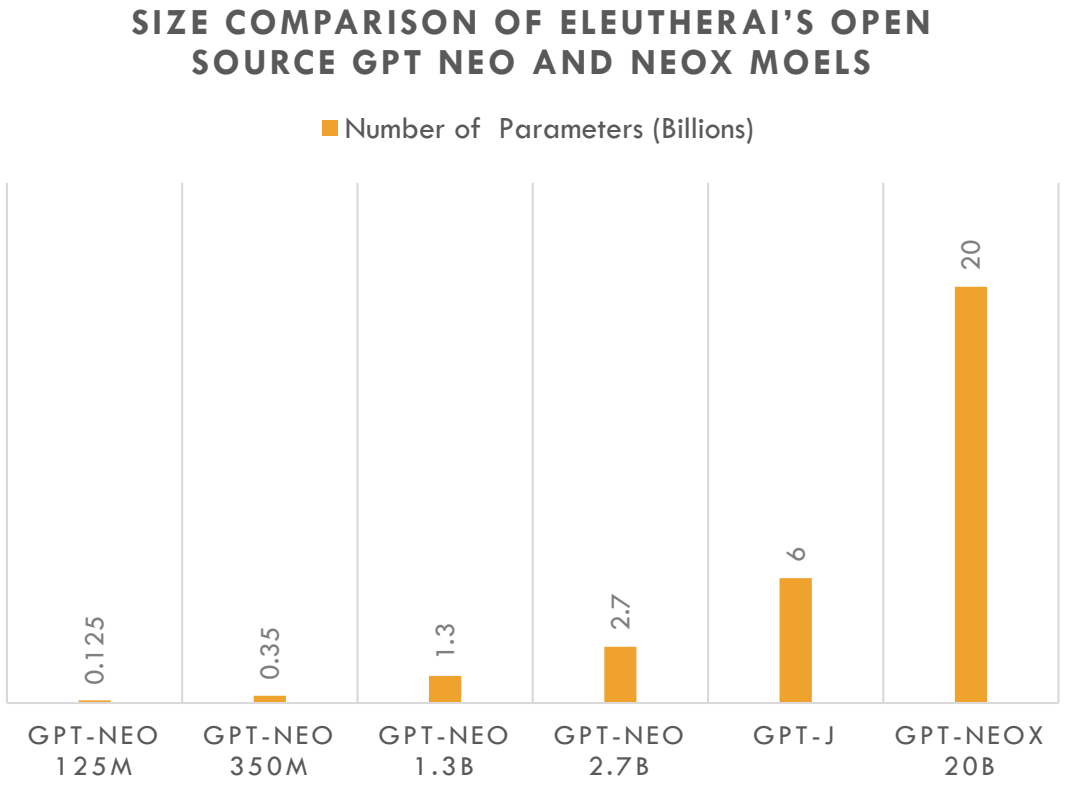
- GPT [1]
  - deep learning
  - autoregressive
  - language model
  - GPT3 [2] – 175 Billion parameters



[1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018
[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020

GENERATIVE PRE-TRAINED TRANSFORMERS (GPT)

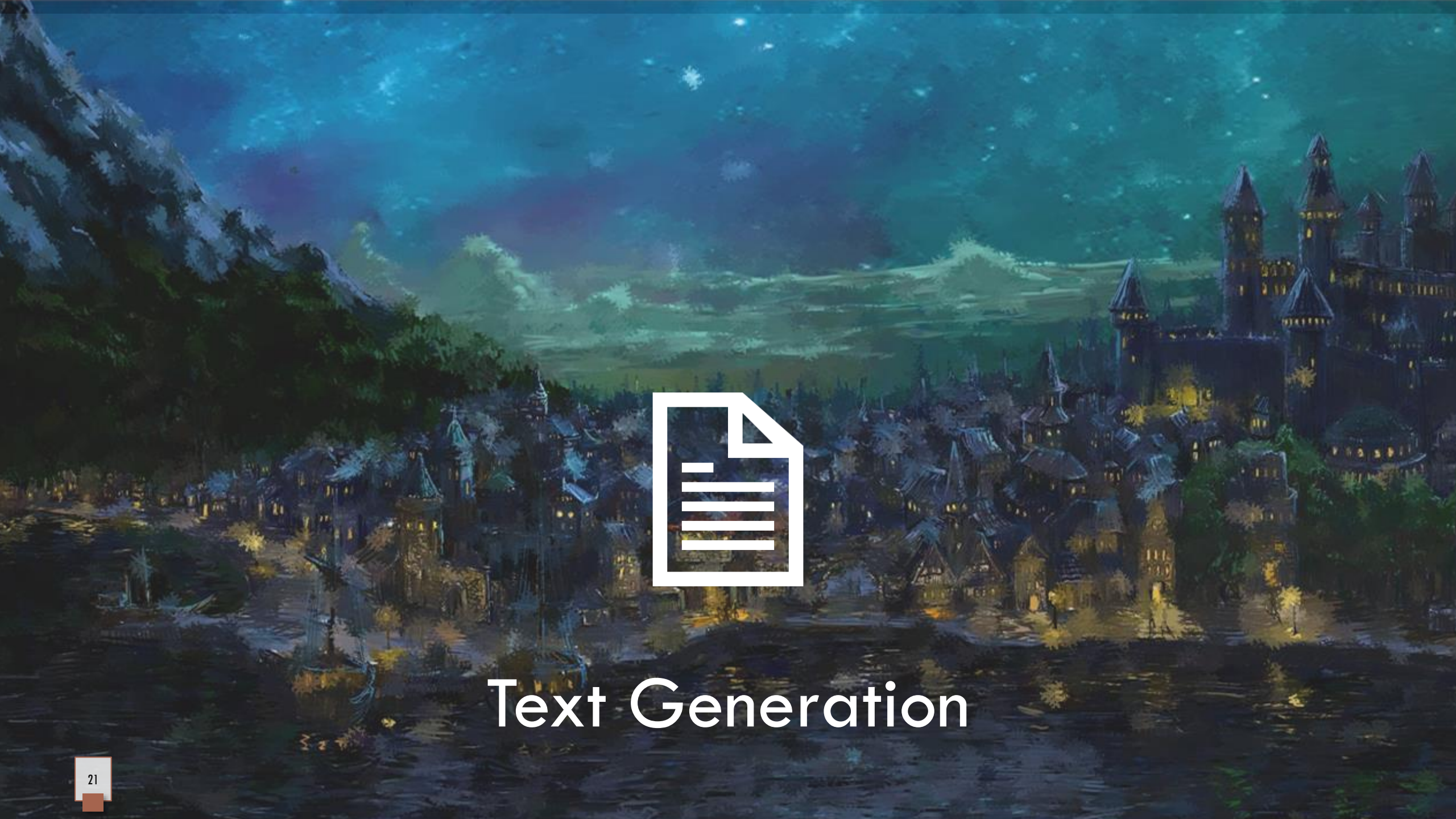
- GPT-Neo [3] (and Neox) [4]
  - Opensource alternative
  - EleutherAI
  - Neo – 2.7 Billion parameters
  - Neox – 20 Billion parameters [5]



[3] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow," Mar. 2021, If you use this software, please cite it using these metadata. [Online]. Available: https://doi.org/10.5281/zenodo.5297715

[4] A. Andonian, Q. Anthony, S. Biderman, S. Black, P. Gali, L. Gao, E. Hallahan, J. Levy-Kramer, C. Leahy, L. Nestler, K. Parker, M. Pieler, S. Purohit, T. Songz, P. Wang, and S. Weinbach, "GPT-NeoX: Large scale autoregressive language modeling in pytorch," 2021. [Online]. Available:http://github.com/eleutherai/gpt-neox

[5] A. Andonian, Q. Anthony, S. Biderman, S. Black, P. Gali, L. Gao, E. Hallahan, J. Levy-Kramer, C. Leahy, L. Nestler, K. Parker, M. Pieler, S. Purohit, T. Songz, P. Wang, and S. Weinbach, "GPT-NeoX: Large scale autoregressive language modeling in pytorch," 2021. [Online]. Available: http://github.com/eleutherai/gpt-neox



Text Generation

TRAINING THE MODEL

- GPT Neo 1.3B (GPT3_XL)
- FRW-J dataset

D&D TEXT GENERATION

Input prompt

"A Bard, A Wizard, and a Paladin walk into a tavern. The bard hides behind a group of tables, drinking and hiding his face. The Wizard hands over the Handbook of All Things Magical to the paladin, who hangs it up and begins to read. The bard steals the book, dissolves the rules and instructions, and replaces them with his own. The party meets up at the next table and learn that the wizard took the book because it was to be penned by him. Bards make a sign of the wizard to politely take the book and tell him the party is there to do his bidding. The wizard read the book, the bard says his or her name and uses it to vanish out into the night. As the wizard takes off, the bard doorstep knights and spells a silent alarm before whispering, "He is a wizard, and his old book was the Handbook of All Things Magical. I will guard it for him. […]"

"A Bard, A Wizard, and a Paladin walk into a tavern. The Bard says that he lives in Baldur's Gate in a tavern dedicated to a wizard named Juiblex. The Bard tells Xanatos about his origin story. In the tavern, the Bard mentions non-existence of its streets. The Bard explains that the people of Baldur's Gate rarely go outside of the city, and are plagued by strange things. The Bard meets a human man. The Bard asks the man why he is "always" being summoned by Juiblex, whereas he is a cast-away person who wandered the Realms. The Bard mentions a dream he had in which he was being magically summoned to the tavern to meet the wizard, and the man says he has ably catalogued the madness of the city. The Bard asks why people don't leave, and the man counters that only the rich can afford to leave. The Bard asks the man if he agrees with the praise of his dream, and he agrees by telling him that 50% of the people in his fantasy life are less important than the other 50% . […]"

D&D TEXT GENERATION

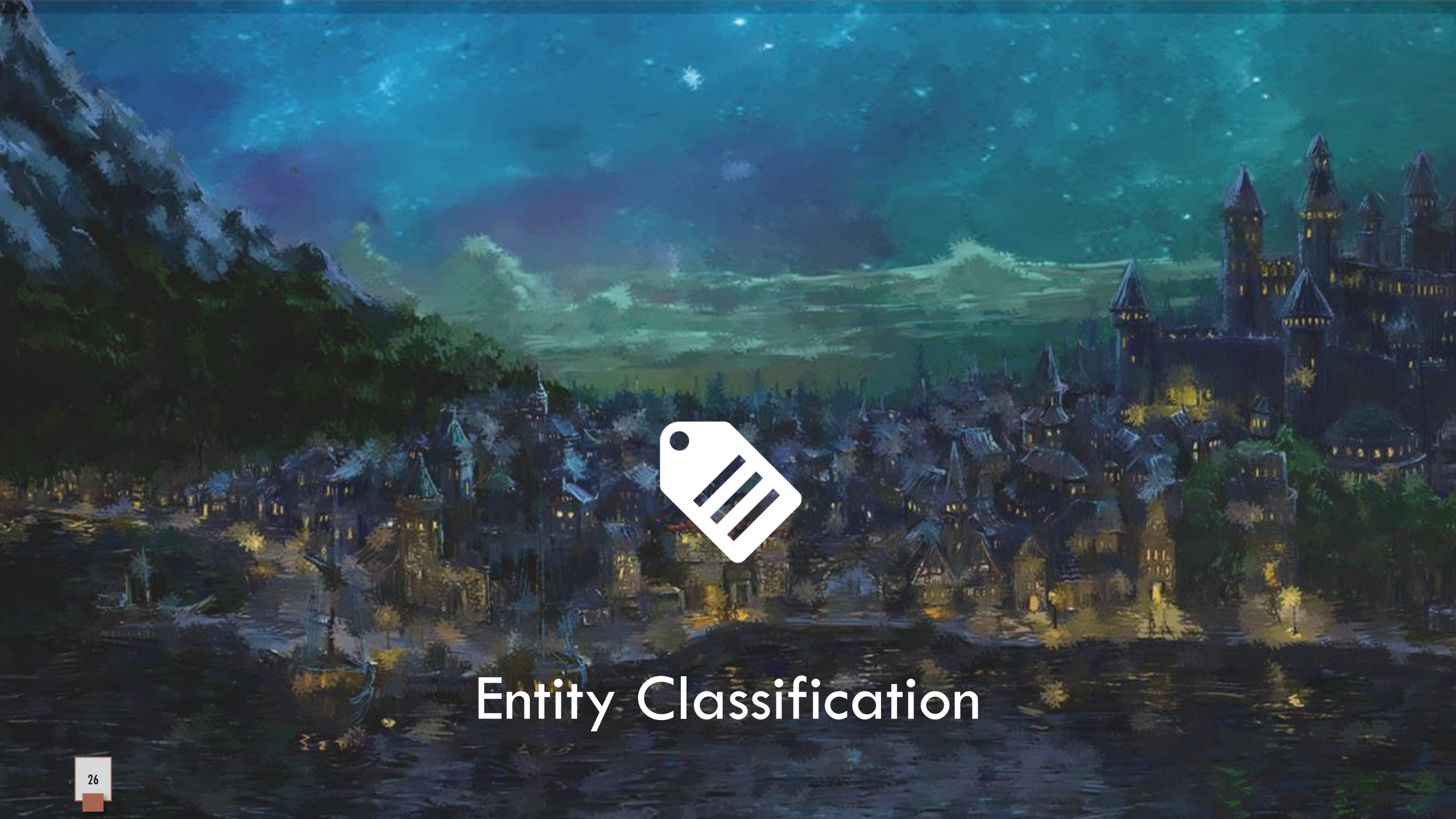
“A Bard, A Wizard, and a Paladin walk into a tavern. The bard hides behind a group of tables, drinking and hiding his face. The Wizard hands over the Handbook of All Things Magical to the paladin, who hangs it up and begins to read. The bard steals the book, dissolves the rules and instructions, and replaces them with his own. The party meets up at the next table and learn that the wizard took the book because it was to be penned by him. Bards make a sign of the wizard to politely take the book and tell him the party is there to do his bidding. The wizard read the book, the bard says his or her name and uses it to vanish out into the night. As the wizard takes off, the bard doorstep knights and spells a silent alarm before whispering, "He is a wizard, and his old book was the Handbook of All Things Magical. I will guard it for him. […]”

“A Bard, A Wizard, and a Paladin walk into a tavern. The Bard says that he lives in Baldur's Gate in a tavern dedicated to a wizard named Juiblex. The Bard tells Xanatos about his origin story. In the tavern, the Bard mentions non-existence of its streets. The Bard explains that the people of Baldur's Gate rarely go outside of the city, and are plagued by strange things. The Bard meets a human man. The Bard asks the man why he is "always" being summoned by Juiblex, whereas he is a cast-away person who wandered the Realms. The Bard mentions a dream he had in which he was being magically summoned to the tavern to meet the wizard, and the man says he has ably catalogued the madness of the city. The Bard asks why people don't leave, and the man counters that only the rich can afford to leave. The Bard asks the man if he agrees with the praise of his dream, and he agrees by telling him that 50% of the people in his fantasy life are less important than the other 50% . […]”

D&D TEXT GENERATION

“A Bard, A Wizard, and a Paladin walk into a tavern. The bard hides behind a group of tables, drinking and hiding his face. The Wizard hands over the Handbook of All Things Magical to the paladin, who hangs it up and begins to read. The bard steals the book, dissolves the rules and instructions, and replaces them with his own. The party meets up at the next table and learn that the wizard took the book because it was to be penned by him. Bards make a sign of the wizard to politely take the book and tell him the party is there to do his bidding. The wizard read the book, the bard says his or her name and uses it to vanish out into the night. As the wizard takes off, the bard doorstep knights and spells a silent alarm before whispering, "He is a wizard, and his old book was the Handbook of All Things Magical. I will guard it for him. […]”

“A Bard, A Wizard, and a Paladin walk into a tavern. The Bard says that he lives in Baldur's Gate in a tavern dedicated to a wizard named Juiblex. The Bard tells Xanatos about his origin story. In the tavern, the Bard mentions non-existence of its streets. The Bard explains that the people of Baldur's Gate rarely go outside of the city, and are plagued by strange things. The Bard meets a human man. The Bard asks the man why he is "always" being summoned by Juiblex, whereas he is a cast-away person who wandered the Realms. The Bard mentions a dream he had in which he was being magically summoned to the tavern to meet the wizard, and the man says he has ably catalogued the madness of the city. The Bard asks why people don't leave, and the man counters that only the rich can afford to leave. The Bard asks the man if he agrees with the praise of his dream, and he agrees by telling him that 50% of the people in his fantasy life are less important than the other 50% . […]”



Entity Classification

ENTITY CLASSIFICATION

- Entities in Generated Text
  - In Forgotten Realms lore
- FRW-I dataset
- Entity, Label, Generated text

CONTENT

References

REFERENCES

[1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018
[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., “Language models are few-shot learners,” Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020
[3] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, “GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow,” Mar. 2021, If you use this software, please cite it using these metadata. [Online]. Available: https://doi.org/10.5281/zenodo.5297715
[4] A. Andonian, Q. Anthony, S. Biderman, S. Black, P. Gali, L. Gao, E. Hallahan, J. Levy-Kramer, C. Leahy, L. Nestler, K. Parker, M. Pieler, S. Purohit, T. Songz, P. Wang, and S. Weinbach, “GPT-NeoX: Large scale autoregressive language modeling in pytorch,” 2021. [Online]. Available:http://github.com/eleutherai/gpt-neox
[5] A. Andonian, Q. Anthony, S. Biderman, S. Black, P. Gali, L. Gao, E. Hallahan, J. Levy-Kramer, C. Leahy, L. Nestler, K. Parker, M. Pieler, S. Purohit, T. Songz, P. Wang, and S. Weinbach, “GPT-NeoX: Large scale autoregressive language modeling in pytorch,” 2021. [Online]. Available: http://github.com/eleutherai/gpt-neox

THANK YOU!

ROLL INVESTIGATION TO ASK A QUESTION

LET THE DICE DECIDE!