



A DUNGEONS AND DRAGONS DATA SET

Akila Peiris

CONTENT



Introduction



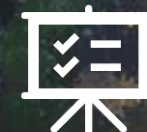
Data
Extraction



Dataset



Evaluation



Discussion

INTRODUCTION

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



INTRODUCTION

FORGOTTEN REALMS

- De-facto default setting for D&D 5e
- High fantasy
- Medieval to Middle Eastern and Asian
- Has the most resources
- Fandom Wikia
- <https://forgottenrealms.fandom.com/>
- 41000 articles



INTRODUCTION

WIKIPEDIA AS A DATA SOURCE

- Corpi of the **relevant** domain
- Community-based collaborative text maintenance.
- Freely available
- Multiple languages
- Sheer volume of content
- Annotated data (infoboxes)

<i>Dungeons & Dragons</i>	
	
Logo used for the 5th edition	
Designers	Gary Gygax Dave Arneson
Publishers	TSR (1974–1997), Wizards of the Coast (1997–present)
Publication	1974 (original) 1977 (Basic Set 1st ver.) ^[1] 1977 (Advanced D&D) 1981 (Basic Set 2nd ver.) 1983 (Basic Set 3rd ver.) 1989 (AD&D 2nd Edition) 1991 (Rules Cyclopedia) 2000 (3rd edition) 2003 (v3.5) 2008 (4th edition) 2014 (5th edition)
Years active	1974–present
Genres	Fantasy
Systems	<i>Dungeons & Dragons</i> d20 System (3rd Edition)
Playing time	Varies
Random chance	Dice rolling
Skills required	Role-playing , improvisation , tactics , arithmetic
Website	dnd.wizards.com  

DATA EXTRACTION



Dump

Forgotten Realms Wiki
Markdown text + XML
schema



Clean

<text>, <title> tags
Templates



Text

Remove templates
Parse markdown



Graph

Links
Markdown



Infobox

Key-value pairs

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

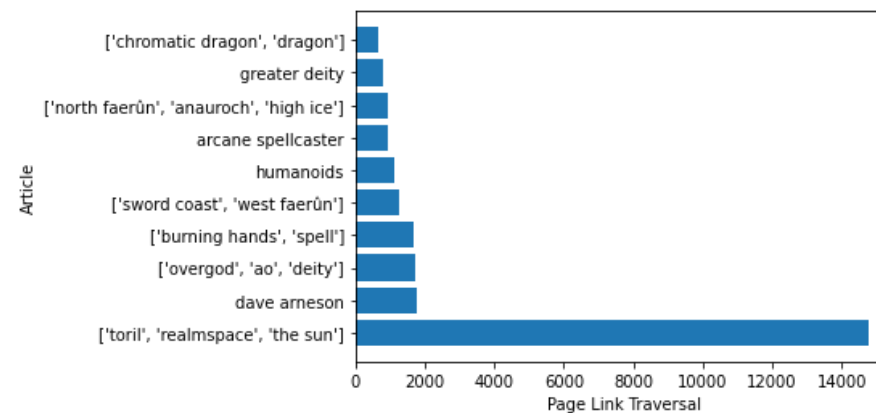
DATASET : FRW-L

- 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

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

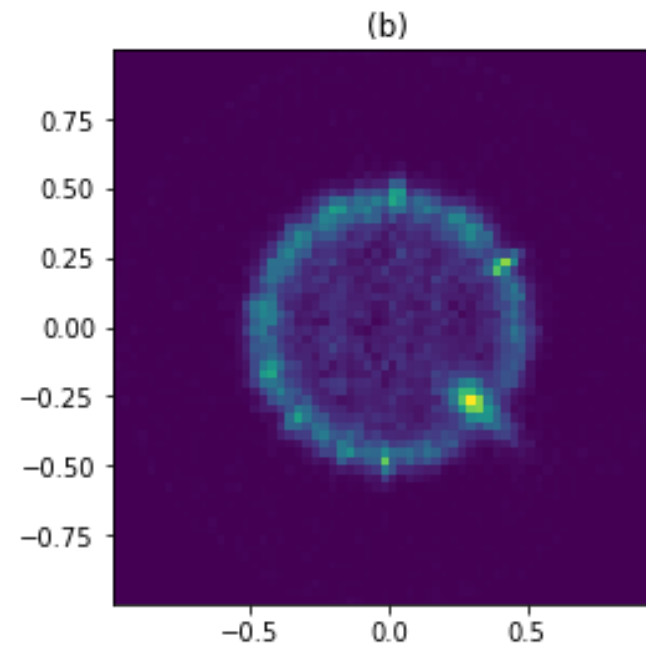
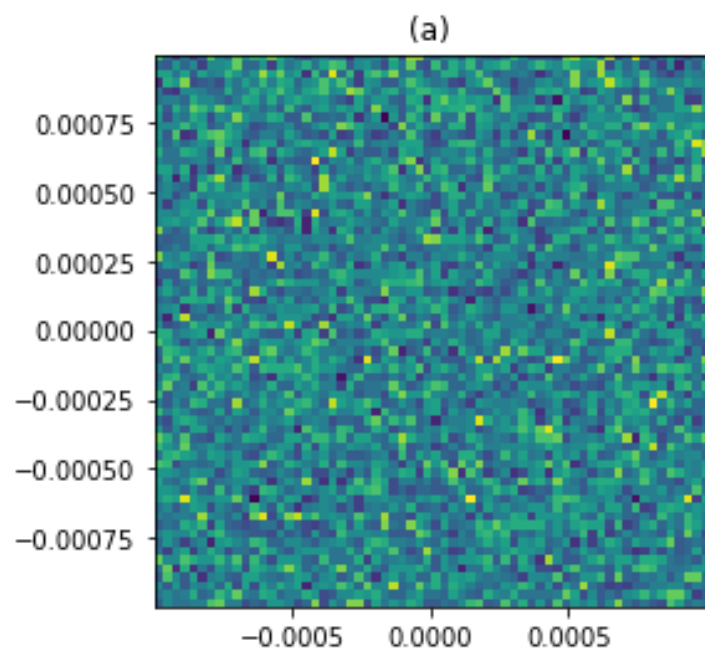
DATASET : FRW-I

- 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

DATASET : FRW-PE

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



DATASETS : WORD AND DOCUMENT EMBEDDING

Embedding	Dataset	Techniques
Word embedding (Word2Vec) ^[3]	FRW-P	CBOW, Skip-gram
Document embedding (Doc2Vec) ^[4]	FRW-J	PV-DBOW, PV-DM
First paragraph document embedding (Doc2Vec)	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. In Advances in neural information processing systems, pages 3111–3119

[4] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning, pages 1188–1196. PMLR.

EVALUATION : PAIRWISE COMPARISON

GENERATING TEXT PAIRS

- Text pair – Article titles
 - Hierarchical similarity
 - Embedded vector similarity
- 1 million pairs
- Weighted random sampling
- With replacement
- Weight updated dynamically

$$P(i) = \frac{w_i}{\sum_{n=1}^N w_j} \quad (1)$$

$$w_{ik} = \sqrt{N} - k \quad (2)$$

N - the total number of pairs we generate

w_i - the current weight of the element i

k - the number of times element i has been selected

EVALUATION : PAIRWISE COMPARISON

HIERARCHICAL SIMILARITY

- FRW-FL
- Add root node
- Wu and Palmer
- Jiang-Conrath
- Poincare embeddings

Unified graph with node $G' = (V', E')$
where,

$$E' = \begin{cases} E \cup (root, v), & \text{if } deg^-(v) = 0 \\ & \text{and } v \neq root \\ E \cup (root, v') \cup (v', v), & \text{if } v \in G_c \\ E, & \text{otherwise} \end{cases} \quad (3)$$

Let G be a graph in the set of disconnected graphs

$$G = (V, E) \in G_1, G_2, \dots, G_n$$

$$E \in E'$$

$$V \in V'$$

$G_c = (V_c, E_c) \in G$ where $e_1, \dots, e_n$ is a trail with vertex
sequence $a_1, \dots, a_n$

EVALUATION : PAIRWISE COMPARISON

EMBEDDED VECTOR SIMILARITY

- FRW-W
- FRW-D
- FRW-FD
- Cosine similarity
- Word2Vec
 - Retrieve content for article title
 - Vectors for each word
 - Average pool
- Doc2Vec
 - Get document vector for article title

EVALUATION : PAIRWISE COMPARISON

			Hierarchical			Embedding						
			WP	JC	P	Word2Vec		Doc2Vec				
						<i>(FRW-J)</i>		<i>(FRW-FJ)</i>		<i>(FRW-J)</i>		
						CBOW	SG	DM	DBOW	DM	DBOW	
Hierarchical		WP	1.0000									
		JC	0.6346	1.0000								
		P	0.0097	0.0624	1.0000							
Embedding	Word2Vec	<i>(FRW-J)</i>	CBOW	0.0581	0.0212	0.0013	1.0000					
			SG	0.0553	0.0188	-0.0043	0.9412	1.0000				
	Doc2Vec	<i>(FRW-FJ)</i>	DM	0.0040	-0.0298	0.0548	-0.0626	-0.0791	1.0000			
			DBOW	0.0466	0.0155	0.0359	0.0362	0.0222	0.5691	1.0000		
		<i>(FRW-J)</i>	DM	0.0259	0.0186	0.0175	-0.1865	-0.2593	0.1724	0.1484	1.0000	
			DBOW	0.0361	0.0287	0.0453	-0.0896	-0.1601	0.1511	0.1825	0.5493	1.0000

WP – Wu & Palmer similarity

JC – Jiang-Conrath Similarity

P – Poincare' metric

CBOW – Continuous Bag of Words

SG – Skip-gram

DM – Distributed Memory

DBOW – Distributed Bag of Words

DISCUSSION

- Similar measures have relatively high coefficients
- First paragraph similarity is comparable to full document similarity
- First link alternatives:
 - Multiple first links (consecutive)
- Novel datasets for D&D domain

THANK YOU!

ROLL 16 OR HIGHER FOR
A CHANCE TO ASK A
QUESTION



LET THE DICE DECIDE!