

EVOLVING WISDOM

Searching for a Personal Sage through NLP

FSDL 2021 Online, Final Project
https://github.com/wtcooper/fsdl_pplm

CAN I EVOLVE TEXT GENERATION TOWARDS BETTER CONTENT?

Vision: a user is journaling about a traumatic event, and the journaling app, *knowing the user's penchant for Eastern philosophy*, recommends a helpful and relevant snippet to aid in their internal processing

Approach Overview:

- Utilize GPT-2 with a spectrum of ancient wisdom texts as auxiliary input to conditionally influence text generation([Uber PPLM](#))
- *World Scripture: A Comparative Anthology of Sacred Texts*
- Test the ability to both generate new text and recommend existing text given a users preferred style and content
- Deploy a simple web app to provide user interaction

METHODS - DETAILS

- https://github.com/wtcooper/fsdl_pplm
 - Readme
 - Appendix of this presentation on Github

RESULTS

×

Configure Style and Content

Choose Style

Buddhism

Custom Keywords (comma separated)

Choose Topic

Divine Law Truth and Cosmic Pri...

Generate Options

Number of samples:

1

Length of output text:

25

Strength of conditioning:

0.04

Number of iterations:

1

Recommender Options

Number of topics to search:

Input:

What do you want hear more about?

In the beginning, the universe

31/1000

Generate new text

Recommend similar passages

Output:

['In the beginning, the universe \xa0was not created. \xa0There was nothing to create. \xa0Nothing to form into a world of existence. ']

Keywords (first 10): ['burning', 'suchness', 'law', 'conquer', 'fool', 'suffering', 'sees

In the beginning, the universe
In the beginning, the universe was
In the beginning, the universe was not
In the beginning, the universe was not created
In the beginning, the universe was not created.
In the beginning, the universe was not created.
In the beginning, the universe was not created.
In the beginning, the universe was not created. There
In the beginning, the universe was not created. There was
In the beginning, the universe was not created. There was nothing
In the beginning, the universe was not created. There was nothing to

NEXT STEPS

1. Fine-tune the base language model on a large corpus of sacred texts
2. Explore alternative approaches to PPLM

APPENDIX

METHODS - OVERVIEW

1. Extract data from a collection of world sacred texts

Buddhism, Hinduism, Christianity, Judaism, Confucianism, Sikhism, Taoism, Jainism, African Traditional Religions

2. For each major tradition, extract top 100 keywords
3. Develop a classification model to predict the text topic
4. **Text Generation:** Using the keywords (style) and classification model (content), apply Uber's PPLM to generate text in a given style and content
5. **Text Recommendation:** Find relevant passages based on user content in their preferred tradition

METHODS - DATA EXTRACTION

Data Source:

- World Scripture: A Comparative Anthology of Sacred Texts
- <http://www.tparents.org/Library/Unification/Books/World-S/0-Toc.htm>

Regex on
html tags

Diagram illustrating the extraction of HTML tags from a text source using regex. The text source is shown as a code block with line numbers 27 to 39. The extraction process is indicated by red arrows pointing from labels to specific HTML tags:

- topic** points to `<p>Traces Of God's Existence </p>` (line 27).
- start** points to `<p>` (line 28).
- text** points to the content of the paragraph: `How can human beings recognize the existence of this transcendent R` (line 28).
- source** points to `<p>` (line 29).
- end** points to `</p>` (line 29).

The code block contains the following HTML content:

```
27 <p>Traces Of God's Existence </p>
28 </b></font><p>How can human beings recognize the existence of this transcendent R
29 <p><hr></p>
30 <p>The door of the Truth is covered by a golden disc. Open it, O Nourisher! Remov
31 <p>O Nourisher, lone Traveller of the sky! Controller! O Sun, offspring of Prajap
32 <p>Hinduism. Isha Upanishad 15-16 </p>
33 <p><hr></p>
34 <p>He who looks inwardly at the self revels in the self; He who revels in the sel
35 <p>Jainism. Acarangasutra 2.173 </p>
36 <p><hr></p>
37 <p>The thing that is called Tao is eluding and vague. Vague and eluding, there is
38 <p>Taoism. Tao Te Ching 21 </p>
39 <p><hr></p>
40 <p>Confucius said, "The power of spiritual forces in the universe--how active it
```


METHODS - KEYWORD EXTRACTION

Apply TF-IDF to gather top 100 Keywords per Tradition

- based on: <https://github.com/kavgan/nlp-in-practice/tree/master/tf-idf>

```
=====
tradition: Hinduism
keywords: [('undivided', 0.95), ('single', 0.859), ('surrender', 0.826), ('phenomenal', 0.801), ('faith', 0.507), ('enticement', 0.772), ('lo
=====
tradition: Jainism
keywords: [('deeds', 0.764), ('right', 0.762), ('knows', 0.72), ('without', 0.698), ('greatness', 0.684), ('neither', 0.674), ('capital', 0.6
=====
tradition: Taoism
keywords: [('faith', 0.76), ('greater', 0.727), ('looks', 0.705), ('everlasting', 0.669), ('wins', 0.659), ('injury', 0.642), ('kill', 0.378)
=====
tradition: Confucianism
keywords: [('seeks', 0.416), ('perseverance', 0.737), ('attack', 0.728), ('sacrifice', 0.693), ('humane', 0.681), ('destiny', 0.67), ('discov
=====
tradition: Buddhism
keywords: [('burning', 0.898), ('suchness', 0.66), ('law', 0.87), ('conquer', 0.659), ('fool', 0.79), ('suffering', 0.857), ('sees', 0.851),
=====
```

METHODS - DISCRIMINATOR MODEL

1. Use Chapter headings as Topics (20 classes)
2. Train as classification head on GPT-2
3. 90% train / 10% test
4. 50 epochs, LR=0.001

Some tuning but low accuracy

Test performance per epoch			
epoch	loss	acc	
1	2.8660716777894555	0.13719512195121952	
2	2.6257752965136274	0.2225609756097561	
3	2.519620360397711	0.25914634146341464	
4	2.4150783783052026	0.3170731707317073	
5	2.327070058845892	0.31402439024390244	
6	2.2657220014711705	0.35365853658536583	
7	2.2868954611987604	0.3323170731707317	
8	2.2773656350810354	0.32926829268292684	
9	2.3066995841700857	0.31402439024390244	
10	2.20122276282892	0.3353658536585366	
11	2.239207543977877	0.3353658536585366	
12	2.1665232268775383	0.3719512195121951	
13	2.1994120900223897	0.3475609756097561	
14	2.229640483856201	0.35060975609756095	
15	2.1632561247523237	0.36890243902439024	
16	2.147991468266743	0.36890243902439024	
17	2.1333967883412432	0.3628048780487805	
18	2.116928574515552	0.38109756097560976	
19	2.1396755009162716	0.36890243902439024	
20	2.1407014829356497	0.36585365853658536	
21	2.119234611348408	0.3597560975609756	
22	2.1410574709496846	0.35365853658536583	
23	2.144318598072703	0.36890243902439024	
24	2.1245930689137156	0.3719512195121951	
25	2.104356405211658	0.36890243902439024	
26	2.1190659389263247	0.3719512195121951	
27	2.1277136017636553	0.36585365853658536	
28	2.11862483547955	0.3597560975609756	
29	2.113829444094402	0.36585365853658536	
30	2.119030280811031	0.3780487804878049	
31	2.121875315177731	0.3902439024390244	
32	2.1285853182397236	0.38109756097560976	
33	2.086743628106466	0.3871951219512195	
34	2.123986322705339	0.3597560975609756	
35	2.102312431102846	0.36585365853658536	

METHODS - TEXT RECOMMENDER

1. Find relevant passages using semantic similarity
2. Uses sentence-transformers

<https://github.com/UKPLab/sentence-transformers>

stsb-roberta-large

3. Steps:

- a. User inputs text
- b. Tool predicts the Content topic of the text using the attribute model
- c. Search raw data for the most similar text passage based on semantic similarity that is of the same Style (Keyword list) and Content (uses the top n predicted class labels)

METHODS - INFRASTRUCTURE

1. Run python scripts on Colab Pro
2. Colab - Drive integration to save weights
3. Streamlit app deployment

local only for now

RESULTS

This project brought me laughter, so I consider that a success.



The religion of Star Wars:

</endofxtxt>In the beginning, the Universe didn't exist. But that all changed in the early universe. It's all about the beginning. In a galaxy far, far...

The wisdom of The Beatles:

[illegible]

ABOUT ME

Background:

- degrees in biology/ecology (stats + simulation modeling)
- researcher turned applied data scientist

Current Role:

- data scientist @ PwC - lead a dev team for ML-enabled products (not a focus on NLP so fun growth edge!)