### **Project Title**

LDA (Latent Dirichlet Allocation) Topic Modeling Playground

### **Description**

### This repository contains a Jupyter Notebook that can be used as a tutorial to learn about how to conduct and evaluate LDA topic modeling, as well as to interpret and present the results. Some helpful readings to gain a deeper understanding of topic modeling are also included in the notebook and presented below as well. These can help you increase your general knowledge of topic modeling and learn different topic modeling approaches.

You can run the topic\_modeling\_playground\_notebook in a shared project space in which all required packages are already installed. The project directory includes the following:

* LDA topic modeling playground Jupyter notebook
* Sample data for topic extraction
* yaml file to set up a conda environment in your own space

## **topic\_modeling\_playground\_notebook.ipynb**

This notebook shows a step-by-step workflow to conduct an LDA topic modeling, including steps to process raw data, run LDA topic modeling, and evaluate and visualize the results. The contents of the notebook are grouped into five parts:

* Data preparation (Steps 1 - 3)
* Model training (Step 4)
* Model evaluation (Steps 5 - 7)
* Visualization of results (Step 8)
* Identification of dominant topics (Step 9)
* Model improvement (Steps 10 - 11)

## **papers.csv**

This is a Kaggle dataset of published COVID-19 papers from May 2020 to May 2021. Topics are generated from this dataset using the LDA topic modeling.

## **topic\_modleing\_env.yml**

This is the conda environment file for a virtual environment which should contain all necessary packages to run the notebook. You don’t need this file if you are using the shared environment in which all required packages and libraries are already installed for you. To create your personal topic\_modeling conda env, run this command in the root directory:

conda env create -f topic\_modeling\_env.yml

For more information on setting up your own conda environment, check out [this page][(http://rcpedia.stanford.edu/topicGuides/sharedCondaEnv.html](http://rcpedia.stanford.edu/topicGuides/sharedCondaEnv.html)).

**How to Use the Notebook**

To use this Topic Modeling Playground Notebook, you need to activate your environment using the full “directory path” and click on the `Topic\_modeling\_playground\_notebook.`

To activate this shared environment, please run the following command:

conda activate /zfs/projects/darc/sandbox\_projects/topic\_modeling/

To deactivate, simply run:

conda deactivate

Once you launch the notebook, you can run the code in each cell and see the results in an interactive way. Please note that this is a read-only notebook. You won’t be able to edit the contents of the notebook; however, you can still run cells interactively. If you would like to have your own copy, make a copy in your own directory and set up your conda environment using the yaml file provided.

**Helpful Readings**

**Understanding Topic Modeling**:

* [Latent Dirichlet Allocation](<https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>)
* [Reading Tea Leaves: How Humans Interpret Topic Models](<http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf>)

**Other topic modeling approaches:**

* **[**LDA2vec: Word Embeddings in Topic Models](<https://towardsdatascience.com/lda2vec-word-embeddings-in-topic-models-4ee3fc4b2843>)
* [Latent Semantic Analysis (LSA)](<https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/>)
* [Non-negative matrix factorization](<https://radimrehurek.com/gensim/models/nmf.html>)
* [Structural Topic Modeling](<https://cbail.github.io/SICSS_Topic_Modeling.html#structural-topic-modeling>)