

General Overview of Repository

- data
 - three different states of the data (raw, preprocessed, feature extracted etc.)
- data_preprocessing
 - one preprocessing .py script
- feature_extraction
 - five .py scripts: 3 different feature extraction methods, base class and main script calling all functions.
- feature_analysis
 - one .py script calling on functions defined in scripts under feature extraction, outputting feature analysis visuals
- model_training
 - two .py scripts: one defining the functions, another one calling them and actually performing the model training.
- test
 - three different test, py files for preprocessing, feature extraction and model training
- pyproject.toml
 - sets up: build, project info, dependencies, optional dependencies, main scripts (for feature extractions and analysis), linting, static typing
- tox.ini
 - testing and automating

```
preprocess_text()
create_folder()
save_post()
save_preprocessed_post()
fetch_additional_posts_from_users()
fetch_posts_from_subreddit()
fetch_posts_from_all()
print_file_counts_for_folder()
```

```
class FeatureExtractor
__init__()
load_documents_and_labels()
preprocess_text()
save_to_csv()

generate_summary_table()
generate_empath_table()
```

```
class EmpathFeatureExtractor(FeatureExtractor)

__init__()
extract_empath_features()
analyze_correlations()
generate_correlation_table()
save_correlation_table()
save_features_and_results()
```

```
class LDAFeatureExtractor(FeatureExtractor)

__init__()
preprocess_documents()
train_lda()
extract_topic_distributions()
topic_distribution_to_matrix()
save_features()
run_tsne()
generate_topic_table()
visualize_tsne()
run_feature_extraction()
run_feature_analysis()
```

```
class NGramFeatureExtractor(FeatureExtractor)

__init__()
extract_features()
compute_pmi()
filter_bigrams_by_pmi()
save_features()
compute_frequencies()
generate_wordclouds()
save_wordclouds()
```

```
class ModelTrainer

__init__()
load_and_combine_data()
preprocess_data()
train_model()
evaluate_model()
run_pipeline()

save_results_as_image()
run_and_save_results()
```

```
class TestDataPreprocessing(unittest.TestCase)

setUp()
test_preprocess_and_save()
```

```
class TestFeatureExtraction(unittest.TestCase)

setUp()
test_ngram_extraction()
test_empath_extraction()
test_lda_extraction()
```

```
class TestModelTraining(unittest.TestCase)

setUp()
test_load_and_combine_data()
test_preprocess_data()
test_train_model()
test_evaluate_model()
test_pipeline_execution()
teardown()
```