```
* b15fc85 (HEAD -> main, origin/main, origin/HEAD) (update completed notebook for
using, 2022-05-30)
* 6643fd8 (update READme, 2022-05-30)
* 9759c68 (create colab notebook for control preparing data, 2022-05-29)
* 6efaa03 (get_perplexity problem, 2022-05-29)
* 23e02a5 (update to device get_perplexity, 2022-05-29)
* a898d4b (update to device get_perplexity, 2022-05-29)
* 197e834 (update to device get_perplexity, 2022-05-29)
* cfd0012 (update to device get_perplexity, 2022-05-29)
* f9fbc6b (add to(device) to perplexity, 2022-05-29)
* f705502 (clean folders update visualization, 2022-05-29)
* 134370d (run time figure, 2022-05-29)
* 216388b (update main_lda.py show ppl each batch, 2022-05-29)
* 8e0372b (remove etm-julia from dir, 2022-05-29)
* c054859 (remove testing_notebooks, 2022-05-29)
* 246e4a7 (remove testing_notebooks, 2022-05-29)
* 1953962 (Update README.md, 2022-05-29)
* 9bb13d8 (Update README.md, 2022-05-29)
* 8a8092a (Update README.md, 2022-05-29)
* 566c9a7 (Update README.md, 2022-05-29)
* bcb00ee (perplexity update, update notebook, 2022-05-29)
* 0a463e2 (update perplexity for lda, main_lda.py command, add new
topicPerplexityNew, 2022-05-28)
* 4397f69 (Update README.md, 2022-05-28)
* f6da1e4 (Update README.md, 2022-05-28)
* 1e7cb84 (add lda-perplexity as batches of 100, 2022-05-28)
* 266481b (visualization runtime update path figures, 2022-05-28)
* d113463 (visualization runtime update path figures, 2022-05-28)
* 725156b (visualization runtime, update normalized perplexity for ETM, 2022-05-28)
* cfab556 (bert problem solved, update not_in_bert_vocab, update read_prefitted for
bert-embedding, 2022-05-28)
* d202402 (perplexity build in, 2022-05-28)
* d136413 (solved bert-vocab for all min_df, 2022-05-27)
*   83b0861 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space into main, 2022-05-27)
|\
| * 7255c3f (notebook for complete running from data to etm, 2022-05-27)
* | f3c36f3 (tiny change self.train(), 2022-05-27)
|/
* c2a1a5e (testdataset check, 2022-05-27)
* 3ddf364 (add visualization=True to train(), 2022-05-27)
* fb3e5f6 (run notebook main.py with all args, 2022-05-27)
* f05ded7 (control args.lr, args.wdecay, update plot for neg-kld, 2022-05-27)
* ad485e3 (change KLD plot, add args.lr, args.wdecay to main, 2022-05-27)
* d6428dc (run bert-prefitted-ETM, save figures, topics by word2vec_model, add
comments to main, 2022-05-27)
* 710fcff (update notebook etm, 2022-05-27)
* 52f5d32 (solved conflict vocab file, 2022-05-27)
* 3373c58 (ETM error found at cross-entropy, 2022-05-27)
* 1dbc897 (remove sorted train_indices from get_doc_in_words, 2022-05-26)
* 7670261 (control prepared_data, update read_prefitted, todo testdataset, 2022-05-
27)
* 479d205 (update notebook main and notebook embeddings, 2022-05-26)
* 5fa13e5 (update notebook_replication cause of implementation changes in /src,
2022-05-26)
* 2f64a47 (results etm by hidden-size and activat func, 2022-05-26)
* 4ba5530 (add new args to main for hidden-size and activat func, 2022-05-26)
* 0051ac3 (add new args to main, save etm-topics with topics und without topics
directory, 2022-05-26)
```

```
* f0fc8a9 (add new args to main args.filter_stopwords and use_bert_embedding, 2022-05-26)
* 459043a (lda new runs stopwords and without stopwords, 2022-05-26)
* c6ef450 (add not_in_bert_vocab, update prepare_dataset.py, 2022-05-26)
* 38d787a (lda run stopwords and without stopwords, 2022-05-25)
*   08a59a4 (update epochs for LDA, 2022-05-25)
|\
| *   17ed1e8 (das was schon oben war übernehmen, 2022-05-25)
| |\
| * | fbf97cf (Fehlerbehebung in Perplexity, 2022-05-25)
* | | d3798bf (update epochs for LDA, 2022-05-25)
| |/
|/|
* | 5be92eb (move runtime files, 2022-05-25)
* | e3bb55a (notebook compare similar words of different embeddings methods, 2022-05-25)
* | c41c744 (remove old BERT notebook, remove old lda corpus function, 2022-05-25)
* | 4d74944 (update documentation for implementation words-embeddings with BERT, 2022-05-25)
* | 742b689 (control of correctness of implementation with bert, 2022-05-25)
* | b0d7dfa (add not_in_bert_vocab, 2022-05-25)
* | a036309 (LDA results by min_df, notebooks documentation, 2022-05-25)
|/
* b01b7d6 (fixing lda gensim error, 2022-05-25)
* 84a9103 (etm and lda num_topis 20 min_df 100, 2022-05-25)
* ac822d2 (etm and lda num_topis 20 min_df 100, 2022-05-25)
* 6316f8b (etm and lda num_topis 20, 2022-05-25)
* a13da3b (etm min_df 2 num_topis 50, 2022-05-24)
* 27fe96b (etm min_df 2 10, 20 epochs, 2022-05-24)
* 5886160 (etm 200 n-topcs 10 epochs results, 2022-05-24)
* be46c13 (etm 100 epochs results, 2022-05-24)
* 5153326 (add torch to embedding.py, 2022-05-24)
* f6363ed (update main_eval_checkpoints, update seed=42 embeddings, 2022-05-24)
* b07dc3a (change args.min_df, 2022-05-24)
* 0faf952 (epochs 100, update main.py with num_toics, update train_etm save by num_topics, add load ckpt, 2022-05-24)
* 981674b (remove textsloader from main_lda, 2022-05-24)
* c9395c7 (add loss, min_df args to main.py, update read_prefitted, update find similar words, 2022-05-24)
* b9f0386 (update bert_preparing remove stopwords list, 2022-05-24)
* c81ab18 (update bert preparing fix splitting long sentences, 2022-05-24)
* 12dfa7b (solved similar problem, 2022-05-24)
* 501dab7 (remove cache topics, 2022-05-24)
* acc32b3 (update gitignore file, 2022-05-24)
* 812765e (problem lda by min_df, 2022-05-24)
* cc4471b (problem mit lda, update evaluierung with number words, 2022-05-24)
* 3e0a268 (not upload min_df dirs, 2022-05-23)
* 91b0010 (remove saving, update ignore file, run notebook, 2022-05-23)
* 1a0cc48 (testing save_path in notebook, 2022-05-23)
* 0ecd32d (embedding.py change save path, 2022-05-23)
* 897a45a (remove save prepared_data/ outputs from code, 2022-05-23)
* fa436be (add checkpoints, 2022-05-23)
* 5bc6be4 (push etm_julia, 2022-05-23)
* 9513197 (test gitignore after rm --cached, 2022-05-23)
* cfb5892 (add prepared_data/ in gitignore, 2022-05-23)
* dee1b63 (test add after conflicts, 2022-05-23)
*   09864d2 (Merge branch 'main' of https://github.com/hanhluukim/replication-topic-modelling-in-embedding-space, 2022-05-23)
|\
| * 55cc7d2 (test push after conflicts, 2022-05-23)
* | 396a07d (adding the exponential for perplexity, 2022-05-23)
|/
* 1ae8189 (fixing syntax errors, 2022-05-23)
* 5a134ac (Test, 2022-05-23)
*   b539281 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space, 2022-05-23)
|\
| * edb902c (change save embedding format, add new read_prefitted embeddings,
todo:notebook, 2022-05-23)
| * b4e4a4e (testing similar words of embeddings, 2022-05-22)
| * 04edcf1 (remove save bert from bert_main, update find_similar_words, 2022-05-
22)
| * f222a4a (test bert_vocab of two sentences, 2022-05-22)
| * 8bccc8d (update embedding, find_similiar_words_self_implemented, 2022-05-22)
| *   6627429 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space, 2022-05-22)
| |\
| | * 0243046 (add BertEmbedding, implemented find_similar_words, 2022-05-22)
| * | 402eebc (remove subprocess bert from main.py, 2022-05-22)
| * | 611f267 (add BertEmbedding, implemented find_similar_words, 2022-05-22)
| |/
| * ba5f1ab (remove run bert_embedding.py, 2022-05-22)
| *   3dfc23a (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space into main, 2022-05-22)
| |\
| | * 7a25673 (check word_ids tokenizerfast, 2022-05-22)
| | * 23a4cc4 (check word-ids tokenizerfast, 2022-05-22)
| * | 6fda8b2 (sort embeddings by order in vocabulary, 2022-05-22)
| |/
| * deddea7 (remove prints in bert, 2022-05-22)
| * a1ad272 (add run bert_main.py to notebook, 2022-05-22)
| * d67fa96 (remove evaluation code of authors, add from src.evaluation_of_authors,
2022-05-22)
| *   77cb6b0 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space into main, 2022-05-22)
| |\
| | * 54c39ac (copy TopicCoherrence and TopicDiversity to notebook, 2022-05-22)
| * | d253a1f (add evaluation of authors, 2022-05-22)
| |/
| * 458e971 (move bert_main.py, update readme, 2022-05-22)
| * d69308f (add fixed seed and deteministic net layers, 2022-05-22)
| * 4fe067e (update readme, 2022-05-22)
| * fb9e1d8 (build bert in main, changed embedding path, update read prefitted,
2022-05-22)
| * e9414de (test completed bert, 2022-05-22)
| * 9e79b83 (add bert completed, 2022-05-22)
| * c48e6a5 (add bert completed, 2022-05-22)
| * aabcfea (bert-embedding completed, 2022-05-22)
| * ee69fc9 (Created using Colaboratory, 2022-05-22)
| * f4d1939 (save bert by sent, 2022-05-22)
| * 04a5763 (rename bert notebook, 2022-05-22)
| * d0cf4b0 (create embedding for each word in a sentence, 2022-05-22)
| * 35d4f11 (create embedding for each unique word in a sentence, 2022-05-22)
| * 3a07e22 (error download bert, 2022-05-21)
| * 2aa9919 (Update README.md, 2022-05-21)
| * 8df3f50 (update bert_embedding, 2022-05-21)
* |   b1bfe20 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space, 2022-05-21)
```

```
|\ \
| |/
| *   8ddd06a (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space into main, 2022-05-21)
| |\
| * | 7eb586d (conflicts solved, first version bert_embedding, 2022-05-21)
* | | 4addede (fixing Syntax error, 2022-05-21)
| |/
|/|
* | 16a20ad (bei perplexity 2.teil mit einbezug des 2. Teil der Dokumente
eingebaut, 2022-05-21)
* | fff935d (todo: same word/belonging subwords in different positions in the
sentence, 2022-05-21)
* | 15c3655 (get subwords-embedding und create wb, 2022-05-21)
|/
*   a62bed3 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space into main, 2022-05-20)
|\
| *   18506c2 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space, 2022-05-20)
| |\
| | * dda4181 (add BERT implemetation, 2022-05-20)
| * | ddd67ca (fixing naming error, 2022-05-20)
| |/
| * 0bbad58 (fixing Syntax errors, 2022-05-19)
* | e625918 (remove cv2, 2022-05-20)
|/
*   3114602 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space into main, 2022-05-18)
|\
| * 58b3ef2 (update figure ETM in notebook, 2022-05-18)
* | e4942af (extend figure with embedding part, 2022-05-18)
|/
* 443f61f (update pipeline, 2022-05-18)
* 11e690a (update notebook, 2022-05-18)
*   8d06aff (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space into main, 2022-05-18)
|\
| * 7a5d7f8 (add more explanation and documentation for notebook, 2022-05-18)
* | 1a6e259 (remove old codes, 2022-05-18)
|/
* 59afcf1 (add new ETM architecture, 2022-05-18)
* 1130980 (add documentation to code, 2022-05-18)
* b5eff27 (test other activation function and weight decay, 2022-05-18)
* 4ca818f (test different loss functions, 2022-05-18)
* 7bdecd1 (test different loss functions, 2022-05-18)
* 424f912 (update loss function, remove K from KLD, 2022-05-18)
* 47f74ee (update loss function, remove K from KLD, 2022-05-18)
* 0265572 (update loss function, remove K from KLD, 2022-05-18)
* 2af324e (add loss-name to args, update def loss_function, 2022-05-18)
* 2999275 (bitbucket dowload nyt, 2022-05-18)
* fadf00e (update loss function followed paper 2020, 2022-05-18)
* 92d0c0a (statistic vocab update, 2022-05-17)
* 53fb54d (save statistics of different corpora, 2022-05-17)
* c6a9faa (update embedding paramerters followed Mikolov paper, 2022-05-17)
* 030c55c (add wordvec arg to main.py, 2022-05-17)
* b2f73b9 (add wordvec arg to main.py, 2022-05-17)
* 04e4ef7 (remove some prints, 2022-05-17)
* 0ecabb4 (etwas für test2 in perplexity, 2022-05-16)
* 4220ff4 (variablenumbennenung perplexity, 2022-05-16)
* 893a21e (kleiner Teil von Perplexity, 2022-05-14)
* 341074e (was ich wahrscheinlich brauch in perplexity function, 2022-05-14)
* 9d5d4f1 (erster Versuch topicDiversity, 2022-05-14)
* 47e15fb (hizufügen evaluierungsgerüste und todo für Perplexity und Diversity,
2022-05-14)
* 8da8fb0 (topicCoherence2 wo pointwiseinf mit anderer marg funktion berechnet
wird, 2022-05-14)
* 04f0f57 (erweiterung pointwiseInf mit anderer marg funktion, 2022-05-14)
* 4048a88 (andere Berechnung für marg hinzugefügt, 2022-05-14)
* e74d7be (Compare the results own TC and original TC of authors, 2022-05-13)
* a098fa6 (Compare the results own TC and original TC of authors, 2022-05-13)
* 04f07d0 (check coherrence of ETM, 2022-05-11)
* 08be922 (test gensim.coherencemodel, 2022-05-11)
* 23b291d (update notebook with lda, topicCoherrence, 2022-05-08)
* 0f6e6d4 (update visualization.py, main.py, 2022-05-08)
* b367d52 (update save figures by min_df, 2022-05-08)
* 905c8a0 (Fall für keine Dokumente mit selben Wort hinzugefügt, 2022-05-08)
* c91c8fc (add visualization.py for embedding space, 2022-05-07)
* 39878e2 (update main.py, 2022-05-07)
* 053ebc5 (check other loss function, 2022-05-07)
* 14e4669 (check show_topics, 2022-05-07)
* 4527ad2 (add save word2vec model, 2022-05-07)
* ea7925c (add show_topics to ETM, 2022-05-07)
* fe6356a (add kld visualization, 2022-05-07)
* 91e6726 (train epoch 500, 2022-05-06)
* ad7f80b (update save path by min_df, update etm predictions, 2022-05-06)
* 8088006 (save bowmat by min_df, sorting of embedding-voca, 2022-05-06)
* d9bf35d (add save checkpoints, 2022-05-06)
* da9b028 (add save checkpoints, 2022-05-06)
* 6f0b1b6 (update notebook, 2022-05-06)
* 287b5dc (add explanation of code, 2022-05-06)
*   76bc015 (save preprocessd docs, check cuda, 2022-05-06)
|\
| * 49ec8c9 (check KL Loss in notebook, 2022-05-06)
* | d863034 (save preprocessd docs, check cuda, 2022-05-06)
|/
* 3871320 (update julia-ETM, 2022-05-06)
* f0e9650 (julia setting, 2022-05-06)
* e9995d5 (Update README.md, 2022-05-06)
* 6db1ed8 (Update README.md, 2022-05-06)
* 843ed42 (for julia read prefitted embedding, create bows, 2022-05-06)
* a8b35d0 (save id2word to file, 2022-05-05)
* b993e36 (save id2word to file, 2022-05-05)
* bcfbb75 (save id2word to file, 2022-05-05)
* c58ed42 (add first julia implementation, 2022-05-06)
* f0bfb71 (compute runtime, check train, 2022-05-05)
*   fa515c6 (check loss function, 2022-05-05)
|\
| * 23177ed (run TrainETM with normalize_data, 2022-05-05)
* | 1166f20 (check loss function, 2022-05-05)
* | 6d054b8 (check loss function, 2022-05-05)
|/
* a576041 (add normalize_data attribute for class DocSet and training, 2022-05-05)
*   2c7eec2 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space into main, 2022-05-05)
|\
| * b2eb69f (add etm architecture in notebook, 2022-05-05)
* | 3c16e56 (add get_normalized_batch to TrainETM(), 2022-05-05)
```

```
|/
* 856831a (add figure for ETM, 2022-05-05)
* 7b09fa6 (prepared_data folder, 2022-05-05)
* 0ac4f77 (update prepare_dataset after name change, 2022-05-05)
* 9dbf96e (add saving preprocessed data in prepared_data, 2022-05-05)
* e840394 (add wb_creator.cluster_words in notebook, 2022-05-05)
* f18bbbb (add umap n_components 2, 2022-05-05)
* 0585c09 (update TrainETM, Umap Visualization, prepare_dataset, 2022-05-05)
* 5cf127b (fixing more syntax errors in evaluierung.py, 2022-05-04)
* 00d5bf4 (fixing tab misposition in evaluierung.py, 2022-05-04)
* 2a73b91 (fixing lda tab and word mispositions, 2022-05-04)
*   3641675 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space, 2022-05-04)
|\
| * 16cbbdc (test ETMTrain() in notebook, 2022-05-04)
* | 7ac40c3 (ausbersserung des imports für lda und mit ids, 2022-05-04)
|/
*   c5845d3 (Merge branch 'main' of https://github.com/hanhluukim/replication-
topic-modelling-in-embedding-space, 2022-05-04)
|\
| * 97d0fd1 (test train etm, update lossfunction, add main args, 2022-05-04)
* | 917ce4f (marginal word, 2022-05-04)
|/
* 078c891 (lda gensim, 2022-05-04)
* a62f3fc (topic Cohorence, 2022-05-04)
* 4a9947f (normalized pointwise Information, 2022-05-04)
* 3a6af95 (Wahrscheinlichkeit wörter in einem Dokument, 2022-05-04)
* 502c70c (test run train in notebook, 2022-05-04)
* 86bae3f (test training first-draft ETM, 2022-05-04)
* dabae6a (make dataset split deterministic, 2022-05-04)
* 34733a0 (change split dataset method, 2022-05-04)
* cd0b5de (update notebook for first-draft-ETM, 2022-05-04)
* d3d2f22 (update main, etm, train_etm, 2022-05-04)
* 688531b (fix bug, 2022-05-04)
* b6267ae (fix bug, 2022-05-04)
* d26488d (update main.py and src, 2022-05-04)
* cdeb9ae (update notebook, 2022-05-04)
* 4ec650c (update main, etm, train_etm, 2022-05-04)
* ac1deee (update main.py and src, 2022-05-04)
* fc628d1 (update main.py and src, 2022-05-04)
* 49b81a3 (update main.py and src, 2022-05-04)
* 114938b (update main.py, 2022-05-04)
* 57bcadd (update DocSet, 2022-05-04)
* 521f53b (update DocSet, 2022-05-04)
* 2488529 (update DocSet, 2022-05-04)
* c0556a7 (add first draft of etm and add DocSet for DataLoader, 2022-05-03)
* 33e8361 (update readme, 2022-05-03)
* de5873b (update readme, 2022-05-03)
* d7d850b (update notebook file, 2022-05-03)
* 3ee5de9 (add visualization embedding space, update requirements.txt, 2022-05-03)
* e6b6065 (add visualization embedding space, update requirements.txt, 2022-05-03)
* 6937906 (visualization words by word2vec, clustering and umap dim-reduction,
2022-05-03)
* f807b55 (visualization words by word2vec, 2022-05-03)
* a11dd87 (remove notebook 2, 2022-05-03)
* 7309af0 (add notebook from colab, 2022-05-03)
* 7938ea6 (Created using Colaboratory, 2022-05-03)
* 69acce0 (test train embedding and save embeddings, 2022-05-03)
* 5e5f5a9 (update embedding.py, 2022-05-03)
* a5f7dd6 (evaluierung worte im selben Dokument, 2022-05-02)
* 842c8ba (testmain, 2022-05-02)
| * 0ca5bb2 (origin/Test) (test visual, 2022-05-02)
|/
* e0b67de (update embedding.py, 2022-05-01)
* c063fea (add WordEmbeddingCreator, 2022-04-30)
* 1c3ce49 (add WordEmbeddingCreator, 2022-04-30)
* a378278 (add WordEmbeddingCreator, 2022-04-30)
* 5060f0e (Update README.md, 2022-04-30)
* 132e3bd (re-create docs in words for embedding-training, 2022-04-29)
* 02e3dca (Update README.md, 2022-04-30)
* 3dbee59 (update notebook_replication, 2022-04-30)
* c0250d3 (re-create docs in words for embedding-training, 2022-04-29)
* 348de3b (re-create docs in words for embedding-training, 2022-04-29)
* 5cccaf9 (update transformation for lda corpus, 2022-04-29)
* 62fd53c (Update README.md, 2022-04-27)
* 7b2fa34 (Update README.md, 2022-04-27)
* cb90028 (update pipeline, 2022-04-25)
* d94340c (update pipeline, 2022-04-25)
* 5aaa43b (finished prepare_dataset.py, 2022-04-24)
* f0f0252 (add notebook_replication to check the outputs, 2022-04-24)
* c72c7f6 (Update README.md, 2022-04-24)
* ed1f0e0 (Update README.md, 2022-04-24)
* f1c3ecb (add parse.arg prepare_dataset.py, 2022-04-24)
* 7952ea4 (finished fixing prepare_dataset.py, 2022-04-23)
* fe7d976 (fix create_bow prepare_dataset.py, 2022-04-23)
* 9fd8094 (update prepare_dataset.py, 2022-04-23)
* bba3723 (add main.py, check prepare_dataset.py, 2022-04-22)
* cd8eb9a (Update README.md, 2022-04-22)
* 471d463 (Update README.md, 2022-04-22)
* 8192c6e (Update README.md, 2022-04-22)
* cdb4730 (test git push from google colab, 2022-04-21)
* 0889b98 (test git push from google colab, 2022-04-21)
* b70e6bc (update testing and preprare_dataset.py, 2022-04-22)
* ef00f9e (update testing and preprare_dataset.py, 2022-04-21)
* 03cf5ee (add update testing preprocessing, 2022-04-21)
* 8b1549d (add notebooks testing preprocessing and lda, 2022-04-20)
* 664e441 (origin/jupyter-branch) (add init codes and todos, 2022-04-19)
* 22450ab (Initial commit, 2022-04-07)
```