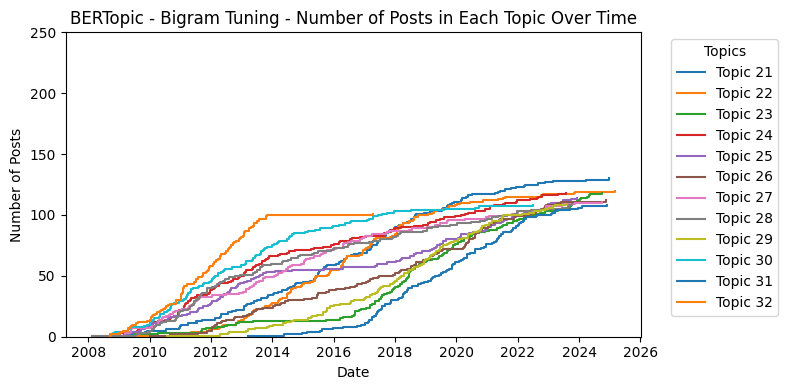
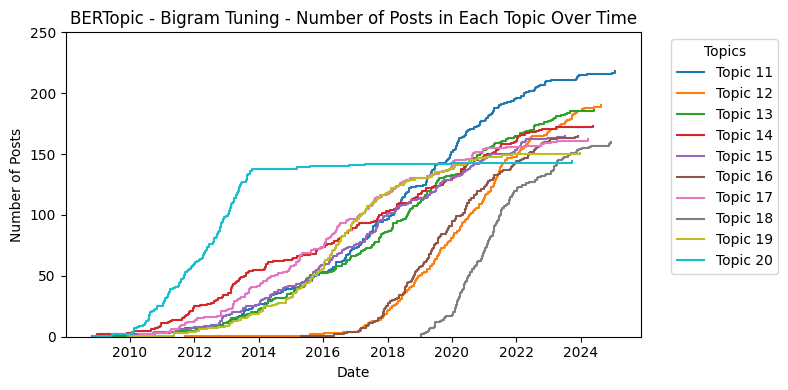
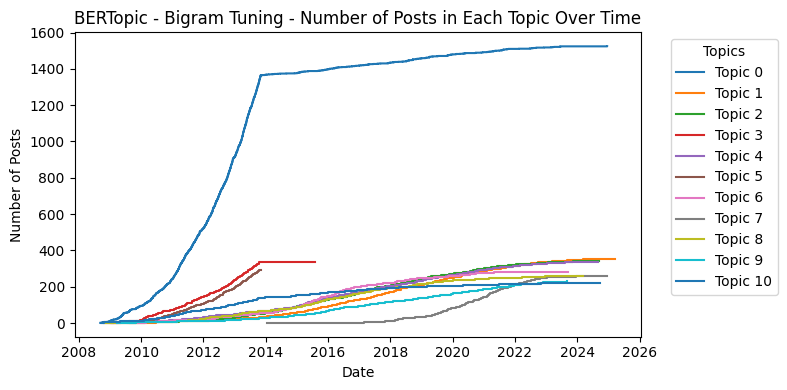
BERTopic Final Model Full Topics Results



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Topics | Topic Name | Top 10 words | Example | Link |
| 0 | 0\_text file\_txt file\_line text\_read text | ['text file', 'txt file', 'line text', 'read text', 'line file', 'line line', 'file text', 'file using', 'file line', 'log file'] | counting the number of lines in a text file | <https://stackoverflow.com/questions/3482064/counting-the-number-of-lines-in-a-text-file> |
| 1 | 1\_text classification\_scikit learn\_test set\_training set | ['text classification', 'scikit learn', 'test set', 'training set', 'machine learning', 'random forest', 'multi label', 'test data', 'classify text', 'training data'] | How to train a model to classify input to one or more classes | <https://stackoverflow.com/questions/62677043/how-to-train-a-model-to-classify-input-to-one-or-more-classes> |
| 2 | 2\_import nltk\_install nltk\_nltk download\_nltk data | ['import nltk', 'install nltk', 'nltk download', 'nltk data', 'nltk package', 'nltk corpus', 'following error', 'use nltk', 'installing nltk', 'nltk python'] | Installing megam for NLTK on Windows | <https://stackoverflow.com/questions/46852316/installing-megam-for-nltk-on-windows> |
| 3 | 3\_using jquery\_text input\_input field\_get text | ['using jquery', 'text input', 'input field', 'get text', 'change text', 'text jquery', 'jquery text', 'text box', 'text node', 'javascript jquery'] | Remove all HTMLtags in a string (with the jquery text() function) | <https://stackoverflow.com/questions/7889765/remove-all-htmltags-in-a-string-with-the-jquery-text-function> |
| 4 | 4\_sentiment analysis\_positive negative\_sentiment score\_analysis using | ['sentiment analysis', 'positive negative', 'sentiment score', 'analysis using', 'negative neutral', 'negative word', 'naive bayes', 'stanford nlp', 'negative sentiment', 'negative positive'] | How to calculate the quality of comments using NLP | <https://stackoverflow.com/questions/76927485/how-to-calculate-the-quality-of-comments-using-nlp> |
| 5 | 5\_html cs\_center text\_align text\_text overflow | ['html cs', 'center text', 'align text', 'text overflow', 'text wrap', 'text align', 'image text', 'nbsp nbsp', 'using cs', 'make text'] | How to center text vertically in HTML using CSS only | <https://stackoverflow.com/questions/2943234/how-to-center-text-vertically-in-html-using-css-only> |
| 6 | 6\_parse tree\_nltk tree\_tree tree\_stanford parser | ['parse tree', 'nltk tree', 'tree tree', 'stanford parser', 'dependency parser', 'dependency tree', 'stanford dependency', 'typed dependency', 'using stanford', 'np tree'] | How to do a Tree Transfer in prolog for MT | <https://stackoverflow.com/questions/19580566/how-to-do-a-tree-transfer-in-prolog-for-mt> |
| 7 | 7\_ner model\_spacy ner\_named entity\_using spacy | ['ner model', 'spacy ner', 'named entity', 'using spacy', 'spacy model', 'custom ner', 'entity recognition', 'model spacy', 'training data', 'custom entity'] | Remove training data from spacy model | <https://stackoverflow.com/questions/60970109/remove-training-data-from-spacy-model> |
| 8 | 8\_po tag\_po tagger\_part speech\_po tagging | ['po tag', 'po tagger', 'part speech', 'po tagging', 'tag word', 'none none', 'stanford po', 'nltk po', 'using nltk', 'word tag'] | NLTK POS tags extraction, tried key, values but not there yet | <https://stackoverflow.com/questions/37234753/nltk-pos-tags-extraction-tried-key-values-but-not-there-yet> |
| 9 | 9\_sentence tokenizer\_using nltk\_nltk tokenize\_tokenize sentence | ['sentence tokenizer', 'using nltk', 'nltk tokenize', 'tokenize sentence', 'nltk sentence', 'nltk word\_tokenize', 'one token', 'sentence word', 'word tokenize', 'want tokenize'] | Why can&#39;t I tokenize text in languages other than English using NLTK? | <https://stackoverflow.com/questions/72279174/why-cant-i-tokenize-text-in-languages-other-than-english-using-nltk> |
| 10 | 10\_unicode text\_unicode character\_utf encoding\_codec decode | ['unicode text', 'unicode character', 'utf encoding', 'codec decode', 'decode byte', 'code point', 'text file', 'text encoding', 'special character', 'character encoding'] | (HTML5) cannot display Chinese characters when using UTF-8 | <https://stackoverflow.com/questions/11738217/html5-cannot-display-chinese-characters-when-using-utf-8> |
| 11 | 11\_lda model\_topic modeling\_topic modelling\_topic model | ['lda model', 'topic modeling', 'topic modelling', 'topic model', 'gensim lda', 'topic distribution', 'lda topic', 'latent dirichlet', 'number topic', 'using gensim'] | LdaModel - random\_state parameter not recognized - gensim | <https://stackoverflow.com/questions/39294921/ldamodel-random-state-parameter-not-recognized-gensim> |
| 12 | 12\_program file\_file microsoft\_install spacy\_appdata local | ['program file', 'file microsoft', 'install spacy', 'appdata local', 'lib site', 'site package', 'microsoft visual', 'microsoft sql', 'file user', 'visual studio'] | install issue with python - spacy package in anaconda environment | <https://stackoverflow.com/questions/28399340/install-issue-with-python-spacy-package-in-anaconda-environment> |
| 13 | 13\_stop word\_remove stopwords\_remove stop\_removing stopwords | ['stop word', 'remove stopwords', 'remove stop', 'removing stopwords', 'stopwords list', 'trying remove', 'remove word', 'list stopwords', 'word list', 'want remove'] | Is there a way to remove a word from a KeyedVectors vocab? | <https://stackoverflow.com/questions/54697748/is-there-a-way-to-remove-a-word-from-a-keyedvectors-vocab> |
| 14 | 14\_count number\_word count\_count word\_word frequency | ['count number', 'word count', 'count word', 'word frequency', 'number time', 'number word', 'bla bla', 'count many', 'frequency word', 'time word'] | Coding Matrix with overlap counts in R | <https://stackoverflow.com/questions/40470933/coding-matrix-with-overlap-counts-in-r> |
| 15 | 15\_bigram trigram\_get bigram\_unigrams bigram\_bi gram | ['bigram trigram', 'get bigram', 'unigrams bigram', 'bi gram', 'unigram bigram', 'find bigram', 'nltk bigram', 'bigram list', 'nltk collocation', 'extract bigram'] | Generate unigrams and bigrams from a trigram list | <https://stackoverflow.com/questions/20035974/generate-unigrams-and-bigrams-from-a-trigram-list> |
| 16 | 16\_embedding layer\_output shape\_input shape\_kera model | ['embedding layer', 'output shape', 'input shape', 'kera model', 'attention layer', 'shape none', 'using kera', 'kera embedding', 'input layer', 'functional api'] | Concatenation of list of 3-dimensional tensors along a specific axis in Keras | <https://stackoverflow.com/questions/47325813/concatenation-of-list-of-3-dimensional-tensors-along-a-specific-axis-in-keras> |
| 17 | 17\_using wordnet\_nltk wordnet\_wordnet synset\_synonym word | ['using wordnet', 'nltk wordnet', 'wordnet synset', 'synonym word', 'get synonym', 'get synset', 'word wordnet', 'using nltk', 'wordnet nltk', 'list synset'] | Python wordnet from nltk.corpus. How can I get the definition of each word in several sentences using wordnet? | <https://stackoverflow.com/questions/62287379/python-wordnet-from-nltk-corpus-how-can-i-get-the-definition-of-each-word-in-se> |
| 18 | 18\_bert model\_fine tuning\_using bert\_hidden state | ['bert model', 'fine tuning', 'using bert', 'hidden state', 'use bert', 'pre trained', 'fine tuned', 'trained bert', 'fine tune', 'train bert'] | Bert NER model start and end position None after fine-tuning | <https://stackoverflow.com/questions/76393971/bert-ner-model-start-and-end-position-none-after-fine-tuning> |
| 19 | 19\_stanford nlp\_edu stanford\_java edu\_stanford corenlp | ['stanford nlp', 'edu stanford', 'java edu', 'stanford corenlp', 'nlp pipeline', 'corenlp server', 'jar file', 'java lang', 'pipeline stanfordcorenlp', 'tagger maxent'] | Import Stanford nlp Intellij | <https://stackoverflow.com/questions/40625438/import-stanford-nlp-intellij> |
| 20 | 20\_using php\_text file\_file php\_php file | ['using php', 'text file', 'file php', 'php file', 'php text', 'php script', 'php code', 'txt file', 'file using', 'script php'] | Parsing a formatted text file in PHP | <https://stackoverflow.com/questions/14875886/parsing-a-formatted-text-file-in-php> |
| 21 | 21\_stem word\_stemmed word\_snowball stemmer\_store result | ['stem word', 'stemmed word', 'snowball stemmer', 'store result', 'word stem', 'stemming word', 'stemming algorithm', 'porter stemmer', 'word stemming', 'nltk stem'] | Compute word n-grams on original text or after lemma/stemming process? | <https://stackoverflow.com/questions/47219389/compute-word-n-grams-on-original-text-or-after-lemma-stemming-process> |
| 22 | 22\_naive bayes\_bayes classifier\_nltk naive\_using naive | ['naive bayes', 'bayes classifier', 'nltk naive', 'using naive', 'text classification', 'multinomial naive', 'term exists', 'bayes text', 'confusion matrix', 'test document'] | python nltk naive bayes probabilities | <https://stackoverflow.com/questions/20773200/python-nltk-naive-bayes-probabilities> |
| 23 | 23\_api ai\_wit ai\_chat bot\_api call | ['api ai', 'wit ai', 'chat bot', 'api call', 'using dialogflow', 'chat message', 'rasa nlu', 'firstname lastname', 'bot framework', 'square root'] | how add a required filed to context parameters in dialog flow | <https://stackoverflow.com/questions/50980489/how-add-a-required-filed-to-context-parameters-in-dialog-flow> |
| 24 | 24\_html document\_html tag\_web page\_text html | ['html document', 'html tag', 'web page', 'text html', 'beautiful soup', 'extract text', 'html file', 'text webpage', 'html page', 'using python'] | Web Scraping Rap lyrics on Rap Genius w/ Python | <https://stackoverflow.com/questions/24873773/web-scraping-rap-lyrics-on-rap-genius-w-python> |
| 25 | 25\_pdf file\_text pdf\_pdf document\_extract text | ['pdf file', 'text pdf', 'pdf document', 'extract text', 'pdf text', 'cosfloat pdfoperator', 'pdfoperator tj', 'cosint pdfoperator', 'cosint cosint', 'tc cosstring'] | how to extract text from a selection of pages in a larger pdf using pymupdf? | <https://stackoverflow.com/questions/67782158/how-to-extract-text-from-a-selection-of-pages-in-a-larger-pdf-using-pymupdf> |
| 26 | 26\_seaside seaside\_extract word\_longest common\_list word | ['seaside seaside', 'extract word', 'longest common', 'list word', 'string list', 'word list', 'word word', 'want check', 'check string', 'word sentence'] | Python searching for two words regex | <https://stackoverflow.com/questions/28004731/python-searching-for-two-words-regex> |
| 27 | 27\_tokenizer parser\_symbol symbol\_next token\_source file | ['tokenizer parser', 'symbol symbol', 'next token', 'source file', 'lexical analysis', 'identifier symbol', 'token type', 'explicit string', 'ling tokensregex', 'lexer parser'] | Referencing a definition name in flex | <https://stackoverflow.com/questions/35159776/referencing-a-definition-name-in-flex> |
| 28 | 28\_detect language\_language detection\_google translate\_language text | ['detect language', 'language detection', 'google translate', 'language text', 'different language', 'english arabic', 'text language', 'language recognition', 'language would', 'american english'] | A PHP Library / Class to Count Words in Various Languages? | <https://stackoverflow.com/questions/2935544/a-php-library-class-to-count-words-in-various-languages> |
| 29 | 29\_tf idf\_idf score\_idf value\_scikit learn | ['tf idf', 'idf score', 'idf value', 'scikit learn', 'idf matrix', 'calculate tf', 'learn tfidfvectorizer', 'idf vectorizer', 'tfidf vectorizer', 'using tfidfvectorizer'] | Updating the feature names into scikit TFIdfVectorizer | <https://stackoverflow.com/questions/25154231/updating-the-feature-names-into-scikit-tfidfvectorizer> |
| 30 | 30\_natural language\_language processing\_nlp library\_quick add | ['natural language', 'language processing', 'nlp library', 'quick add', 'information extraction', 'nlp tool', 'semantic network', 'get started', 'processing tool', 'open source'] | What are the prerequisites to learning natural language processing? | <https://stackoverflow.com/questions/2006763/what-are-the-prerequisites-to-learning-natural-language-processing> |
| 31 | 31\_word vec\_vec model\_gensim word\_using gensim | ['word vec', 'vec model', 'gensim word', 'using gensim', 'gensim model', 'fasttext model', 'model gensim', 'load model', 'doc vec', 'model wv'] | Gensim: how to load precomputed word vectors from text file | <https://stackoverflow.com/questions/49750112/gensim-how-to-load-precomputed-word-vectors-from-text-file> |
| 32 | 32\_data type\_text field\_sql server\_text blob | ['data type', 'text field', 'sql server', 'text blob', 'text column', 'text type', 'text varchar', 'mysql text', 'varchar text', 'blob text'] | Convert BLOB into TEXT with foreign characters | <https://stackoverflow.com/questions/3457414/convert-blob-into-text-with-foreign-characters> |

LDA

Zhang, Weinan & Liu, Ting & Yang, Yang & Cao, Liujuan & Zhang, Yu & Ji, Rongrong. (2014). A Topic Clustering Approach to Finding Similar Questions from Large Question and Answer Archives. PloS one. 9. e71511. 10.1371/journal.pone.0071511.

<https://towardsdatascience.com/nlp-with-lda-latent-dirichlet-allocation-and-text-clustering-to-improve-classification-97688c23d98/>

<https://www.kaggle.com/code/panks03/clustering-with-topic-modeling-using-lda>

<https://medium.com/exness-blog/topic-modeling-for-surveys-or-how-we-solved-the-problem-of-text-clustering-with-lda-ef4896e6f905>

BERTopic

<https://medium.com/data-science/topic-modeling-with-bert-779f7db187e6>

<https://github.com/MaartenGr/BERTopic>

<https://towardsdatascience.com/topic-modelling-with-berttopic-in-python-8a80d529de34/>

<https://maartengr.github.io/BERTopic/algorithm/algorithm.html>

<https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2>

<https://medium.com/@lukei_3514/dealing-with-contractions-in-nlp-d6174300876b>

https://github.com/MIND-Lab/OCTIS