Semantic Search in articles

Data Set

Medium is a prominent platform for disseminating knowledge across a wide range of topics. It is particularly renowned for articles on Machine Learning, Artificial Intelligence, and Data Science. This dataset comprises a collection of approximately 350 articles within these domains.

The dataset originally contains a total of **192,368** articles. For the purpose of testing the pipeline, a subset of **200** articles was selected. These **200** articles were further split into sentences, from which **400** sentences were chosen to create pairs for training a sentence transformer.

The statistics for the pairs created are as follows:

- **Training Set:** 64,160 pairs
- **Validation Set:** 14,436 pairs
- **Testing Set:** 1604 pairs

Data Set

- To assess the similarity between pairs, the dataset utilizes **Term Frequency-Inverse Document Frequency (TF-IDF)** and **cosine similarity matrices**. These measures help in evaluating and refining the quality of the sentence pairs for the sentence transformer model.

Preprocessing

- **Tokenization:** The text is split into individual tokens or words.
- **Lowercasing:** All text is converted to lowercase to ensure uniformity.
- **Removal of Stop Words:** Common, non-informative words are removed.
- **Punctuation Removal:** Punctuation marks are eliminated to focus on meaningful words.

Fine Tune Sentence Transformer

- The **all-MiniLM-L6-v2** model This is a sentence-transformers model: It maps sentences & paragraphs to a 768 dimensional dense vector space and can be used for tasks like clustering or semantic search
- Loss **Cosine Similarity Loss for pairs.**
- **Training Results:**

Step	Training Loss	Validation Loss	Pearson Cosine	Spearman Cosine	Pearson Manhattan	Spearman Manhattan	Pearson Euclidean	Spearman Euclidean	Pearson Dot	Spearman Dot	Pearson Max	Spearman Ma
800	0.004100	0.004052	0.772340	0.291697	0.880912	0.275734	0.891707	0.291697	0.772340	0.291699	0.891707	0.29169
1600	0.003700	0.003741	0.786892	0.302689	0.887709	0.286841	0.899778	0.302689	0.786892	0.302689	0.899778	0.30268
2400	0.003600	0.003562	0.794998	0.306467	0.891635	0.290215	0.903894	0.306467	0.794998	0.306466	0.903894	0.30646
3200	0.003700	0.003410	0.803787	0.317407	0.896580	0.301854	0.909131	0.317408	0.803787	0.317407	0.909131	0.31740
4000	0.003400	0.003342	0.806801	0.323789	0.897954	0.307302	0.910576	0.323789	0.806801	0.323789	0.910576	0.32378

Fine Tune Sentence Transformer

- **Testing Results:**
  a. **Pearson Correlation Coefficient:**
     - **Cosine Similarity**: 0.767
     - **Manhattan Distance**: 0.878
     - **Euclidean Distance**: 0.895
     - **Dot Product**: 0.767
     - **Max Similarity**: 0.895
  b. **Spearman Rank Correlation Coefficient:**
     - **Cosine Similarity**: 0.335
     - **Manhattan Distance**: 0.330
     - **Euclidean Distance**: 0.335
     - **Dot Product**: 0.335
     - **Max Similarity**: 0.335

FAISS Vector Database

- **Embedding Generation:** A vector database is created for the article texts using the FAISS library and the all-mpnet-base-v2 model from the Sentence Transformers library. This involves generating embeddings for each article using the Hugging Face API.

Search

- **Query Embedding:** An embedding is generated for the search query (key).
- **Article Matching:** The most relevant articles are identified by finding the top N matches based on cosine similarity within the FAISS database.
- **Generate N-Hot Keywords:**
  - **Keyword Extraction:** From the most relevant articles, a **TF-IDF/Count Vectorizer** is used to extract the top N hot keywords.
- **Visualization:**
  - **Word Cloud:** A visual representation of the N-hot keywords is created using a word cloud to highlight the most significant terms.

Results

- **Search Query:** "machine learning"
- **Top Matched Article:**
  - *"A machine learning method uncovered a hidden clue in people's language predictive of the later manifestation of psychosis: the frequent use of words associated with sound. A paper published by the journal npj Schizophrenia released the findings by scientists from Emory University and Harvard University. Hidden details The researchers developed a new machine-learning methodology to more precisely quantify the semantic richness of people's conversational language (a known indicator for psychosis). Their results indicated that automated analysis of the two language variables (more frequent use of words associated with sound and speaking with low semantic density, or vagueness) can predict if an at-risk person will later develop psychosis with an impressive 93 percent accuracy."*

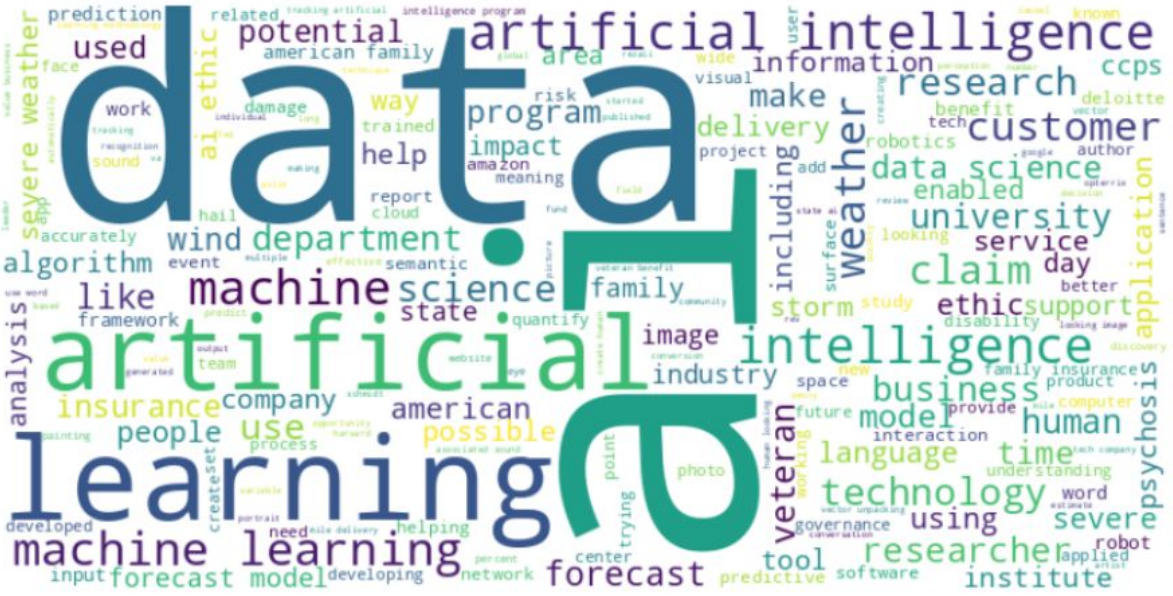
Results

- **TF-IDF Top matched words in N-Articles:**

	word	tfidf
47	ai	0.159174
256	data	0.122734
1102	weather	0.104149
571	learning	0.102854
603	machine	0.099461
1062	using	0.077359
82	artificial	0.076220
530	intelligence	0.076220
639	min	0.074468
601	lstm	0.074468
684	norris	0.074468
910	series	0.074468
977	stauffer	0.074468
1022	time	0.070330
608	make	0.069066
141	business	0.068558
250	customer	0.064225
778	prediction	0.063718
1013	technology	0.062085
406	forecast	0.060297
809	psychosis	0.058661
167	claim	0.058083
1059	use	0.055799
560	language	0.055678
733	people	0.049939
1111	wind	0.049334
863	researcher	0.044740
915	severe	0.043852
477	human	0.043560
69	application	0.043002

Results

- **Count Vectorizer Top matched words in N-Articles:**



ai	55
data	34
learning	25
artificial	25
artificial intelligence	25
intelligence	25
machine	24
machine learning	22
weather	19
claim	16
customer	16
technology	14
research	14
science	14
use	13
business	12
department	12
researcher	12
university	11
forecast	11
dtype: int64	