NLP on Scientific Articles for Information Retrieval

Xin Zhao

Problem statement

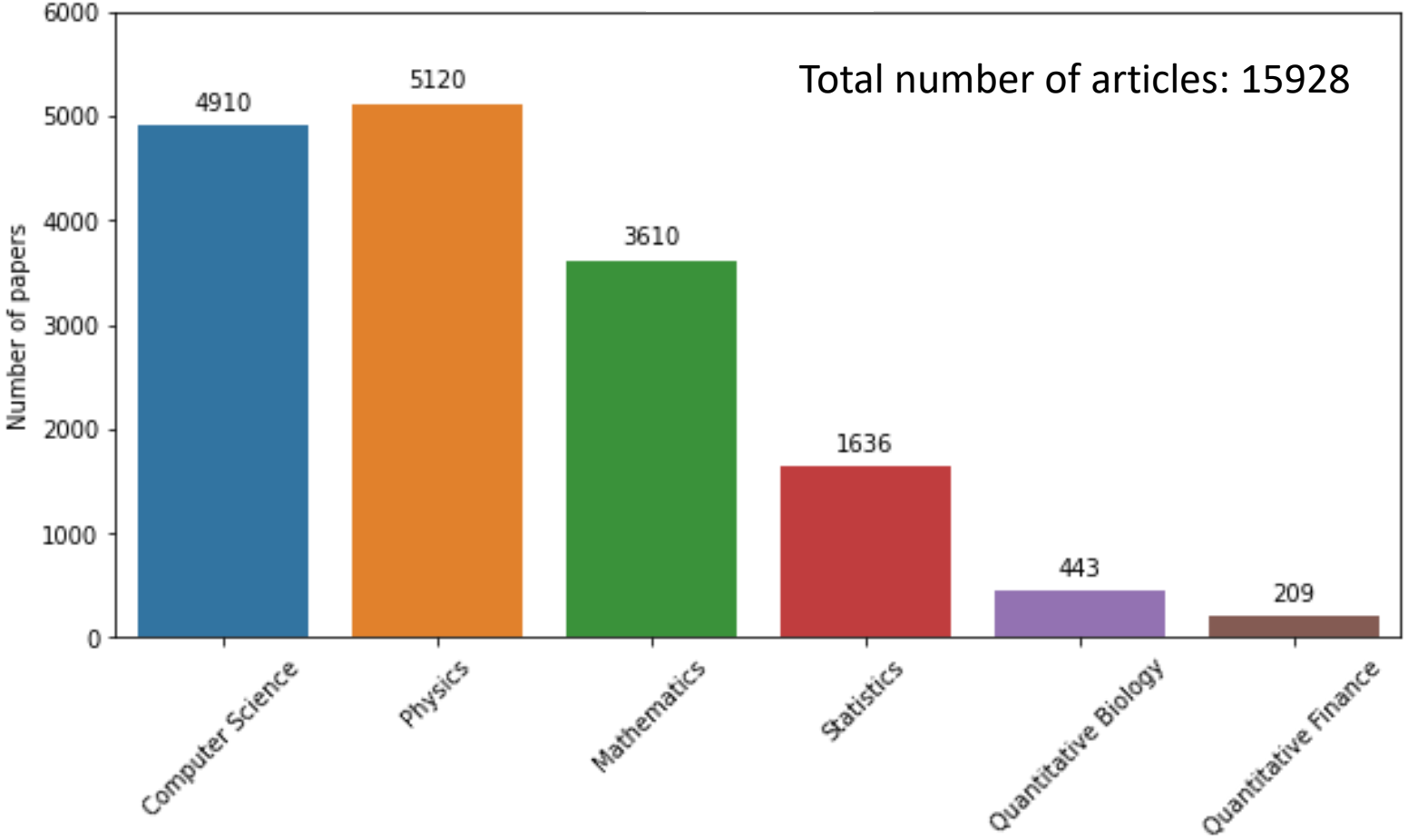
- How to facilitate recommendation and search process of scientific articles?
  - To give a token of identification to the articles

- Word embedding by NLP to give the tokens
  - NLP = natural language processing

Data

- Title and abstract for a set of research articles from Kaggle

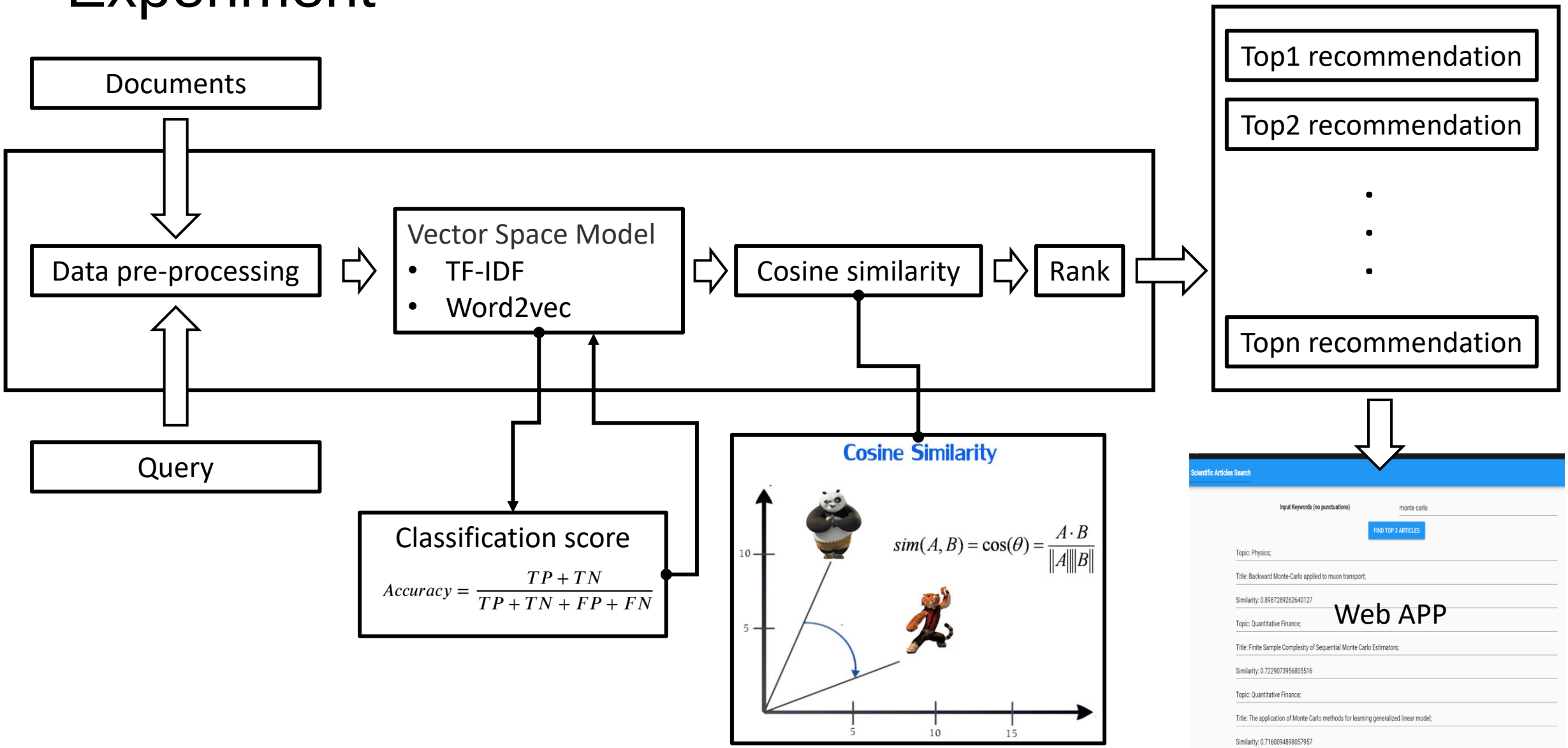
Topic



Example

Title: Rotation Invariance Neural Network
Abstract: Rotation invariance and translation invariance have great values in image recognition tasks. In this paper, we bring a new architecture in convolutional neural network (CNN) named cyclic convolutional layer to achieve rotation invariance in 2-D symbol recognition. We can also get the position and orientation of the 2-D symbol by the network to achieve detection purpose for multiple non-overlap target. Last but not least, this architecture can achieve one-shot learning in some cases using those invariance.

Experiment

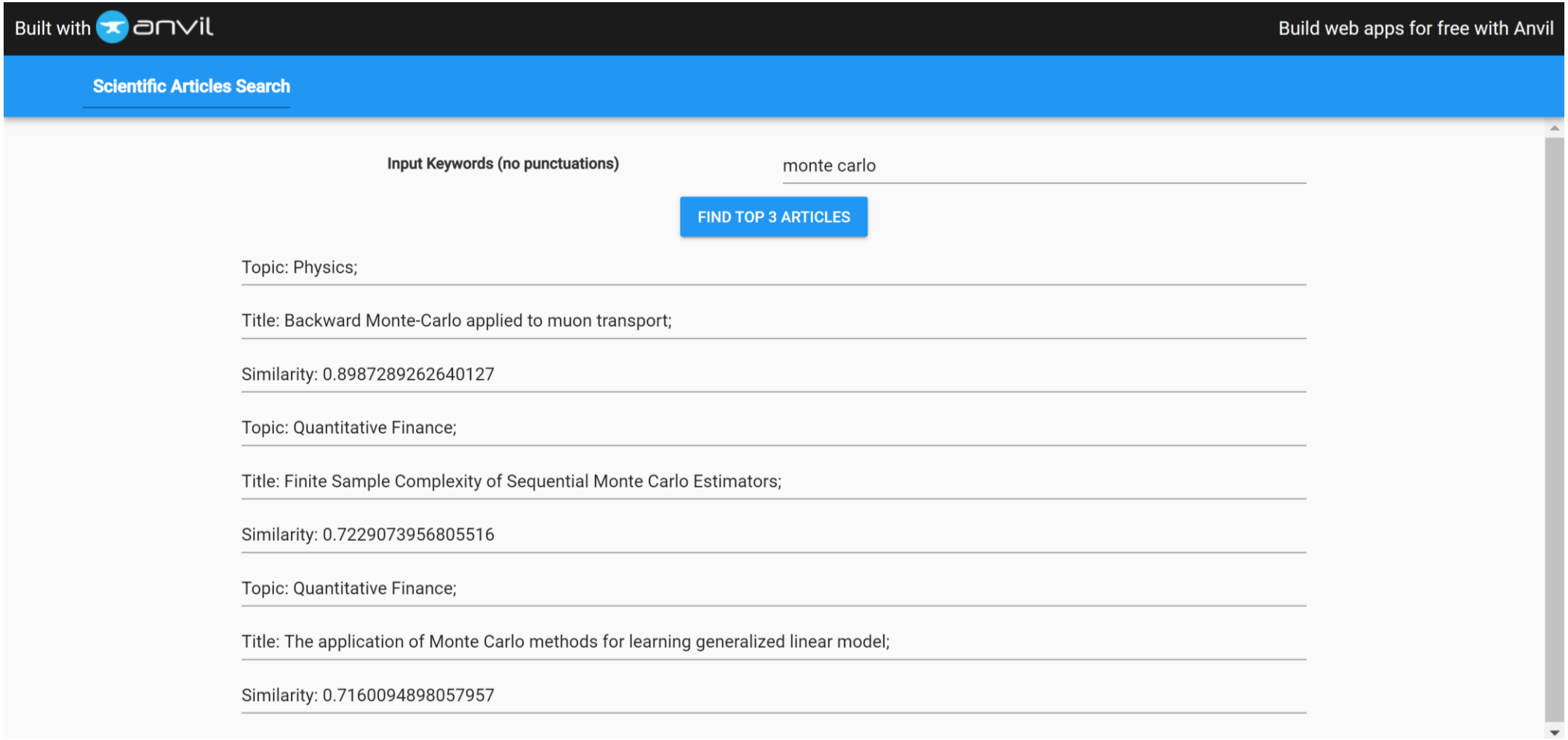

Documents
Data pre-processing
Query
Vector Space Model
• TF-IDF
• Word2vec
Cosine similarity
Rank
Top1 recommendation
Top2 recommendation
Topn recommendation
Classification score
$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Cosine Similarity
$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$
10
5
5
10
15
Scientific Articles Search
Input Keywords (no punctuations)
monte carlo
FIND TOP 3 ARTICLES
Topic: Physics;
Title: Backward Monte-Carlo applied to muon transport;
Similarity: 0.8987289262640127
Topic: Quantitative Finance;
Title: Finite Sample Complexity of Sequential Monte Carlo Estimators;
Similarity: 0.7229073956805516
Topic: Quantitative Finance;
Title: The application of Monte Carlo methods for learning generalized linear model;
Similarity: 0.7160094898057957
Web APP

Word space model selection

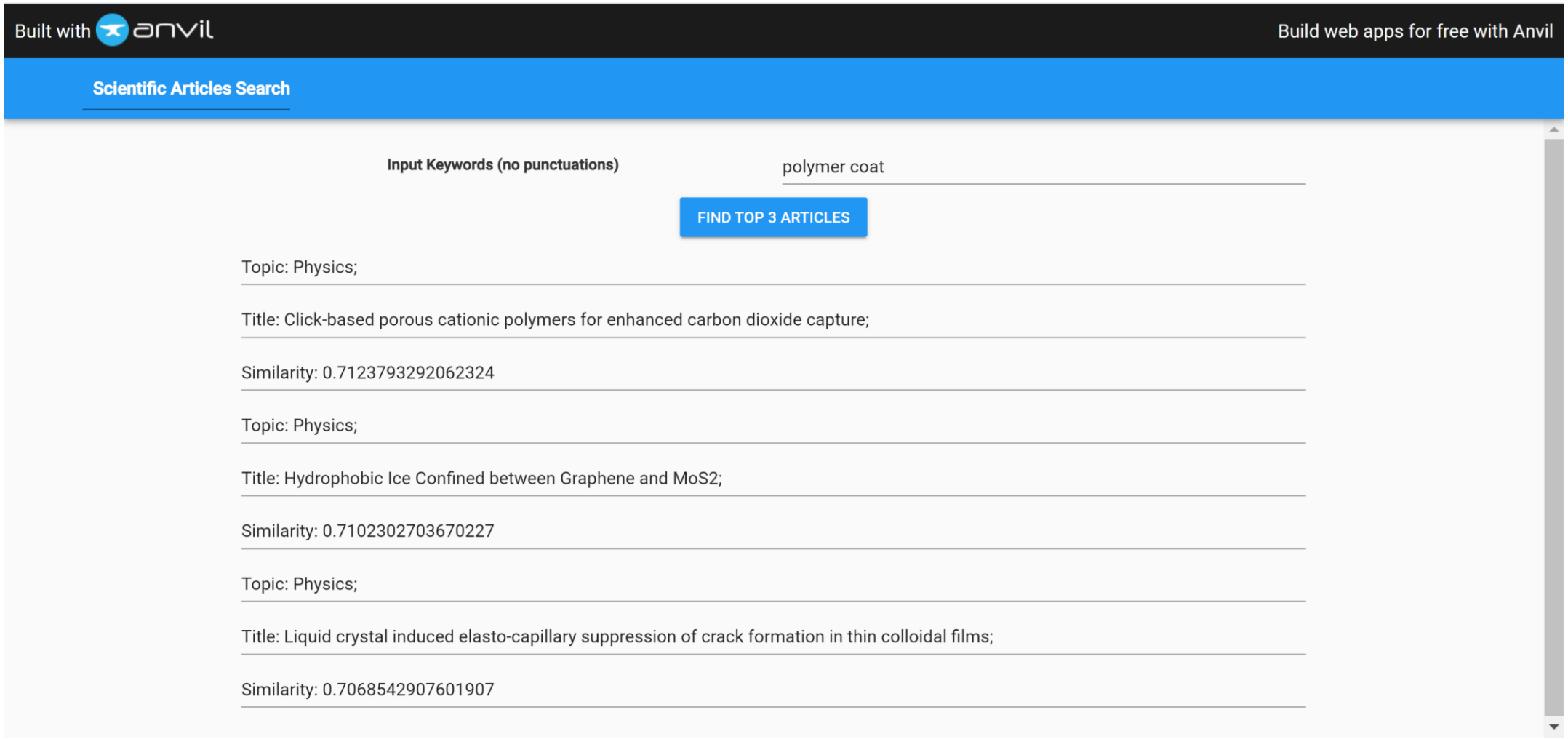
	TF-IDF	word2vec
Feature number	16293	300
Train set score	0.92	0.86
Test set score	0.84	0.84

Model complexity
Simpler vectors
Less computing cost

Web APP https://scientific-articles-search.anvil.app/



Reasonable recommendations



Conclusions

- Word vectorization using the NLP algorithm of word2vec
  - Classification accuracy score = **0.84**

- A product to realize article search
  - Word vectorization + cosine similarity ranking
  - Implementing to database of scientific articles

To improve the work

- Larger documents to improve the balance?

- Playing with weight between the title and abstract?