VECTORIZATION OF DOCUMENTS AND ANALYSIS OF THEIR IDENTITY USING A NEURAL NETWORK

Anton Rogozin
Ural Federal University

Marina Medvedeva
Ural Federal University

Vitaly Ford
Arcadia University
Purposes and objectives
The purpose of this article is to design a convenient and fast system for searching for similar documents.
Objectives

- Text processing.
- Studying the theory of word embedding.
- Designing an algorithm to determine the optimal model parameters.
- Model development.
- Application of the model in practice.
- Discussion of the results of the designed model.
Relevance
The user wrote a question

Living people are on the forum

Answers appear after a while
The user wrote a question

The system analyzes the text of the question

The system offers similar questions
Solution
Plan

- Text processing (tokenization, stemming, removal of stop words)
- Creating a model architecture based on the doc2vec neural network to select the optimal parameters
- Determining the Best Quality Metrics
- Implementation on the website
- Model training every n days
- Using the model on new data
Vocabulary: Man, woman, boy, girl, prince, princess, queen, king, monarch

word2vec. Word embedding
Doc2vec architecture
System architecture
Quality metric

Arithmetic average precision

\[
AP @ k(q) = \frac{\sum_{i=1}^{k} \text{Precision} @ k(q)}{k}
\]

\[
\text{Precision} @ k(q) = \frac{1}{k} \sum_{i=1}^{k} y(q, d_q^i)
\]
**numpy** is a library for convenient work with arrays.

**pandas** is library for working with data frames.

**sklearn** is library containing various classification, regression, clustering, downsizing algorithms using t-SNE.

**nltk** is a package of libraries and programs for symbolic and statistical processing of a natural language.

**pymystem3** is a library for lemmatizing tokens of the Russian language.

**scipy** is a library for performing scientific and engineering calculations.

**matplotlib** is a library for multidimensional data.

**genism** is a library containing the implementation of doc2vec.
Data sets
**Toy set**
- 70 documents
- Each document contains 3-5 words
- 3 classes
- 300 words

**Books**
- 700 documents
- Each document contains 100 thousand words
- 8 classes
- 75 million words

**News**
- 600 thousand documents
- Each document contains 300 words
- 20 classes
- 120 million words
Results
Toy set. Dependence of the error on the length of the vector
Toy set. Visualization graphs on the train and test data sets
Books. Dependence of the error on the length of the vector
Books. Visualization graphs on the train and test data sets
News. Dependence of the quality on the length of the vector
• all classes should be present in the training and test data sets.
• the algorithm does not work correctly on a small text corpus.
• train and test data sets should be large enough for the quality of the model to be good.
• vector representation allows to determine the hidden (latent) meaning of texts based on the occurrence of words with each other. This technique gave an accuracy of about 90% when searching for similar news.
• the error decreases sharply, reaches its global minimum, slowly grows or remains at the same level.