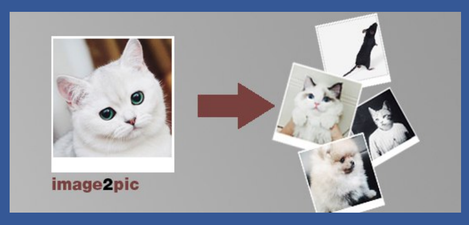
General approach to semantic search across media and texts

Menshikh Ivan && Gennady Shtekh

Nonexistent Chair Of Our Dream, Ural Federal University



github.com/
menshikh-iv/image2pic

Abstract

We provide easy way to build a multimodal search in common semantic space considering both text and media information.
The tricky part is to extract descreete annotation from media. In most cases(images and audio) it can be done by getting activations of pre-trained model from first fully connected layer after convolution.

The demo application scheme

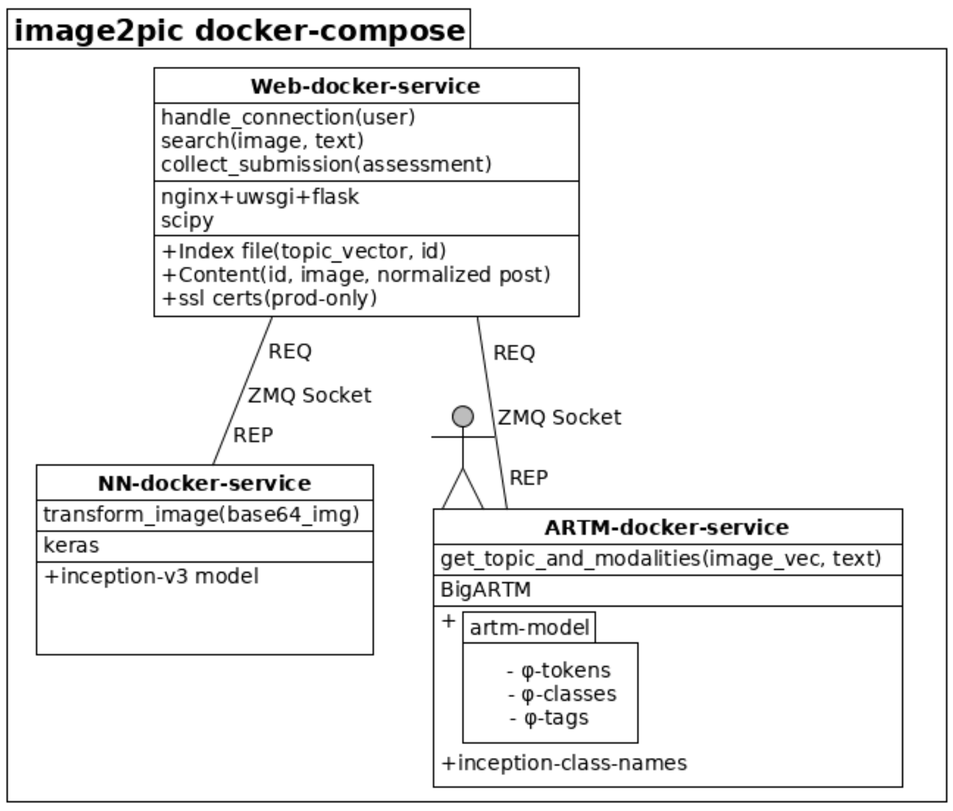
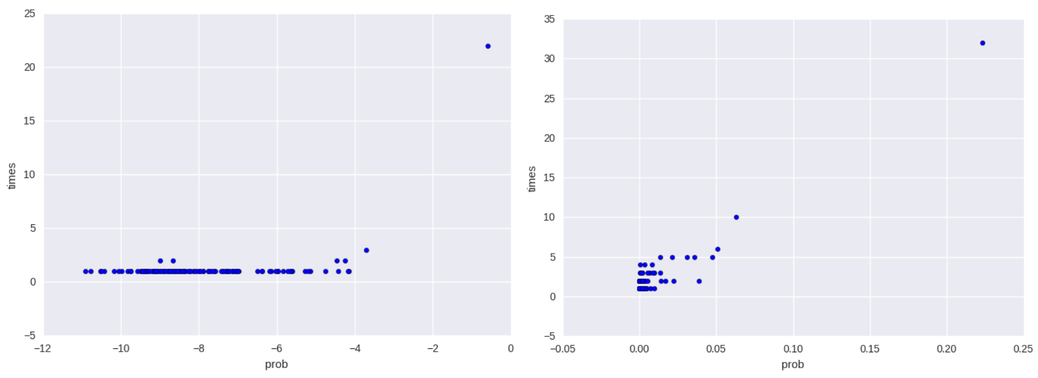


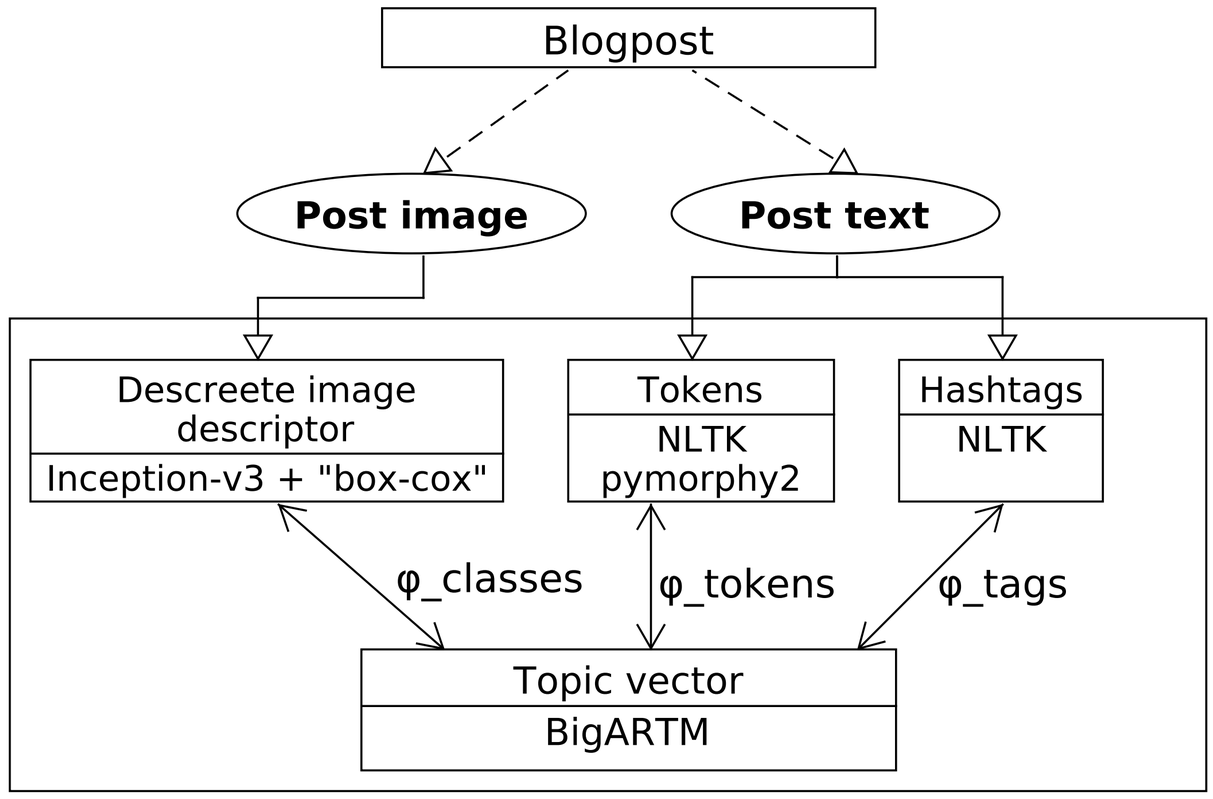
Image descriptors values distribution, before and after transformation



Conclusion

After gathering vector representation of search items it easy to leverage fuzzy vector-space indexer(annoy, faiss, NMSLib) to build scalable search app for your own purposes.
For long enougth text query we achieved **nDSG score = 0.37** for image search scenario.

Data Pipeline



BigARTM modalities conception *

Let D denote a finite set (collection) of texts and W^1 denote a finite set (vocabulary) of all terms from these texts. Each term can represent a single word or a key phrase. A document can contain not only words, but also terms of other modalities. Each modality is defined by a finite set (vocabulary) of terms

$$W^m, m = 1, \dots, M$$

Examples of not-word modalities are: authors, class or category labels, date-time stamps, references to/from other documents, entities mentioned in texts, objects found in the images associated with the documents, users that read or downloaded documents, advertising banners, etc.
Assume that each term occurrence in each document refers to some latent topic from a finite set of topics T. Text collection is considered to be a sample of triples (w_i, d_i, t_i), $i = 1, \dots, n$, drawn independently from a discrete distribution $p(w, d, t)$ over the finite space $W \times D \times T$, where $W = W^1 \sqcup \dots \sqcup W^m$ is a disjoint union of the vocabularies across all modalities. Terms w_i and documents d_i are observable variables, while topics t_i are latent variables.
To learn parameters Φ^m, Θ from the multimodal collection we maximize the log-likelihood for each m-th modality:

$$\mathscr{L}_m(\Phi^m, \Theta) = \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln p(w \mid d) \to \max_{\Phi^m, \Theta},$$

where n_{dw} is the number of occurrences of the term $w \in W^m$ in the document d. Note that topic distributions of documents Θ are common for all modalities.

* Vorontsov, Konstantin and Frei, Oleksandr and Apishev, Murat and Romov, Peter and Suvorova, Marina and Yanina, Anastasia, A Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections 2015