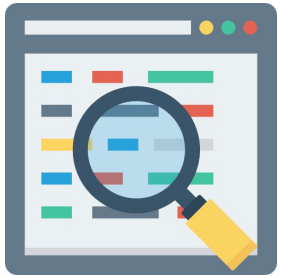
Keyword Extraction

Mehrdad Mohammadian

Contents

- What is the keyword extraction?
- Statistical approaches
- Graph-based approaches
- Machine learning approaches
- Topic modeling
- SBERT
- N-shot learning



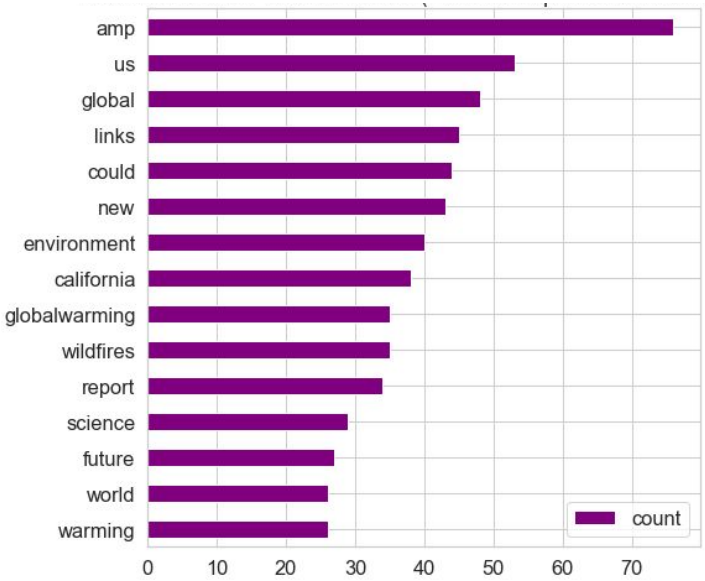
What is the Keyword Extraction?

Keywords describe the main topics expressed in a document

Statistical Approaches

1) Word Frequency

Listing the words and phrases that repeat the most within a text.



Advantages:

- Unsupervised
- Very easy to implement!

Disadvantages:

Missing valuable informations like:

- Meaning
- Grammer
- Synonyms
- Sequence of words
- etc ...

2) Word Collocation and Co-occurrence

Word collocations are words that frequently go together like “video calls”.

As knows as: N-grams

- Uni-gram (ex. book)
- Bi-grams (ex. machine learning)
- Tri-grams (ex. easy to use)

Co-occurrences are words that tend to co-occur in the same document that they don't necessarily have to be adjacent.

Advantages:

- Unsupervised
- Understand the semantic structure
- Count separate words as one

Disadvantages:

- Content lost: when n in small
- Sparsity problem: when n is big
  Denominator or Numerator can be 0.
- Storage problem

3) TF-IDF

Measures how important a word is to a document in a collection of documents.

TF - term frequency:
nd = number of times a word appears in a document
na = all words in document
nd / na

IDF - inverse document frequency:
nd = number of documents that have this word
na = number of all documents
log (na / nd)

TF-IDF = TF * IDF

Advantages:

- Unsupervised
- Useful for search engines

Disadvantages:

- Can not capture:
  Semantics similarities between words
  Co-occurrences
- Slow for large documents

4) RAKE

Rapid Automatic Keyword Extraction (RAKE)

Uses a list of stopwords and phrase delimiters to detect the most relevant words or phrases in a piece of text.

Advantages:

- Fast
- Unsupervised
- Low complexity
- Generate more complicated phrases
- Domain and language independent

Disadvantages:

- Comprehensive list of stop words
- Allow phrases have more weights
- Miss multi-words contains stopwords

5) YAKE

Yet Another Keyword Extractor

Keyword extraction method which rests on text statistical features extracted from single documents to select the most important keywords of a text.

Demo: http://yake.inesctec.pt/demo/usr

Advantages:

- Unsupervised
- Corpus-Independent
- Domain and language independent
- Good in multilingual

Graph-based Approaches

1) TextRank

Inspired from PageRank but for ranking texts.

we measure the relationship between two or more words.

Advantages:

- Unsupervised
- Language independent

---

2) SignleRank
3) ExpandRank
4) TopicRank

Machine Learning Approaches

1) CRF (Conditional Random Field)
2) LSTM (RNN)
3) Transformers (BERT, ALBERT & etc)

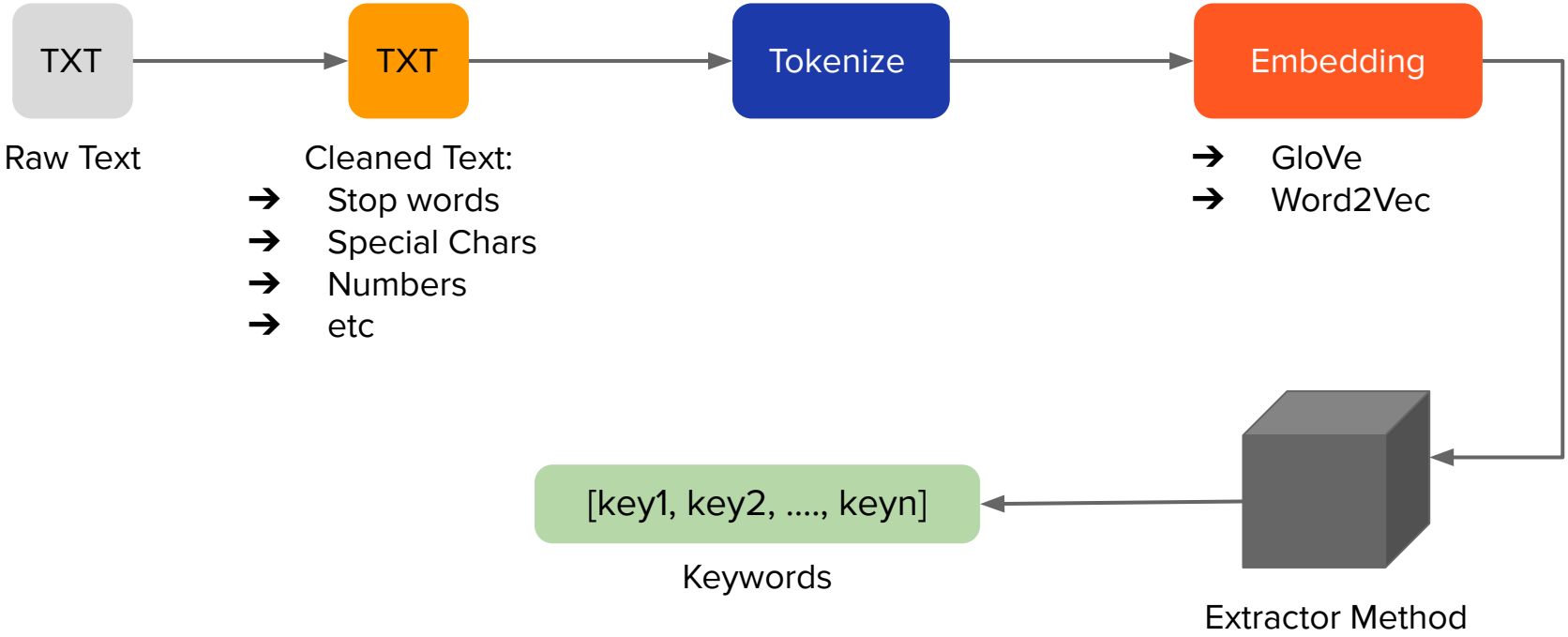
Advantages:

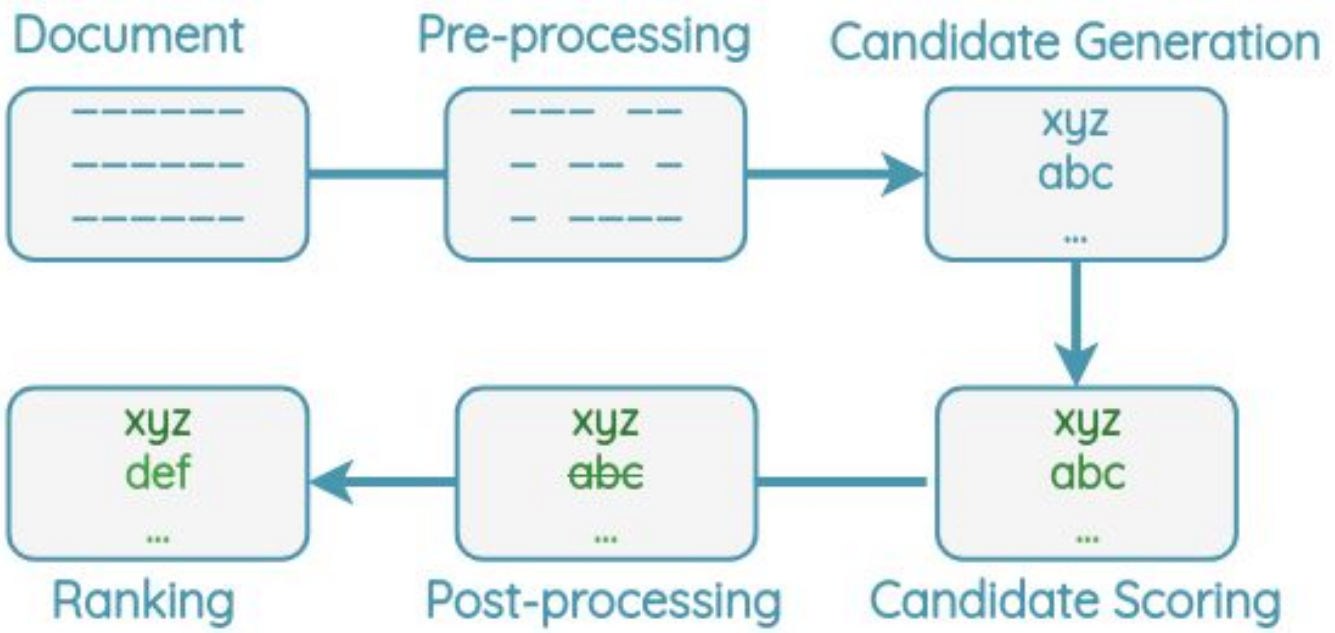
- Accurate
- Understand meaning

Disadvantages:

- Need maaaanyyyy data!
- Very sensitive in pre-processing steps

Py Text


TXT
TXT
Tokenize
Embedding
Raw Text
Cleaned Text:
➔ Stop words
➔ Special Chars
➔ Numbers
➔ etc
➔ GloVe
➔ Word2Vec
[key1, key2, …., keyn]
Keywords
Extractor Method



Zero-Shot Learning PosTagger to delete verbs

https://gist.github.com/mehrdad-dev/620baa214d887993635287bc55dab356

Tools

For Keyword Extraction

- NLTK
- Gensim
- PyTextRank
- PyTopicRank
- Scikit-learn
- YAKE
- KeyBERT
- Hugging Face

Ideas

- Summarization before keyword extraction
- Find tags just from title (ex. news)
- Tag prediction

Topic Modeling

What is the Topic Modeling?

A machine learning technique that automatically analyzes text data to determine cluster words for a set of documents

Advantages:

- Unsupervised

Disadvantages:

- Short-term solution

LSA

Latent Semantic Analysis

LSA takes text documents and creates *n* different parts where each part expresses a different meaning in the text.

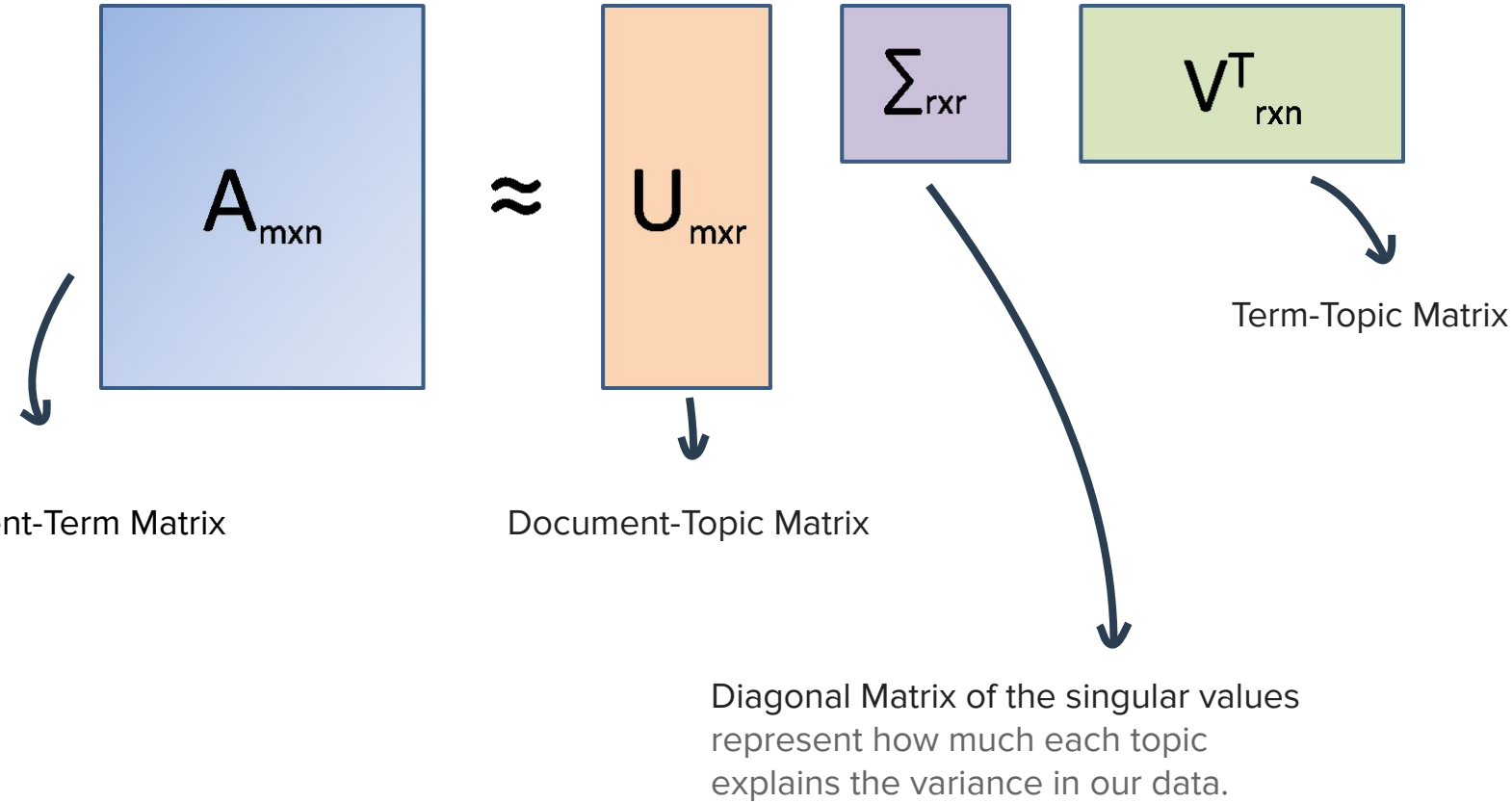
LSA reduces our table of data to a table of latent (hidden*)* concepts.

Advantages:

- Unsupervised

Disadvantages:

- Can’t capture multiple meanings of a word



r = number of topics

Document-Term Matrix	Document 1	Document 2	Document 2	Document 2	
Lebron	0.4	0	0	0	...
Senate	0.01	0.9	0	0	...
Celtics	0.2	0	0	0	...
Sprain	0	0	0.2	0.2	...
Cancer	0	0.02	0.3	0.3	...
...	...	...	...	...	...

The numbers in the table reflect how important that word is in the document. Can be computed by TF-IDF.

SVD

Document-Topic	T1	T2	T3	T4	
D1	0.8	0.2	0	0	...
D2	0	0.7	0	0	...
D3	0.1	0	0	0	...
D4	0.6	0	0.2	0.2	...
...	...	...	...	...	...

Term-Topics	T1	T2	T3	
Lebron	0.8	0.1	0.1	...
Senate	0	0.7	0.3	...
Celtics	0.9	0.1	0	...
Sprain	0.6	0	0.4	...
Cancer	0		0.7	...
...	...	...	...	...

LDA

Latent Dirichlet Allocation

The goal of LDA is to map all the documents to the topics

Each document can be described by a distribution of topics and each topic can be described by a distribution of words.

LDA assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities

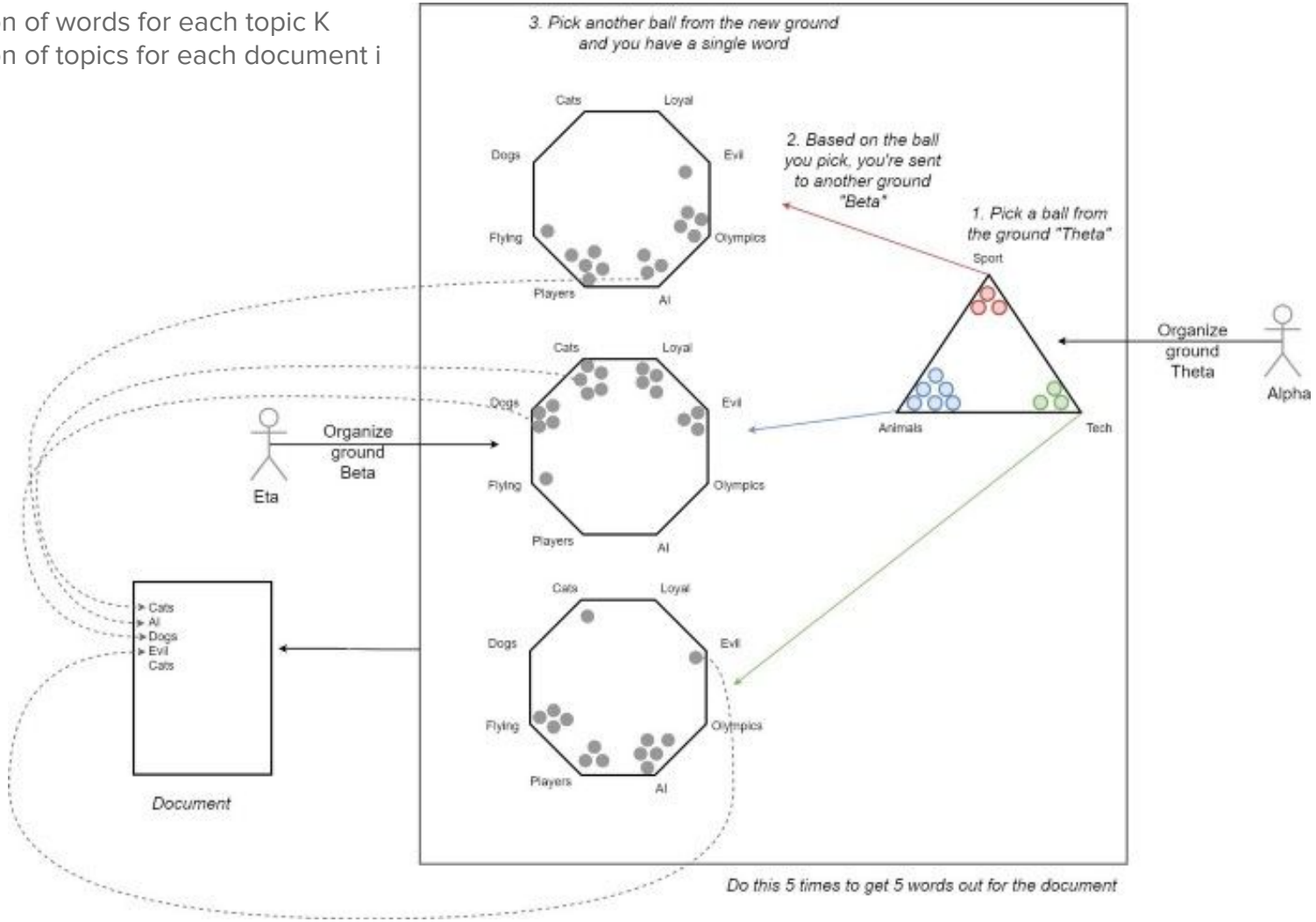
Advantages:

- Unsupervised

Disadvantages:

- Fixed K

the distribution of words for each topic K
the distribution of topics for each document i



BERT

SBERT Library

this version of BERT is specifically designed for tasks like semantic similarity search and clustering

Using Pretrained Model:

`distiluse-base-multilingual-cased-v2`

https://www.sbert.net/docs/pretrained_models.html

N-Shot Learning

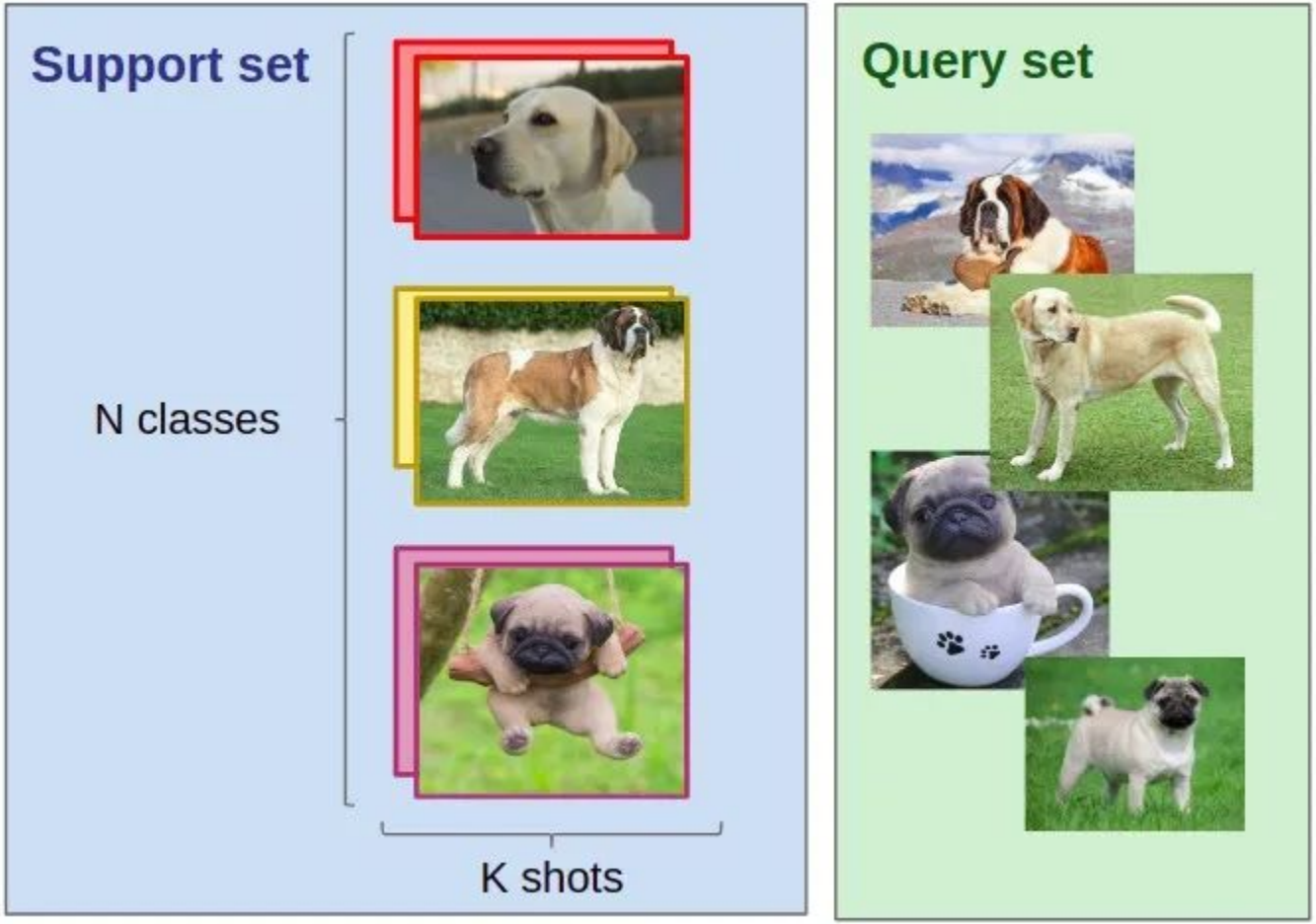
Zero-Shot
One-Shot
Few-Shot

Find similarity function
Using concept of learn to learn.
Support set instead of training set.

We can use hugging face
zero-shot-classification pipeline.

Advantages:

- Need few data samples


Support set
N classes
K shots
Query set

Query:

sim = 0.2	sim = 0.9	sim = 0.7	sim = 0.5	sim = 0.3	sim = 0.4