Steps

1. Indexing documents

Purpose: a document -> a vector of features

d\_i is represented as a d-dimensional vector (f\_1, f\_2, …, f\_d)

1. Applying dimension reduction (optional - depending on the data)

Purpose: reduce number of dimensions required to represent a document

d-dimensional vector -> k-dimensional vector (k<<d)

1. Applying learning methods
   1. Cost function-based classifiers: e.g., SVM
   2. Probability-based: e.g., Bayes, Mixtured models, etc
   3. Heuristic-based: e.g., Decision tree, random tree forest
   4. Neural network-based: e.g., feed forward network, or LSTM neural network (without step 1,2, ok)

First step first!

Indexing documents

Basic techniques:

* Classical (from 1950 (: ): tf-idf

Reference:

<http://www.math.unipd.it/~aiolli/corsi/0708/IR/Lez08.pdf>

<https://arxiv.org/pdf/cs/0110053.pdf> (page 12, 13, 14)

* Controlled vocabulary: Darmstadt Indexing Approach

Reference: <https://arxiv.org/pdf/cs/0110053.pdf> (page 14, 15)

* Modern: Word2Vec, Word Embedding (however, there may be some limitations of Vietnamese support libraries)