Document Classification Based on Word Vectors

Liu Rong
CSLT RIIT TSINGHUA
2014-09-13

Outline

- Document Classification
  - Introduction
  - Approach
- Document Vector(Text Representation)
  - LDA
  - **Word2Vec**
- Experiment
- Conclusion
- References

Document Classification-Introduction

- Introduction
  - Task

    to classify documents into predefined classes
  - Relevant Technologies

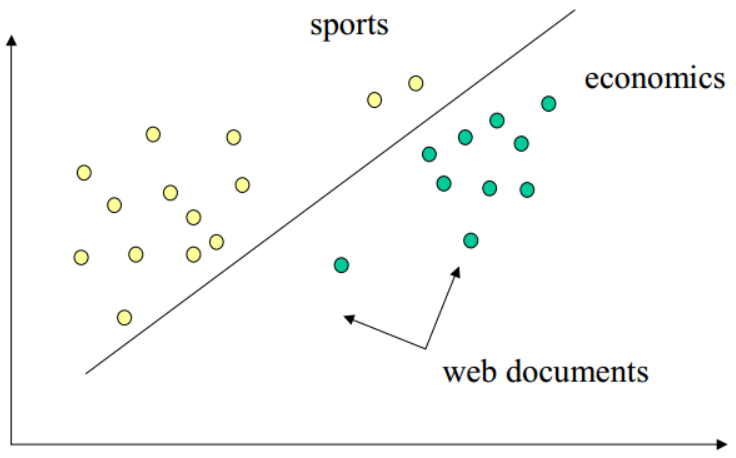
    Text Clustering, Information retrieval,

    Information filtering , Information Extraction.
  - Application

    QA, Categorize newspaper articles and newswires into topics.

    Organize Web pages into hierarchical categories.

    Sort journals and abstracts by subject categories



Document Classification-Introduction

- Approaches
  - Rule-based

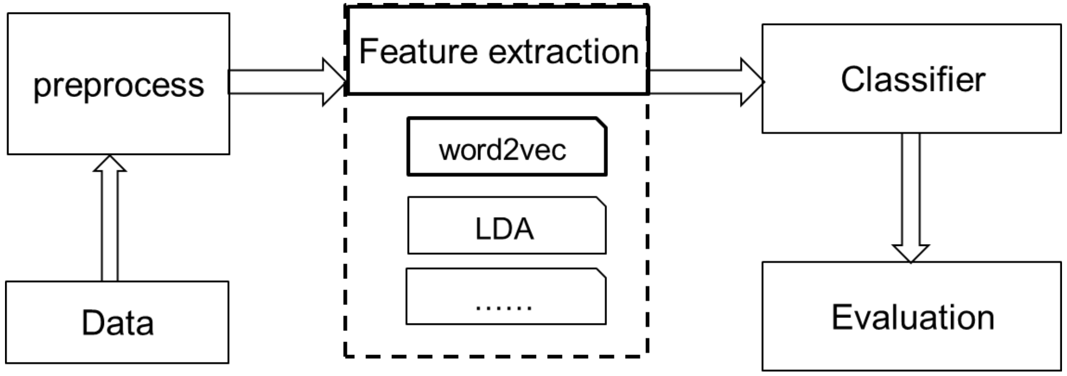
    Rule 1: $\text{"ball"} \in \mathrm{d} \rightarrow t(d) = sports$

    Rule 2: $\text{"ball"} \in \mathrm{d} \,\&\, \text{"dance"} \notin \mathrm{d} \,\&\, \text{game} \in d \,\&\, \text{"}play\text{"} \in d \rightarrow t(d) = sports$
  - Machine learning-based
    - Text preprocessing

      removing stop word and predefined words
    - **Feature Extraction**

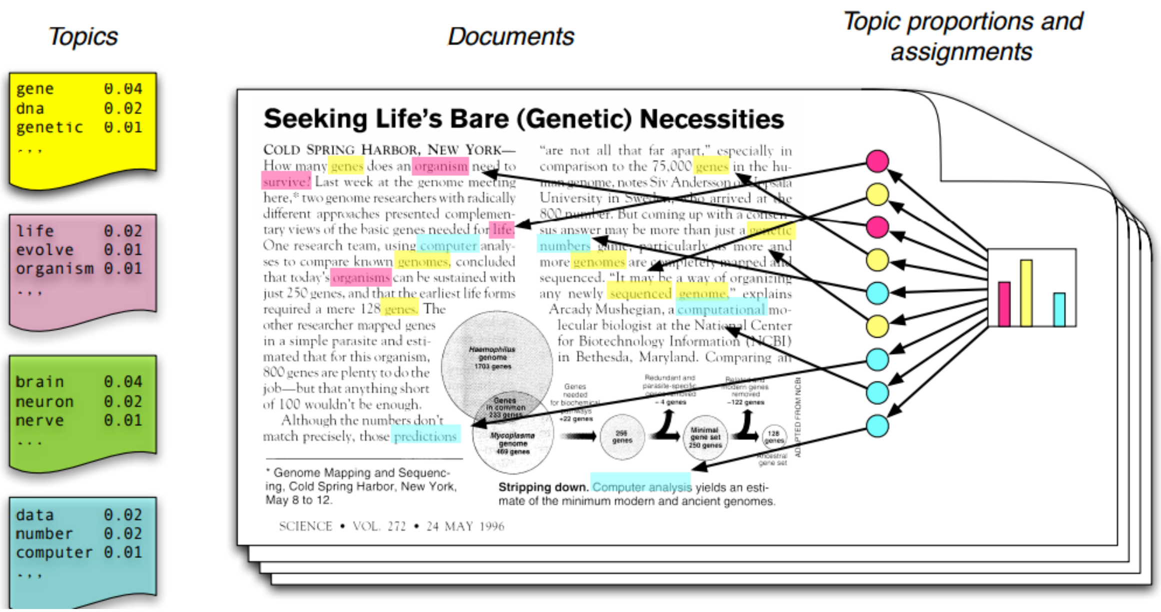
      TF-IDF(Bag-of-word), LDA, LSI, **word2vec**
    - Classifier Construction

      Native Bayes , KNN , SVM
    - Classifier Evaluation



Document Classification-LDA

- Introduction

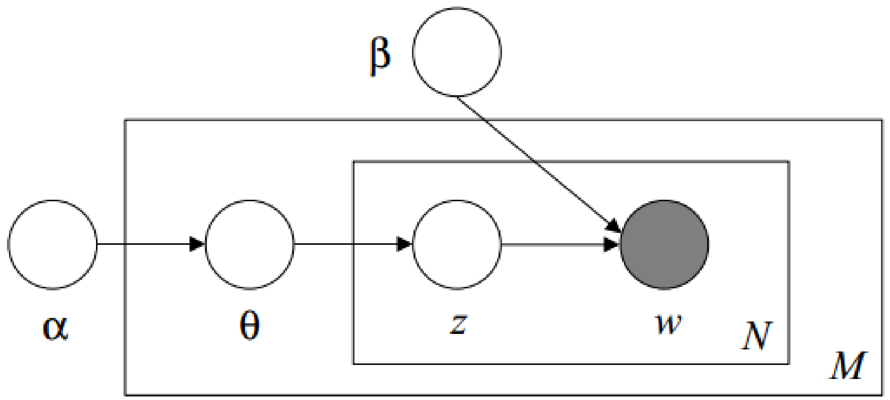


- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Document Classification-LDA

- Model
  - Topic Function

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1},$$



  - Document Function

$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta),$$

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta.$$

  - Corpus Function

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d.$$

The goal of training is to get the α and β when the corpus function get the maximum value.

  - Predict

$$p(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta) = \frac{p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)}{p(\mathbf{w}|\alpha,\beta)}.$$

Document Classification-LDA

- Document Vector
  - The topic distribution in a document

    $$document\ vector = \ \theta = [\mathrm{T}_1, \mathrm{T}_2 \cdots \mathrm{T}_K]$$

    where T_k $is\ the\ problity\ of\ k_{ih} topic\ in\ a\ document$

- Problem
  - Learning structure is uncertain
  - LDA is sensitive with initial value
  - High computational complexity
  - loss semantic information that lda don’t consider the word sequence

Document Classification-w2v

- One-hot Representation

  dog => [0 0 0 0 1 0 0 0 0 0 0]

  cat => [1 0 0 0 0 0 0 0 0 0 0]

- Distributed Representation

  dog => [0.792 -0.177 0.98 -0.9 …..]

  cat => [0.76 0.12 -0.54 0.9 0.65 ….]

- Method

  NNLM:

  C&W:

  M&H: Log-Bilinear /Hierarchical Log-Bilinear Model

  RNNLM:

  Huang: add document information

  Glove:

Document Classification-w2v

- Google Word2vec
  - Skip-gram

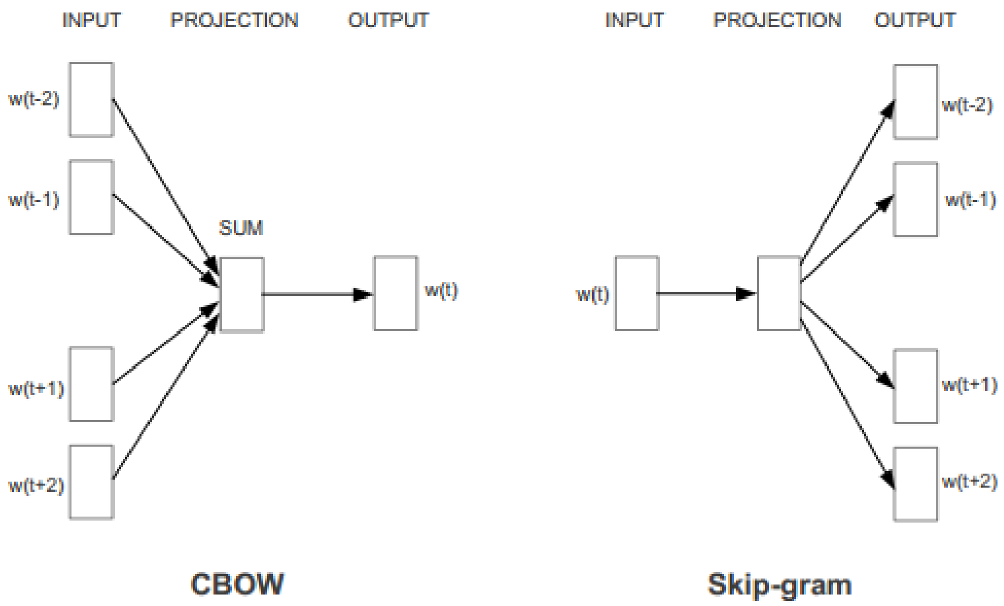
$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

where $w_1, w_2 \cdots w_T$ is sequence words ,c is size of contenx.

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^\top v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^\top v_{w_I}\right)}$$

where v_w and v'_w are the input and output vector representation.

w is the number of words in the vocabulary.

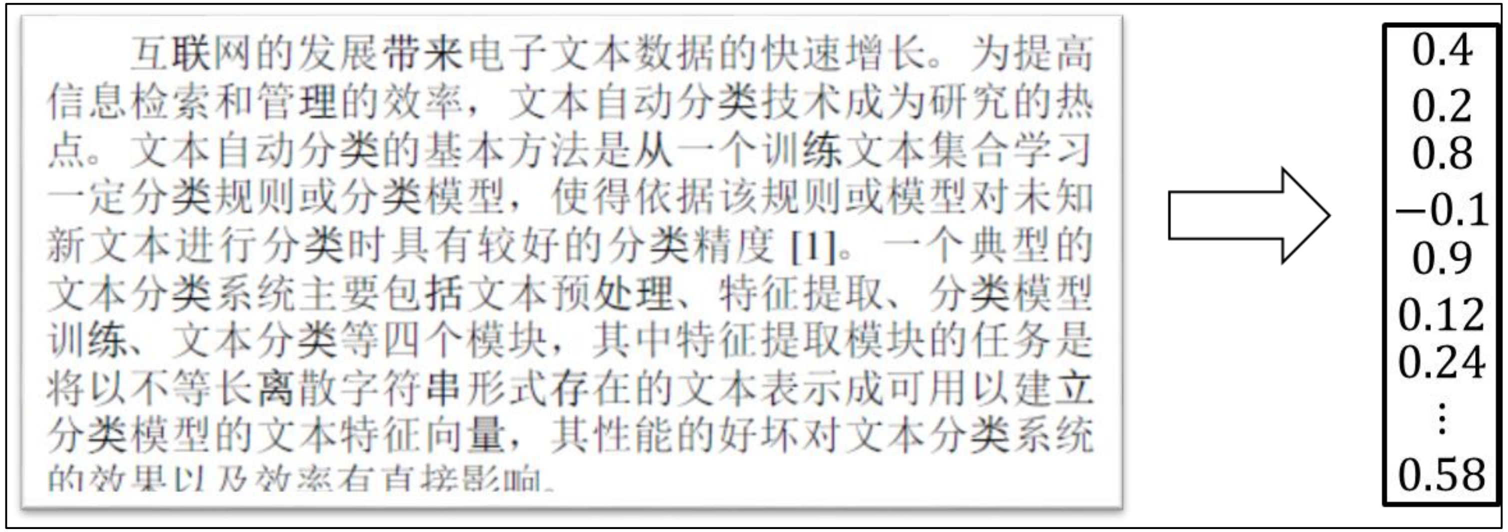


Document Classification-w2v

- Document Vector

$$document\ vector = \frac{1}{|d|}\sum_{w \in d} c_w$$

where |d| is the number of words in the document. c_w is the word vector of w.



Document Classification-Experiment

- Data

  SogouLab :

  1. car, economics, IT, health, sports, travel, education, Recruitment , culture and military
  2. train:14301(65M),test:1809

  w2v: train word vector on People's Daily(5G)

  w2v-ex: train word vector on train data of SougouLab(65M)

- Tool

  segment word : http://www.xunsearch.com/scws/index.php

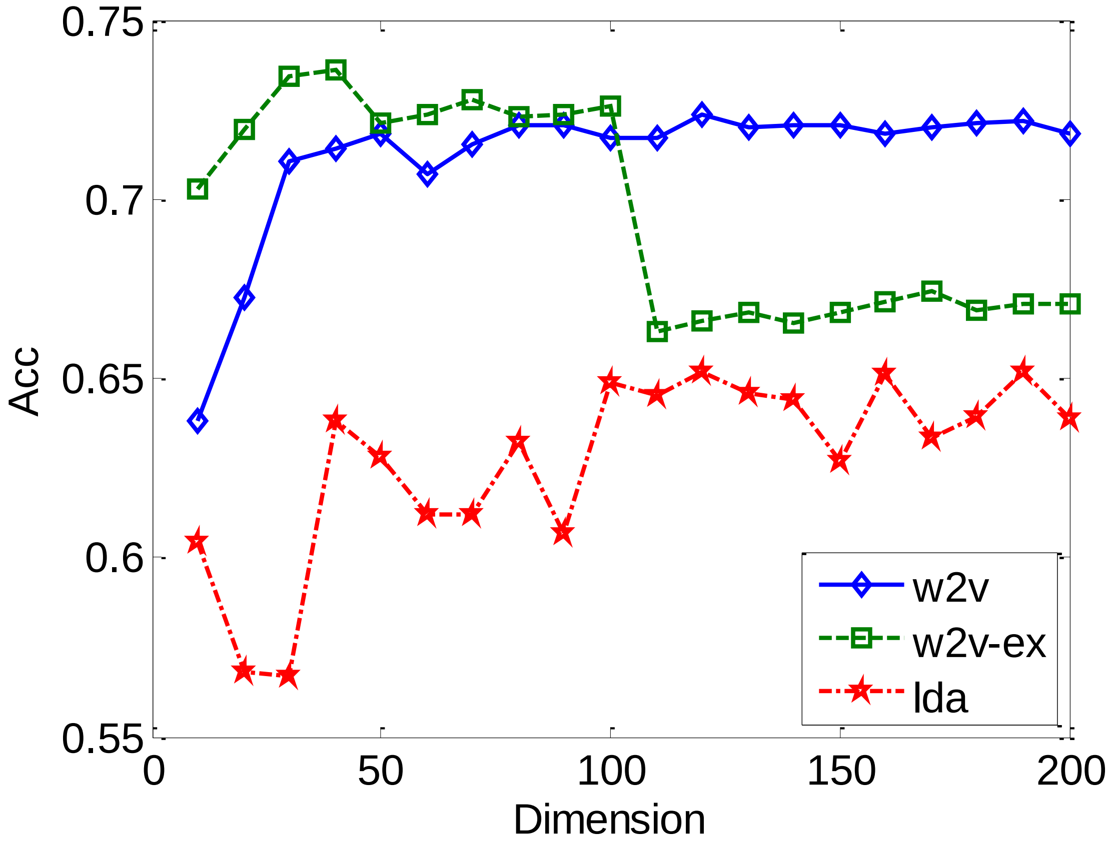
  word2vec: https://code.google.com/p/word2vec

  classifier/weka: http://www.cs.waikato.ac.nz/ml/weka

  LDA: http://www.cs.princeton.edu/ blei/lda-c

Document Classification-Experiment

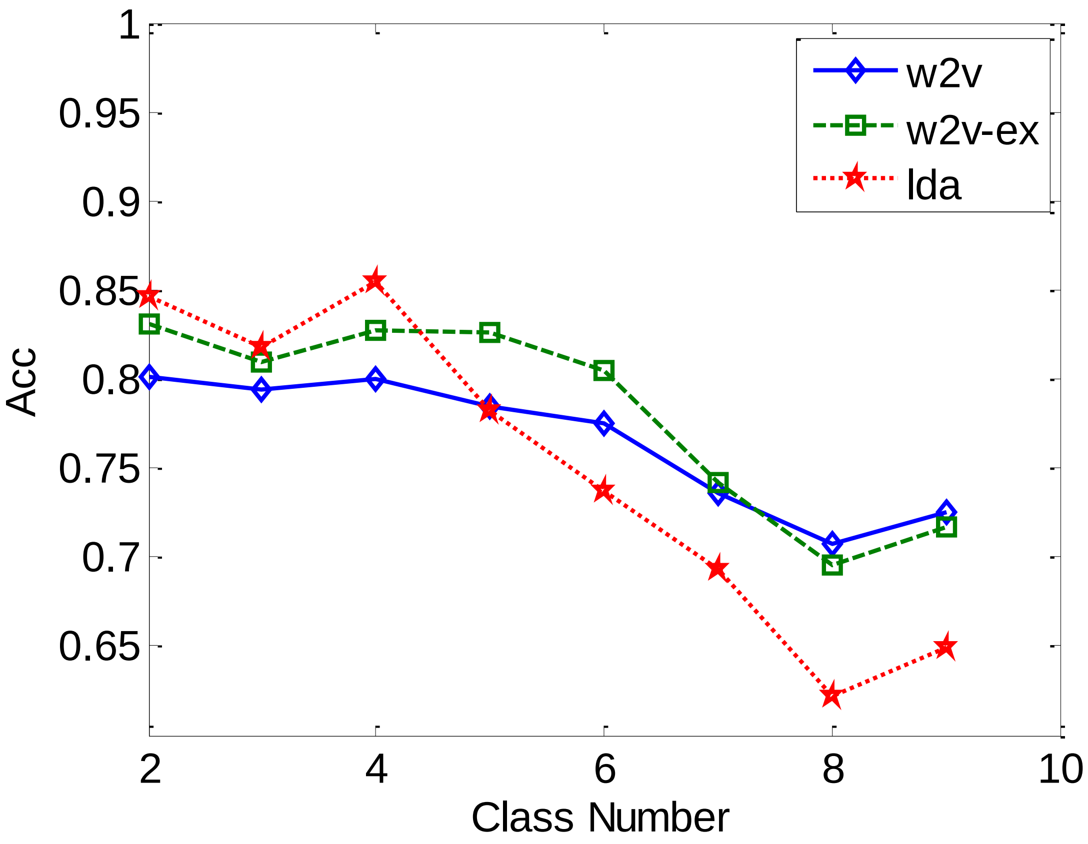
- Different dimensions of LDA and w2v
  - The w2v get higher accuracy than LDA
  - The w2v is more stable than LDA
  - The w2v need more data with higher dimension



w2v: train word vector on People's Daily(5G)
w2v-ex: train word vector on train data of SougouLab(65M)
lda: train lda on train data of SougouLab

Document Classification-Experiment

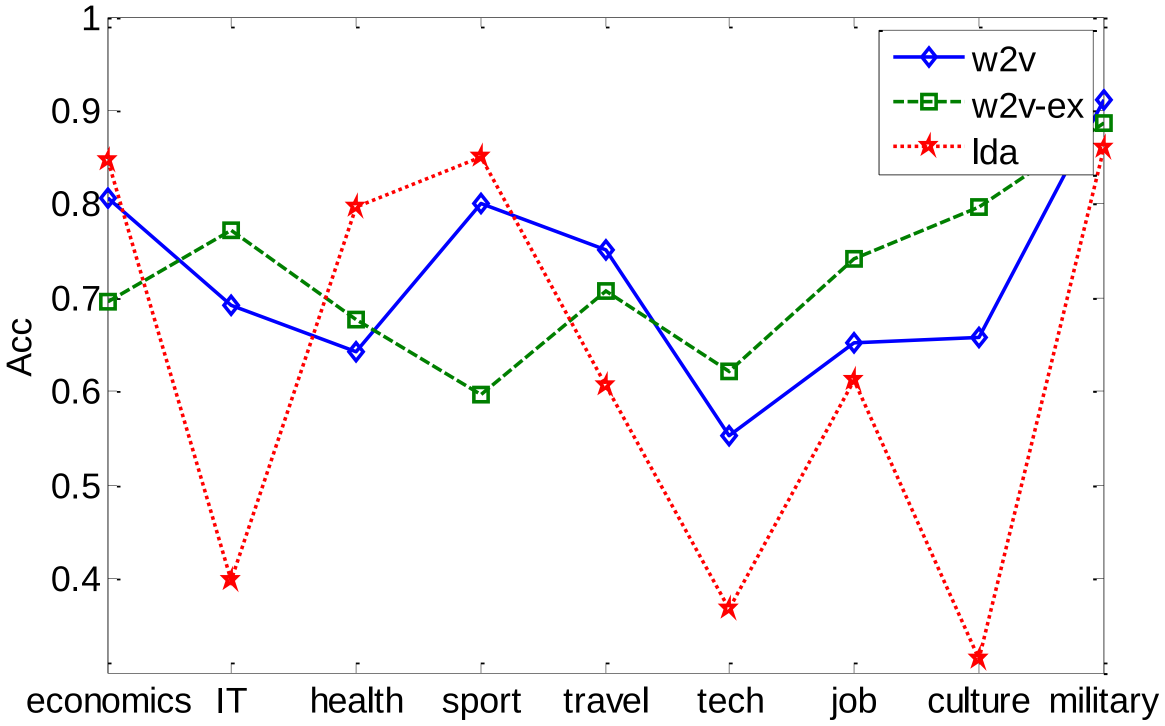
- Different classification task
  - w2v is equal with LDA from 2-classes to 4-classes
  - w2v get higher accuracy from 5-classes to 9-classes
  - w2v is more general



w2v: train word vector on People's Daily(5G)
w2v-ex: train word vector on train data of SougouLab
lda: train lda on train data of SougouLab

Document Classification-Experiment

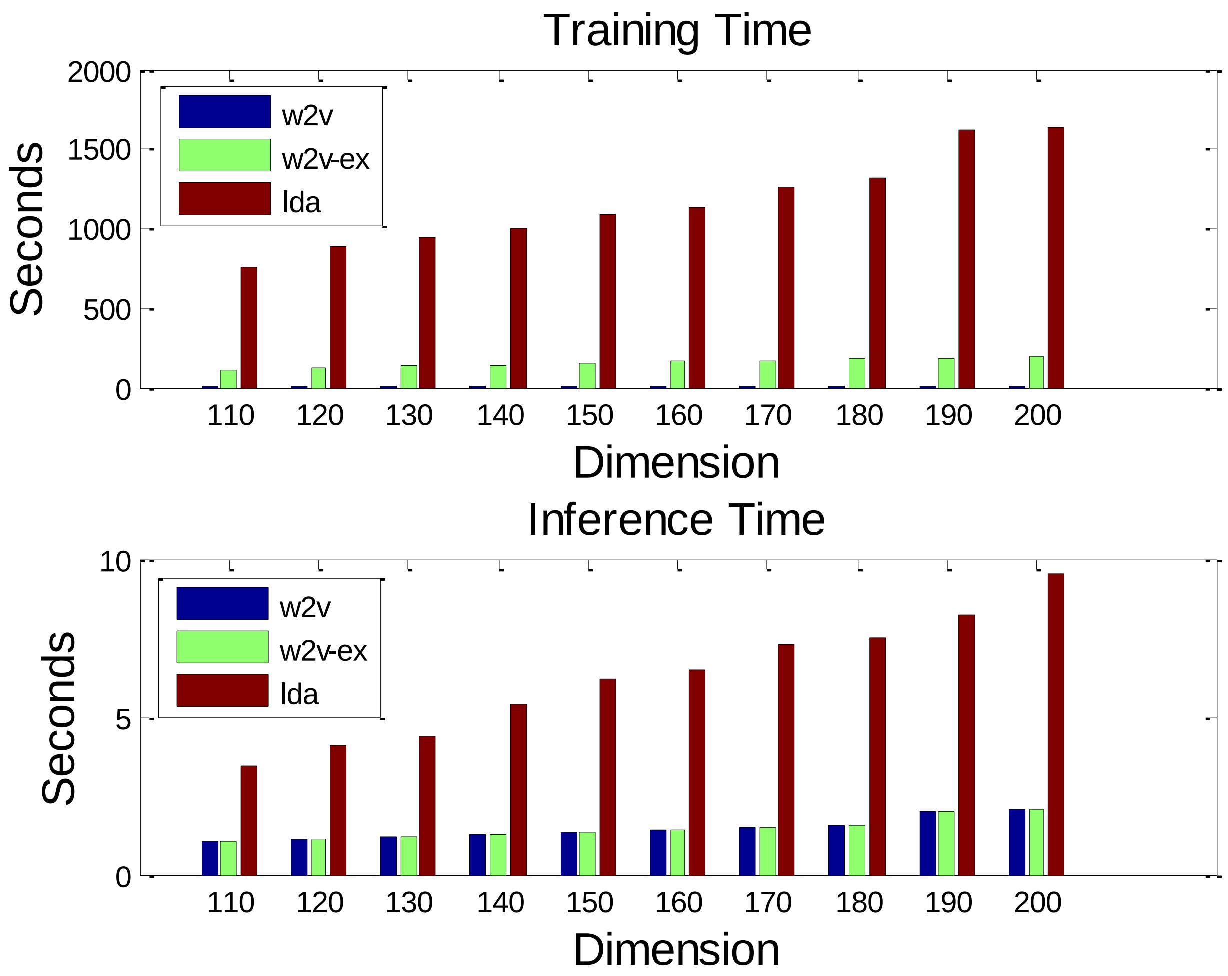
- Different classes Accuracy
  - w2v is more stable



w2v: train word vector on People's Daily(5G)
w2v-ex: train word vector on train data of SougouLab
lda: train lda on train data of SougouLab

Document Classification-Experiment

- Efficiency



Document Classification-Conclusion

- Conclusion

  • Introduce the word vector to document classification and analysis the different of semantic generation between word vector and LDA.

  • Experiment show that document classification based on word vector superior to LDA in classifier accuracy, computational complexity , scalability field, processing capacity
 in complex classification task and representation of content.

Document Classification-Reference

- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases andtheir Compositionality. In Proceedings of NIPS, 2013.
- G. Salton, A. Wong, and C.-S. Yang, “A vector space model for automatic indexing,” Communications of the ACM, vol. 18, no.11, pp. 613–620, 1975
- D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,”the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.
- T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprintarXiv:1301.3781, 2013.

Document Classification-QA

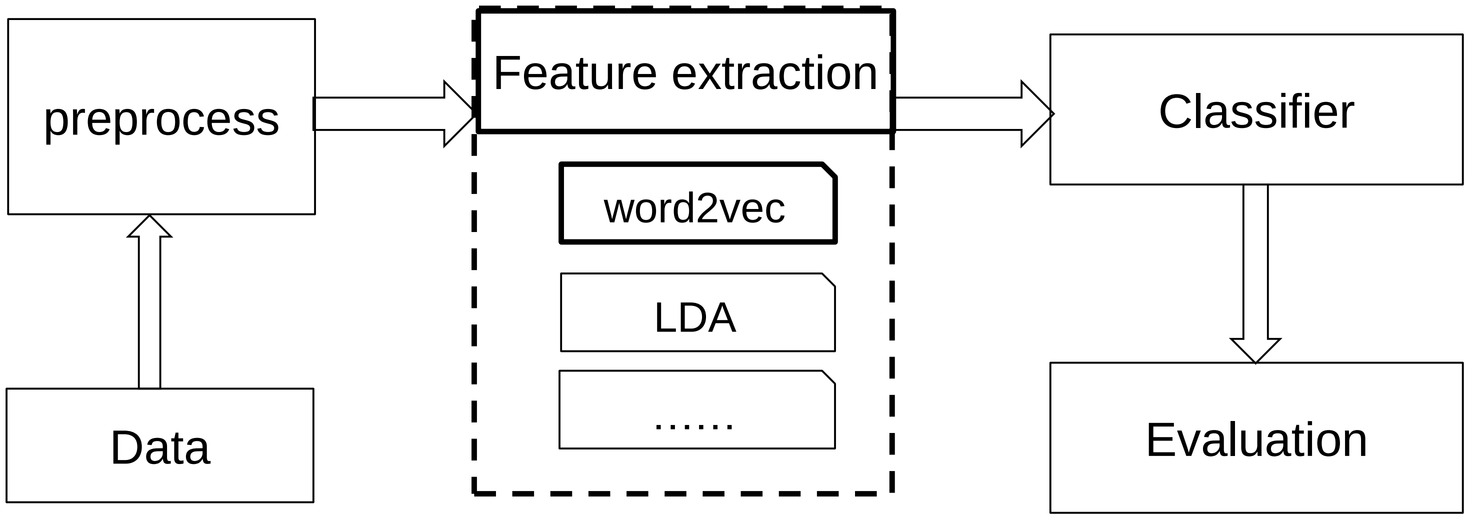
Question and answer

Document Classification-VSM

- Introduction
  - Document Vector

$$document\ vector = [\mathrm{T}_1, \mathrm{T}_2 \cdots \mathrm{T}_K]$$

where $T_j = n \times \log(\frac{M}{m})$, n is TF, M/m is IDF.


preprocess
Feature extraction
word2vec
LDA
……
Classifier
Data
Evaluation

Document Classification-Introduction

- Approaches
  - Rule-based

    Rule 1: $\text{"ball"} \in \mathrm{d} \rightarrow t(d) = sports$

    Rule 2: $\text{"ball"} \in \mathrm{d}\ \&\ \text{"dance"} \notin \mathrm{d}\ \&\ \text{game} \in d\ \&\ \text{"}play\text{"} \in d \rightarrow t(d) = sports$
  - Machine learning-based
    - Text preprocessing

      removing stop word and predefined words
    - **Feature Extraction**

      TF-IDF(Bag-of-word), LDA, LSI, **word2vec**
    - Classifier Construction

      Native Bayes , KNN , SVM
    - Classifier Evaluation