Document Classification: 20 Newsgroup

INT3406 21 - Group 3

Group 3
Pham Truong Giang - 1802xxxx
Nguyen Manh Dung - 18020370
Nguyen Phuc Hai - 1802xxxx
Le Bang Giang - 1802xxxx

Abstract

Your abstract should motivate the problem, describe your goals, and highlight your main findings. Given that your project is still in progress, it is okay if your findings are what you are still working on.

1 Introduction

1.1 abstract

1.2 abstract

2 Preprocessing

2.1 Lowercasing

2.2 Map different word to canonical form

2.2.1 Stemming

2.2.2 Lemmatization

2.3 Stop word

2.4 Noise removal

2.5 Summary

3 Feature Extraction

3.1 Bag of Word (BoW)

3.1.1 Brief Explaination

3.1.2 Algorithm

3.1.3 Implementation

3.1.4 Discussion

3.2 Term Frequency – Inverse Document Frequency (TF – IDF)

3.2.1 Brief Explaination

3.2.2 Algorithm

3.2.3 Implementation

3.2.4 Discussion

3.3 Word Embedding

3.3.1 Brief Explaination

3.3.2 Algorithm

3.3.3 Implementation

3.3.4 Discussion

4 Classification

- 4.1 Linear Model
  - 4.1.1 Naive Bayes
  - 4.1.2 Logistic Regression (LR)
  - 4.1.3 Ridge Classification
  - 4.1.4 Perceptron
  - 4.1.5 Passive-Aggressive
- 4.2 Non-parametric
  - 4.2.1 K-nearest neighbor (KNN)
  - 4.2.2 Support Vector Machine (SVM)
  - 4.2.3 Linear Support Vector Machine (LinearSVC)
- 4.3 Tree-based Classifiers
  - 4.3.1 K-nearest neighbor (KNN)
  - 4.3.2 Support Vector Machine (SVM)
  - 4.3.3 Linear Support Vector Machine (LinearSVC)
- 4.4 Graphical Classification
  - 4.4.1 Conditional Random Fields (CRFs)
  - 4.4.2 ...
- 4.5 Neural Network

5 Summary

Abstract

References