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

Data preprocessing is an essential step in building Machine Learning models. In natural language processing (NLP), text preprocessing can simply be understood as the process to transform raw text data into a form that is ***predictable*** and ***analyzable***.
However, the preprocess steps depend mostly on the task. One task's ideally preprocessing can become another task "nightmare". So it's important to keep in mind that preprocessing is not a one-size-fits-all approach.

2.1 Lowercasing

The simplest technique to start preprocessing data is lowercasing ALL the text. Although simple as it is, lowercasing is the most effective form of text preprocessing that can be applicable to most NLP problems.
It's easy to confuse the model that "Vietnam" and "vietnaM" are 2 different words. Although it has the same meaning, refer to the same country. Here is the example of how lowercasing solves the issue.

Raw	Lowercasing
Vietnam vietNam VIETNAM vietnaM	vietnam
Autumn autumn AuTuMn autumn	autumn

2.2 Map different word to canonical form

Languages we speak and write are made up of several words often derived from one another. When a language contains words that are derived from another word as their use in the speech changes is called ***Inflected Language***.
For simple, we can simply understand that an inflected word will have a *common root*.

Inflected	Root
playing played player	play
better best good	good

2.2.1 Stemming

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.

Stemming uses a crude heuristic process that chops off the ends of words in the hope of correctly transforming words into its root form.

2.2.2 Lemmatization

Stemming and Lemmatization are widely used in *tagging systems, indexing, SEOs, Web search results, and information retrieval*. For example, searching for fish on Google will also result in fishes, fishing as fish is the stem of both words.

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