University of California Davis
Department of Statistics

Machine Learning (STA-208)

Ying-Chen Chou
Chia-Hui Shen
Jiahui Tan
Pei-Ying Ling

Contents

1 Introduction

2 Description of Data

2.1 Raw Data

Year (1999 - 2007)
Sample (Total)
Source (Difficulity of cleaning, how to deal with it(combination of different source, HTML. . . .)

2.2 Top 50 Words Per Year

2.3 Word Cloud

2.4 Mathematics

3 Previous Studies

4 Method

4.1 Feature Engineering

4.2 Fit Models

4.3 Conclusion and exploration