Scalable Document Classification on Spark

tooYoungTooSimple

Department of Statistics
The University of Georgia

Overview

Introduction

Data Cleaning

Naive Bayes Modeling

Prediction

Introduction - Who we are?

- Xiaodong Jiang, Ph.D. Student in Statistics
- Yang Song, M.S. Student in Statistics
- Yaotong Cai, Ph.D. Candidate in Statistics
- Jiankun Zhu, Ph.D. Student in Statistics

Data Cleaning

- Remove special chars, punctuations, and stop words
- Stem and lemmatize all single words, i.e. grouping
- Generate all consecutive bigrams

Naive Bayes Modeling

- Calculate word and bigram tf-idf scores
- Add Laplace smoothing
- Add nonlinear transformation of tf-idf scores
- Classical Naive Bayes Modeling

Thoughts

- More data, better results?
- More grams, better results?
- Nonlinear transformation, better?

Thank You !