Text Classification with Python

Majid Hajiheidari

April 18, 2019

Text Classification with Python

Majid Hajiheidari

Introduction: Divar Datase

Categorizati

No. of Classe

First approach: Naive Bayes

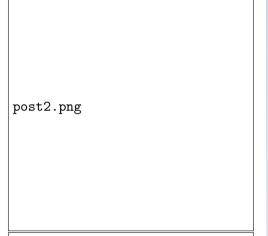
Bayes Classifier

Hyperparameters

Second approach: CNN

Sequence of vectors
How We Embed
Patterns with CNN
Dense Classifier

Divar Posts Dataset



- ► Released for DataDays 2019
- ▶ One million posts

Text Classification with Python

Majid Hajiheidari

Introduction: Divar Dataset

The Problem:

Features
No. of Class

No. of Class

ive Bayes

ayes Classifier

Second appro

CNN
Sequence of vectors

equence of vectors

How We Embed

Patterns with CNN

Dense Classifier

Columns

- ▶ id
- archive_by_user
- published_at
- ► cat1
- ► cat2
- ► cat3
- city title
- desc
- price ▶ image_count
- platform
- mileage
- brand
- year type

Majid Hajiheidari Introduction:

Text Classification

with Python

Divar Dataset

The Problem: Categorization

- ▶ We need to categorize posts based on other posts features;
- ▶ We only use text features(title & description)!

Text Classification with Python

Majid Hajiheidari

Introduction: Divar Dataset

The Problem: Categorization

Features

First approa

Vaive Bayes

Bayes Classifier

Hyperparameters

Hyperparameters

Second approach: CNN

Sequence of vectors
How We Embed
Patterns with CNN

This slide is temp.

Text Classification with Python

Majid Hajiheidari

Introduction Divar Datase

The Problem:

Features

No. of Classes

First approach: Naive Bayes

Count Vectorizer Bayes Classifier

Hyperparameters

Second approach: CNN

Sequence of vectors
How We Embed
Patterns with CNN

This slide is temp.

Text Classification with Python

Majid Hajiheidari

Introduction Divar Datase

ategorizat

No. of Classes

First approach: Naive Baves

Bayes Classifier

Hyperparameters

Second approach: CNN

Sequence of vectors
How We Embed
Patterns with CNN

First Approach: Naive Bayes Classifier bayes_formula.jpg Thomas_Bayes.png

Majid Hajiheidari

Text Classification

with Python

Introduction: Divar Dataset

The Problem: Categorization

First approach: Naive Bayes

Count Vectorizer

Bayes Classifier

econd approach:

CNN Sequence of vectors How We Embed

How We Embed
Patterns with CNN
Dense Classifier

Vectorizing the Text: Count Vectorizer

An example: We want to vectorize these 4 setences¹:

- 1. Hello, how are you!
- 2. Win money, win from home.
- 3. Call me now
- 4. Hello, Call you tomorrow?

Text Classification with Python

Majid Hajiheidari

Introduction: Divar Datase

> he Problem: ategorization

No. of Classes

First approach: Naive Bayes

Count Vectorizer

Bayes Classifier Hyperparameters

Second approach: CNN

Sequence of vectors
How We Embed

How We Embed
Patterns with CNN

¹Example from Rahul Vasaikar

Vectorizing the Text: Count Vectorizer

1. We first build a vocabulary: vocabulary =

{are, call, from, hello, home, how, me, money, now, tomorrow, win, you}

2. Then, we vectorize each sentence based on the occurness of each word:

								money				you
1	1	0	0	1	0	1	0	0	0	0	0	1
2	0	0	1	0	1	0	0	1	0	0	2	0
3	0	1	0	0	0	0	1	0	1	0	0	0
4	0	1	0	1	0	0	0	0	0	1	0	1
1	•	-	•	-	•	•	•	•	•	-	•	-

Text Classification with Python

Majid Hajiheidari

ntroduction: Divar Datase

The Problem: Categorization Features

First approach:

Count Vectorizer
Bayes Classifier

Second approac

Sequence of vectors
How We Embed
Patterns with CNN

Vectorizing the Text: Count Vectorizer

N pair of samples

Text Classification with Python

Majid Hajiheidari

Introduction: Divar Dataset

The Problem Categorization

First approach: Naive Baves

Count Vectorizer Bayes Classifier

Hyperparameters

Second approach: CNN

Sequence of vectors
How We Embed
Patterns with CNN

Bayes Classifier: Naive One!

It is possible to show that accuracy is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values. In other words, we should simply assign a test observation with predictor vector x_0 to the class j for which

$$P(Y = j \mid \mathbf{X} = \mathbf{x})$$

is largest.

Text Classification with Python

Majid Hajiheidari

ntroduction: Divar Datase

The Problem: Categorization Features No. of Classes

First approach: Naive Bayes Count Vectorizer

Bayes Classifier

Second approach:

Sequence of vectors How We Embed Patterns with CNN

Bayes Classifier: Naive One!

We make two assumptions:

- 1. $X_1, X_2, \ldots, and X_m$ are independent from each other:
- 2. $X_1, X_2, \ldots, X_m \mid Y \sim MN(\cdot, p_1, p_2, \ldots, p_m)$

$$P(Y = j \mid \mathbf{X} = (x_1, x_2, \dots, x_m)) = \frac{P(\mathbf{X} = (x_1, x_2, \dots, x_m) \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})}$$

$$= \frac{P(X_1 = x_1 \mid Y = j) \cdot \dots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})}$$
Exist approximate Solution Provided P

$$\hat{y} = \underset{j \in classes}{\operatorname{arg max}} \frac{P(X_1 = x_1 \mid Y = j) \cdot \ldots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})}$$

$$= \underset{j \in classes}{\operatorname{arg max}} P(X_1 = x_1 \mid Y = j) \cdot \ldots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j).$$

Text Classification with Python

Majid Hajiheidari

Bayes Classifier: Naive One!

Let's dive into code!

Text Classification with Python

Majid Hajiheidari

Introduction Divar Datase

ne Problem lategorization

First approach: Naive Bayes

Bayes Classifier

yperparameters

Second approach:

Sequence of vectors
How We Embed
Patterns with CNN

Hyperparameters

Two important hyperparameters:

- 1. Size of the vocabulary:
- 2. Laplace/Lidstone smoothing parameter(α).

Text Classification with Python

Majid Hajiheidari

Hyperparameters

Size of Vocabulary

plot1.png

Text Classification with Python

Majid Hajiheidari

Introduction: Divar Dataset

ategorizatio

First approach: Naive Bayes

Bayes Classifier

Hyperparameters

Second approach

Sequence of vectors
How We Embed
Patterns with CNN

Laplace/ Lidstone Smoothing Parameter(α) plot1.png

Text Classification with Python

Majid Hajiheidari

Introduction: Divar Dataset

> Categorization Features

First approach: Naive Bayes

Bayes Classifier

Hyperparameters

Second approach:

Sequence of vectors How We Embed Patterns with CNN

Grid Search

Text Classification with Python

Majid Hajiheidari

Introduction Divar Datase

The Problen Categorizati

No. of Class

No. of Class

Naive Bayes

Bayes Classifier

Hyperparameters

rryperparameters

Second approach: CNN

Sequence of vector
How We Embed
Patterns with CNN
Dense Classifier

This slide is temp.

Text Classification with Python

Majid Hajiheidari

Introduction Divar Datase

The Problem Categorizatio

No. of Class

No. of Classe

First approach: Naive Bayes

Bayes Classifier

Hyperparameters

Second approach: CNN

Sequence of vectors

How We Embed
Patterns with CNN
Dense Classifier

This slide is temp.

Text Classification with Python

Majid Hajiheidari

Introduction Divar Datase

he Problem Categorizatio

No. of Classe

First approach: Naive Bayes

Bayes Classifier

Hyperparameters

econd approach:

Sequence of vectors

How We Embed

Patterns with CNN

This slide is temp.

Text Classification with Python

Majid Hajiheidari

Introduction Divar Datase

The Problem Categorizatio

No. of Classi

First approach: Naive Bayes

Bayes Classifier

Hyperparameters

Second approach: CNN

Sequence of vectors How We Embed

How We Embed
Patterns with CNN

Dense Classifier

This slide is temp.

Text Classification with Python

Majid Hajiheidari

Introduction Divar Datase

The Problem Categorizatio

No. of Classe

First approach: Naive Baves

Bayes Classifier

Hyperparameters

Second approach: CNN

Sequence of vectors
How We Embed
Patterns with CNN
Dense Classifier

les East

Thanks for your attention!

Codes in slides (in my GitHub):(github link) Divar posts dataset:(divar link) Any questions?

Text Classification with Python

Majid Hajiheidari

Introduction: Divar Dataset

The Problem: Categorization

First approach: Naive Bayes

Bayes Classifier

Second approach:

Sequence of vectors
How We Embed
Patterns with CNN