

# An Experience with Text Classification in *Datadays 2019*

Majid Hajiheidari   Amirmohammad Asadi

April, 2019

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

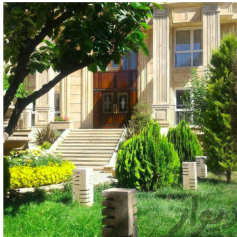
The End

# Divar Posts Dataset

- ▶ Released for DataDays 2019
- ▶ One million posts

بارگشت

همه آگهی ها / املاک / اجاره مسکونی (آپارتمان، خانه، زمین) / آپارتمان



190متر/4خواب/فول محدوده کاج

دقایقی پیش

دریافت اطلاعات تماس


شروع چت

نشان کردن

دسته بندی	آپارتمان
محل	تهران سعادت آباد
نوع آگهی	ارائه
آگهی دهنده	شخصی
تعداد اتاق	چهار
متراژ	۱۹۰

بارگشت

همه آگهی ها / سرگرمی و فراغت / دوچرخه/اسکیت/اسکوتر



دوچرخه مریدا BIG 7-300 سال ۲۰۱۷

۲ ساعت پیش

دریافت اطلاعات تماس

شروع چت

نشان کردن

دسته بندی	دوچرخه/اسکیت/اسکوتر
محل	تهران میدان آزادی
نوع آگهی	فروشی
قیمت	۵,۸۰۰,۰۰۰ تومان

بارگشت

همه آگهی ها / سرگرمی و فراغت / دوچرخه/اسکیت/اسکوتر

با سلام یک دستگاه دوچرخه مریدا BIG 7-300 سال ۲۰۱۷ در حد آک ایک سایز 27/5 تنه 18/5 با کمک باد ست اورژم دنده=طوق و

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Columns

- ▶ id
- ▶ archive\_by\_user
- ▶ published\_at
- ▶ **cat1**
- ▶ **cat2**
- ▶ **cat3**
- ▶ city
- ▶ **title**
- ▶ **desc**
- ▶ price
- ▶ image\_count
- ▶ platform
- ▶ mileage
- ▶ brand
- ▶ year
- ▶ type

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# The Problem: Categorization

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

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
TF-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Features

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# No. of Classes

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Feature Extraction

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Feature extraction is a dimensionality reduction process, where an initial set of raw variables is reduced to more manageable groups (features) for processing, while still accurately and completely describing the original data set.

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Vectorizing the Text: Count Vectorizer

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

An example: We want to vectorize these 4 sentences<sup>1</sup>:

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

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

---

<sup>1</sup>Example from Rahul Vasaikar



# Vectorizing the Text: Count Vectorizer

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

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:

	are	call	from	hello	home	how	me	money	now	tom...	win	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

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
TF-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Vectorizing the Text: Count Vectorizer

N pair of samples

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Tf-idf Vectorizer

- ▶ Tf-idf stands for term frequency-inverse document frequency
- ▶ a statistical measure used to evaluate how important a word is to a document in a collection or corpus
- ▶ the tf-idf weight is composed by two terms:

**TF Term Frequency**, which measures how frequently a term occurs in a document.

$$TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}$$

**IDF Inverse Document Frequency**, which measures how important a term is

$$IDF(t) = \ln \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}$$

# Tf-idf Vectorizer: An Example

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Consider a document containing 100 words wherein the word *cat* appears 3 times. The term frequency (i.e., *tf*) for *cat* is then  $tf(cat) = \frac{3}{100} = 0.03$ . Now, assume we have 10 million documents and the word *cat* appears in one thousand of these. Then, the inverse document frequency (i.e., *idf*) is calculated as  $idf(cat) = \ln \frac{10,000,000}{1,000} = 4$ . Thus, the Tf-idf weight is the product of these quantities:  $tf - idf(cat) = 0.03 * 4 = 0.12$ .

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer

Tf-idf Vectorizer

**Embedding**

Classification  
Algorithms

Naive Bayes

SVM

Passive Aggressive

CNN

A Comparison  
among Models

The End

# Naive Bayes Classifier

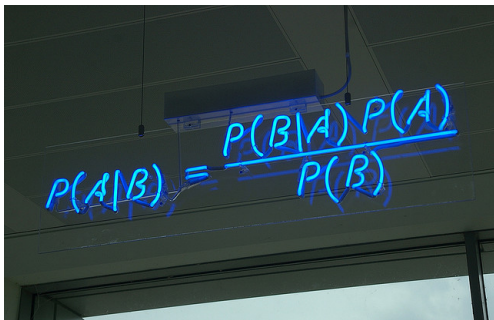

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Photo by Matt Buck



An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes

SVM

Passive Aggressive

CNN

A Comparison  
among Models

The End

# Bayes Classifier: Naive One!

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

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  $\mathbf{x}_0$  to the class  $j$  for which

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

is largest.

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Bayes Classifier: Naive One!

We make two assumptions:

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

$$\begin{aligned} P(Y = j \mid \mathbf{X} = (x_1, x_2, \dots, x_m)) &= \frac{P(\mathbf{X} = (x_1, x_2, \dots, x_m) \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})} \\ &= \frac{P(X_1 = x_1 \mid Y = j) \cdot \dots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})}. \end{aligned}$$

$$\begin{aligned} \hat{y} &= \arg \max_{j \in \text{classes}} \frac{P(X_1 = x_1 \mid Y = j) \cdot \dots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})} \\ &= \arg \max_{j \in \text{classes}} P(X_1 = x_1 \mid Y = j) \cdot \dots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j). \end{aligned}$$



# Bayes Classifier: Naive One!

Let's dive into code!

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

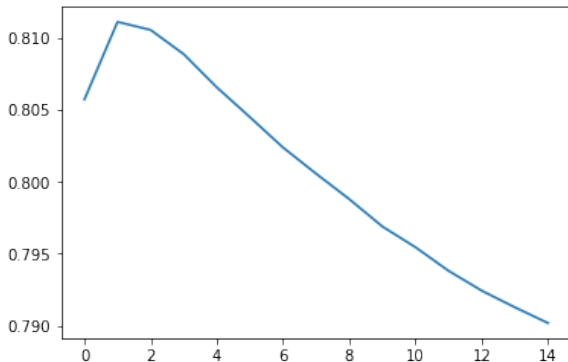
The End

# Hyperparameters

Two important hyperparameters:

1. Size of the vocabulary;
2. Laplace/ Lidstone smoothing parameter( $\alpha$ ).

# Size of Vocabulary



It is convex! (to be completed)

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes

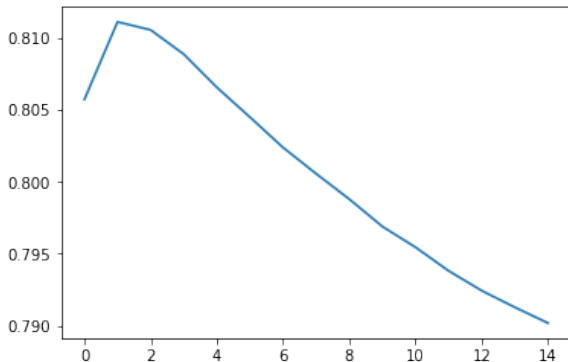
SVM

Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Laplace/ Lidstone Smoothing Parameter( $\alpha$ )



It is convex! (to be completed)

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes

SVM

Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Grid Search

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
**SVM**  
Passive Aggressive  
CNN

A Comparison  
among Models

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
**Passive Aggressive**  
CNN

A Comparison  
among Models

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

**Majid Hajiheidari,  
Amirmohammad  
Asadi**

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
**CNN**

A Comparison  
among Models

The End



# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

**Majid Hajiheidari,  
Amirmohammad  
Asadi**

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
**CNN**

A Comparison  
among Models

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
**CNN**

A Comparison  
among Models

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
**CNN**

A Comparison  
among Models

The End

# Thanks for your attention!

Codes in slides (in my GitHub):(github link)

Divar posts dataset:(divar link)

Any questions?

An Experience  
with Text  
Classification in  
*Datadays 2019*

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Feature Extraction

Count Vectorizer  
Tf-idf Vectorizer  
Embedding

Classification  
Algorithms

Naive Bayes  
SVM  
Passive Aggressive  
CNN

A Comparison  
among Models

The End