An Experience with Text Classification in Datadays 2019

Majid Hajiheidari Amirmohammad Asadi

April, 2019

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Divar Posts Dataset

- Released for DataDays 2019
- ► One million posts







دوحرخه مریدا BIG 7-300سال ۲۰۱۷ ۲ ساعت بیش دوحرخه/اسكيت/اسكوتر دستهبندي تهان مبدان آزادی محل فروشي نوع آگھی oleni Arkensee Cons

با سلام یک دستگاه دوجرخه مریدا BIG 7-300سال ۲۰۱۷ در حد آک آک سایز 27/5 تنه 18/5یا کمک باد ست اوازم دنده=طبق و

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction: Divar Dataset

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 Extracti

Count Vectorizer Tf-idf Vectorizer

Classification

Naive Bayes

aive Bayes NN

inear SVM

inear SVIVI Sective Aggrective Cl

A Comparison

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

The Problem: Categorization

Features

This slide is temp.

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

The Problem: Categorization

No. of Classes

This slide is temp.

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

The Problem: Categorization

Feature Extraction

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.

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction: Divar Dataset

The Problem:

Feature Extraction

Count Vectorizer Tf-idf Vectorizer

Classification

Classification Algorithms

aive Bayes

ive Bayes IN

near SVM

Passive Aggressive Classifie

among Models

Vectorizing the Text: Count Vectorizer

An Experience with Text Classification in Datadavs 2019

Majid Hajiheidari, Amirmohammad Asadi

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?

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:

	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

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction:

Categorization

Feature Extraction

Count Vectorizer

Embedding

lgorithms

laive Bayes

NN

near SVM

assive Aggressive Clas

mong Models

Vectorizing the Text: Count Vectorizer

N pair of samples

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Count Vectorizer

Tf-idf Vectoizer

- ► 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.

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

Inverse Document Frequency, which measures how important a term is

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

An Experience with Text Classification in Datadavs 2019

Majid Hajiheidari, Amirmohammad Asadi

Tf-idf Vectorizer

Tf-idf Vectorizer: An Example

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.

An Experience with Text Classification in Datadavs 2019

Majid Hajiheidari, Amirmohammad Asadi

Tf-idf Vectorizer

This slide is temp.

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Embedding

Classification Algorithms

We used different classifiers and applied different models on the data. The classifiers we tested are:

- Naive Bayes
- Linear Support Vector Machine(SVM)
- Passive Aggressive Classifier
- Convolutional Neural Network(CNN)

An Experience with Text Classification in Datadavs 2019

Majid Hajiheidari, Amirmohammad Asadi

Classification **Algorithms**

Naive Bayes Classifier

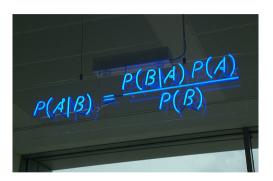


Photo by Matt Buck



An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Naive Baves

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 i for which

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

is largest.

An Experience with Text Classification in Datadavs 2019

Majid Hajiheidari, Amirmohammad Asadi

Naive Baves

Bayes Classifier: Naive One!

We make two assumptions:

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

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

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

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

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

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction:
Divar Dataset

The Problem: Categorization

Feature Extract

Count Vectorizer
Tf-idf Vectorizer

lassification

Naive Bayes

live Dayes

inear SVM

Passive Aggressive

among Models

Bayes Classifier: Naive One!

Let's dive into code!

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Naive Baves

Hyperparameters

Two important hyperparameters:

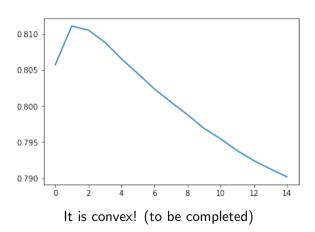
- 1. Size of the vocabulary;
- 2. Laplace/Lidstone smoothing parameter(α).

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Naive Baves

Size of Vocabulary



An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction:

Categorization

Feature Extra

Count Vectorizer
Tf-idf Vectorizer

Classification

Naive Bayes

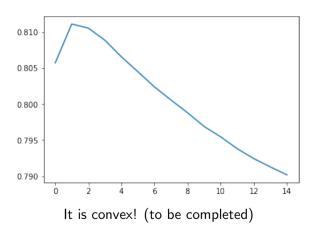
NN

inear SVM

Passive Aggressive Classifi

A Comparison among Models

Laplace/ Lidstone Smoothing Parameter(α)



An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Naive Baves

Grid Search

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Naive Bayes

This slide is temp.

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

CNN

This slide is temp.

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction: Divar Dataset

The Problem: Categorization

Feature Extrac

ount Vectorizer

Tf-idf Vectorize

Classification

Classification Algorithms

aive Bayes

CNN

near SVM

inear SVIVI 'accivo Δσστος

A Comparison

The End

This slide is temp.

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction: Divar Dataset

The Problem:

Feature Extrac

Count Vectorizer Tf-idf Vectorizer

Classification

Algorithms

aive Bayes

CNN

Linear SVM

A Comparison

This slide is temp.

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction: Divar Dataset

The Problem: Categorization

Feature Extrac

unt Vectorizer idf Vectorizer

Embedding

lassification lgorithms

aive Bayes

CNN

near SVM

Passive Aggress

A Comparison among Models

The End

....

Linear SVM

- A non-probablistic classifier
- A discriminative classifier formally defined by a separating hyperplane
- The algorithm outputs an optimal hyperplane which categorizes new examples
- ▶ A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class

An Experience with Text Classification in Datadavs 2019

Majid Hajiheidari, Amirmohammad Asadi

Linear SVM

Linear SVM: Behind the Scene

Given training vectors $x_i \in \mathbb{R}^p$, i=1,...,n, in two classes, and a vector $y \in \{1,-1\}^n$, SVM classifier solves the following primal problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^n \zeta_i$$

subject to $y_i(w^T \phi(x_i) + b) \ge 1 - \zeta_i$,
 $\zeta_i \ge 0, i = 1, ..., n$

Its dual is:

$$\min_{\alpha} \frac{1}{2} \alpha^{T} Q \alpha - e^{T} \alpha$$
subject to $y^{T} \alpha = 0$

$$0 \le \alpha_{i} \le C, i = 1, ..., n$$

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction: Divar Datase

The Problem: Categorization

Feature Extrac

Count Vectorizer Tf-idf Vectorizer

lassification

Algorithms

Naive Bayes

Linear SVM

Passive Aggressive Classific

A Comparison among Models

Linear SVM: Behind the Scene(cont'd)

where e is the vector of all ones, C > 0 is the upper bound, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here, training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ . The decision function is:

$$\operatorname{sgn}(\sum_{i=1}^n y_i \alpha_i K(x_i, x) + \rho)$$

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction: Divar Dataset

The Problem: Categorization

Feature Extraction

Count Vectorize

Classification

laive Bayes

aive Bayes NN

Linear SVM

ear SVIVI ssive Appressive CI

A Comparison among Models

Passive Aggressive

- A margin based online learning algorithm
- Perfect for classifying massive streams
- Easy to implement and very fast
- **Passive**: if correct classification, keep the model;
- **Aggressive**: if incorrect classification, update to adjust to this misclassified example
- See http://koaning.io/passive-agressive-algorithms.html for further reading

An Experience with Text Classification in Datadavs 2019

Majid Hajiheidari. Amirmohammad Asadi

Passive Aggressive Classifier

A Comparison among Models

An Experience with Text Classification in Datadays 2019

Majid Hajiheidari, Amirmohammad Asadi

Introduction:

The Problem: Categorization

Feature Extrac

Tf-idf Vectorizer

Classification

Algorithms

aive Bayes

near SVM

near SVM assive Aggressive C

A Comparison among Models

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