

Класифікація тексту

Misha Beherksy

28 травня 2017 р.

This is abstract

1 All about

На відміну від штучно створених мов, наприклад мов програмування чи математичних нотацій, мови, які ми використовуємо для спілкування, розвивалися з покоління в покоління, постійно видозмінюючись, а тому досить складно відслідкувати і встановити набір чітких конкретно визначених правил. Розробка алгоритмів, що дозволяють "розуміти" людські висловлювання дає змогу покращити велику кількість аспектів взаємодії людини та комп'ютера: передбачення вводу, розпізнавання тексту, пошук інформації в неструктурованому тексті, переклад з однієї мови на іншу, аналіз емоційного забарвлення тексту та багато іншого. Створюючи інтерфейси, що дозволяють людині більш ефективно використовувати комп'ютер, ми прискорюємо розвиток багатомовного інформаційного суспільства.

2 Вступ

Зі стрімким ростом об'єму інформації онлайн, класифікація тексту стала однією з ключових технік для обробки та впорядкування даних. Галузі застосування є досить широкими: починаючи від класифікації новин і закінчуючи персоналізованим пошуком відповідно до потреб користувача. Оскільки побудова власного класифікатора є досить складним та часозатратним процесом, доцільно розглянути приклади уже існуючих класифікаторів. Нижче будуть розглянуті особливості Support Vector Machines (SVMs) класифікатора в контексті класифікації текстів. Метод був запропонований Володимиром Вапником *-* та має значні переваги над іншими в швидкодії та у відсутності довгого процесу тонкого налаштування параметрів моделі.

3

The text classification problem

In text classification, we are given a description \mathbb{X} of a document, where \mathbb{X} is the document space ; and a fixed set of classes \mathbb{C} Classes are also called categories or labels . Typically, the document space \mathbb{X} is some type of high-dimensional space, and the classes are human defined for the needs of an application, as in the examples China and documents that talk about multicore computer chips above.

4 Exploratory data analysis

Візуалізація для наступних цілей: * Комунікативна - представлення даних та ідей - проінформувати - підтримати і аргументувати - вплинути і переконати * Дослідницька - вивчити (дослідити) дані - проаналізувати ситуацію - визначити наступні кроки - прийняти рішення стосовно деякого питання

$$\alpha = \sqrt{\beta} \quad (1)$$

4.1 Класифікація тексту

Метою класифікації текстів є розподіл документів на групи наперед визначених категорій. *-*

5 Висновки

Результати показують, що стабільно показують чудові результати для завдань класифікації текстів, суттєво перевищуючи показники інших методів.

№ п/п	Показники стану ринку (найменування)	Характеристика
1	Кількість головних гравців, од	1
2	Загальний обсяг продаж, грн/ум. од	914 218 млн грн
3	Динаміка ринку (якісна оцінка)	Спадає
4	Наявність обмежень для входу (вказати характер обмежень)	Висока доля невизначеності, відсутність попереднього досвіду та необхідних статистичних даних
5	Специфічні вимоги до стандартизації та сертифікації	-
6	Середня норма рентабельності в галузі (або по ринку), %	18-20%

Табл. 1: Попередня характеристика потенційного ринку стартап-проекту

5.1 Розділ 4. Стартап

Таблиця 1. Опис ідеї стартап-проекту
With width specified:

Day	Min Temp	Max Temp	Summary
Monday	11C	22C	A clear day with lots of sunshine. However, the strong breeze will bring down the temperatures.
Tuesday	9C	19C	Cloudy with rain, across many northern regions. Clear spells across most of Scotland and Northern Ireland, but rain reaching the far northwest.
Wednesday	10C	21C	Rain will still linger for the morning. Conditions will improve by early afternoon and continue throughout the evening.

Зміст ідеї	Напрямки застосування	Вигоди для користувачів
1. Покращення якості зображень	1. Покращення зображень для систем відеонагляду	Отримання більшої кількості інформації
2. Покращення якості зображень	Дає змогу покращити кадр	
3. Покращення якості мрт	Віднайдіння життя громадян	

numeric literals	integers	in decimal	8743
		in octal	0o7464
			0O103
		in hexadecimal	0x5A0FF
	0xE0F2		
	fractionals	in decimal	140.58
			8.04e7
			0.347E+12
5.47E-12			
47e22			
char literals			'H'
			'\n'
			'\x65'
string literals			"bom dia"
			"ouro preto\nmg"

Ринок є доволі привабливим для входження: пристойна середня норма рентабельності, що трохи вище за середній банківський відсоток на вклади у гривні, а спадання ринку потенційно відкриває його для нестандартних інноваційних рішень, оскільки існує дуже висока необхідність в розробці універсального методу для відновлення зображень.