Класифікація тексту

Misha Beherksy

29 травня 2017 р.

Анотація

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 класифікатору 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

$$\alpha = \sqrt{\beta} \tag{1}$$

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 застовування	Вигоди для кори
	1. Покращення зображень для систем відеонагляду	Отримання більш
2. Відновлення деяких фотознімків	Дає змогу покращити кадру	
3. Покращення якості мрт	Віднайдення 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 вищ аза 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.

№ п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Необхідність для інвесторів знайти перспективний метод для вкладень	Люди, які мають фінансову можливість та зацікавленість робоити інвестиції у інноваційні проекти	Люди, які мають фінансову можливість та зацікавленість робити інвестиції у інноваційні проекти мають на меті збільшення свого капіталу, підвищення свого іміджу, а також долучитися до новітніх технологій, щоб бути у тренді	Необхідно розробити методику оцінювання та рекомендації, які б з високою ймовірністю розраховували потенційні необхідні інвестиції та шляхи попередження ключиових ризиків
2	Необхідність команди для побудови цього	Активні люди, які бажають втілити у життя свій проект	Необхідність проаналізувати всі ключові фактори, щоб визначити, чи доцільно реалізовувати проект та чи вдасться залучити спонсорів	Високоточний метод оцінки відновлення зображень, щоб визначити доцільність реалізації відновлення зображень

Табл. 2: Характеристика потенційних клієнтів стаптап-проекту

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Попит	Не вдасться розробити унікальний метод, який би можна було застосовувати для будь-яких відновлення зображень	Розробка максимально універсального методу
2	Науково-технічні	Поява нових технологій, виникнення нових ринкових умов та факторів, які дуже сильно впливають на відновлення зображень	Активне використання наввних рішень; у випадку, якщо наше рішення буде одним з перших та матиме суттєві відмінності від аналогів, захист інтелектуальної власності розробників, патентування цієї технології та додання її до інтелектуальних активів проекту
3	Соціально-культурні	Велика популярність відновлення зображень	Адаптація системи до розширення ринку, появи нових умов та технологій

Табл. 3: Фактори можливостей

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентноспроможною)
1. Тип конкуренції - чиста конкуренція	Велика кількість методів відновлення зображень, частина з яких є запатентованою інтелектувальною власністю	Звертати увагу на якість та універсальність методу відновлення зображень
2. За рівнем конкурентної боротьби - національний	Відновлення зображень не буде прив'язуватися до географічних показників	Акцент в рекламі на потреби жителів великих міст (столиці), таргетування на науковців та молодих дослідників, а також на високозабезпечених людей - потенційних інвесторів
3. За галузевою ознакою - внутрішньогалузева	Конкуренцію складають подібні методики розробки прогностичних моделей	Акцентувати увагу на незвичайність подачі послуг, а також зручність у використанні та надійність, яку вони забезпечують
4. Конкуренція за видами товарів - між бажаннями	Потенційні клієнти роблять вибір між звичними методами побудови моделей (яких дуже велика кількість) і відчувають складність у виборі найбільш доцільного методу	Чітко зрозуміти потреби та бажання кожної з груп цільової аудиторії та розробляти гнучку систему, яка задовольнятиме потреби всіх груп користувачів
5. За характером конкурентних переваг - неціновa	Акцент знаходиться на унікальності та якості послуг, що надаються, а також на перевагах, які отримує клієнт під час використання наших послуг	Робота над покращення методики побудови прогностичних моделей та підвищенням її універсальності
3. За інтенсивністю - не марочна	Продається втілення ідеї, а не певний бренд	Просування ідеї у соціальних мережах

Табл. 4: Ступеневий аналіз конкуренції на ринку