

Автоматизований аналіз мотиваційних листів на основі сентимент аналізу

Рожко Ростислав

Мета

Метою цієї роботи була розробка системи автоматизованого аналізу мотиваційних листів для вступу у вищі навчальні заклади. У ході цієї роботи були розглянуті сучасні методи роботи таких технологій, проблеми обробки текстів природньою мовою. Сфери його застосування та принцип роботи.

Сентимент-аналіз

- Сентимент-аналіз або аналіз тональності тексту (англ. Sentiment analysis, англ. Opinion mining) відноситься до застосування обробки природньої мови, комп'ютерної лінгвістики та аналітики тексту для виявлення та класифікації суб'єктивних думок у вхідних джерелах.

Застосування сентимент-аналізу

Думки є центральними у майже всій діяльності людини, оскільки вони є ключовими чинниками нашої поведінки. Щоразу, коли нам потрібно прийняти рішення, ми хочемо знати думки інших. У реальному світі підприємства та організації завжди хочуть дізнатися думки споживачів чи громадськості щодо своїх товарів та послуг. Споживачі також хочуть дізнатися думку існуючих користувачів товару перед його придбанням, а також думки інших щодо політичних кандидатів, перш ніж приймати рішення про голосування на політичних виборах.

Рівні sentiment-аналізу

- Рівень документа
- Рівень речення
- Рівень сутності та аспекту

Сентимент-лексикон

- Найважливішими показниками тональності є сентимент-слова (англ. sentiment words). Це слова, які зазвичай використовуються для вираження позитивних чи негативних настроїв. Наприклад, хороший, чудовий та дивовижний - це позитивні сентиментальні слова, а поганий та жахливий - слова негативної тональності. Крім окремих слів, існують також фрази та ідіоми.

Визначення думки

$(e_i, a_{ij}, s_{ijkl}, h_k, t_l),$

де e_i - назва сутності, a_{ij} - це аспект e_i , s_{ijkl} - це настрої щодо аспекту a_{ij} сутності e_i , h_k - власник оцінки, а t_l - час, коли думка була висловлена h_k . Настрій s_{ijkl} є позитивним, негативним чи нейтральним або вираженням з різними рівнями сили / інтенсивності, наприклад, від 1 до 5 зірок, як використовується більшість сайтів відгуків в інтернеті.

Завдання сентимент-аналізу

- вилучення та категоризація сутності
- вилучення та категоризація аспектів
- вилучення та категоризація власників думок
- вилучення та стандартизація часу
- категоризація настрою на аспекти
- Скласти усі думки висловлені в документі d на основі результатів вищезазначених завдань

Аналіз мотиваційних листів

Мотиваційний лист - це спосіб пояснити, чому абітурієнт ідеально підходить для навчання у вибраному університеті.

Це можливість особисто описати мотивацію та досвід, який він отримав, що призвело до прийняття цього рішення.

Препроцесинг

- Перетворення слів у символи нижнього регістра
- Токенізація
- Видалення стоп-слів
- Лематизація

Класифікація

Гібридна класифікація поєднує базовий класифікатор, який навчається за допомогою машинного навчання, та систему на основі правил, яка використовується для подальшого покращення результатів. Ці гібридні системи можна легко налаштувати, додавши конкретні правила для тих конфліктуючих тегів, які не були правильно змодельовані базовим класифікатором.

Принцип роботи

- Прізвище, ім'я, по-батькові абітурієнта
- 10 найбільш використаних слів
- Список фраз, які стосуються навчальних здобутків до університету.
- Декілька числових показників (значення від 0 до 1):
- грамотність, тональність (де 0 – негативна, 1 – позитивна)
- Загальна рекомендаційна оцінка якості мотиваційного листа абітурієнта.

Наступні кроки

Точність результатів можна значно покращити збільшивши кількість тренувальних даних. Для цього потрібно ввести механізми за яким людина, яка використовує систему, могла додатково виставляти оцінку якості мотиваційного листа.