Практичне заняття № 5

Тема: Аналіз текстової інформаці

Мета: Розуміти специфіку аналізу текстів. Володіти методикою витягу ключових понять з тексту Володіти засобами аналізу текстової інформації.

Теоретична частина

Перш ніж заглибитися в питання попередньої обробки текстових даних, що передують застосуванню машинного навчання, коротко розглянемо різні типи текстових даних, з якими можна зіткнутися на практиці. Текст, як правило, представлений в наборі даних у вигляді звичайної рядки, проте далеко не всі строкові ознаки обробляються як текст. Строковий ознака іноді може являти собою категоріальні змінні. Обробка строкових ознак неможлива без попереднього аналізу даних.

На практиці можна зустріти чотири типи строкових даних:

- Категоріальні дані
- Неструктуровані рядки, які за змістом можна згрупувати в категорії
- Структуровані рядки
- Текстові дані

Категоріальні дані (categorical data) представляють собою дані, які беруться з фіксованого списку. Наприклад, ви збираєте дані за допомогою онлайн-опитування, в ході якого просите людей назвати їх улюблений колір і для реєстрації відповідей використовуєте випадає з 8 значень («червоний», «зелений», «синій», «Жовтий», «чорний», «білий», «фіолетовий» і «рожевий»), дозволяє респондентам вибрати потрібний варіант.

Остання категорія строкових даних - це текстові дані (text data), які складаються з фраз або пропозицій. Прикладами таких даних можуть бути твіти, логи чату або відгуки про готелі, а також зібрання творів Шекспіра, зміст Вікіпедії або проекту «Гутенберг», що включає 50000 електронних книг. Всі ці колекції містять інформацію, представлену переважно у вигляді пропозицій, складених із слів.

Один з найпростіших, але ефективних і широко використовуваних способів підготовки тексту для машинного навчання - це уявлення текстової інформації у вигляді «мішка слів» (bag-ofwords).

Використовуючи цю виставу, ми видаляємо структуру джерела, таким чином, глави, параграфи, пропозиції, форматування, і лише підраховуємо частоту зустрічальності кожного слова в кожному документі корпусу. Видалення структури і підрахунок частоти кожного слова дозволяє отримати образне уявлення тексту в вигляді «мішка слів». Отримання вистави «мішок слів» включає наступні три етапів:

1. Токенізація (tokenization). Розбиваємо кожен документ на слова, які зустрічаються в ньому (токени), наприклад, за допомогою пробілів і знаків пунктуації.

2. Побудова словника (vocabulary building). Збираємо словник усіх слів, які з'являються в будь-якому з документів, і нумерувати їх (наприклад, в алфавітному порядку).

3. Створення розрідженій матриці (sparse matrix encoding). Для кожного документа підраховуємо, як часто кожне зі слів, занесене в словник, зустрічається в документі.

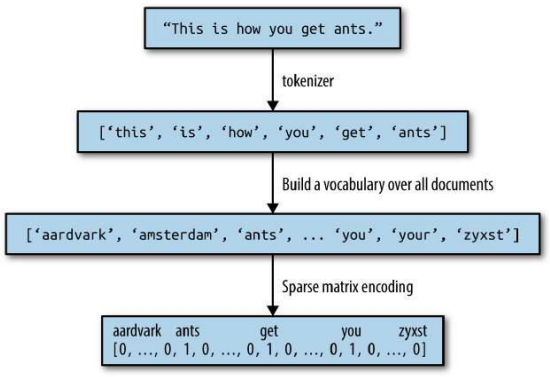

Рисунок 1 – Створення розрідженій матриці

Стоп слова

Ще один спосіб, за допомогою якого ми можемо позбутися від неінформативних слів - виняток слів, які зустрічаються дуже часто, щоб бути інформативними. Існують два основні підходи: використання списку стоп-слів (на основі відповідної мови), або видалення слів, які зустрічаються дуже часто.

Маштабування даних за допомогою tf-idf

Наступний підхід замість виключення несуттєвих ознак намагається масштабувати ознаки в залежності від ступеня їх інформативності. Одним з найбільш поширених способів такого масштабування є метод частота терміна- зворотня частота документа (term frequency-inverse document frequency, tf-idf).

Ідея цього методу полягає в тому, щоб привласнити велику вагу терміну, який часто зустрічається в конкретному документі, але при цьому рідко зустрічається в інших документах корпусу. Якщо слово часто з'являється в конкретному документі, але при цьому рідко зустрічається в інших документах, воно, ймовірно, буде описувати вміст цього документа краще.

У більш складних завданнях обробки тексту часто виникає необхідність поліпшити токенізацію, яка є першим етапом створення моделі «мішка слів». Цей етап визначає, що являє собою слово в плані вилучення ознак.

Раніше ми бачили, що словник часто містить одночасно однину і множину однакових за змістом слів, наприклад, "drawback" і "drawbacks", "drawer" і "drawers", "drawing" і "drawings". При побудові моделі «мішка слів» необхідно враховувати близькість слів "drawback" і "drawbacks" за змістом, присутність цих слів у вигляді окремих ознак лише збільшить перенавчання замість того, щоб дозволити моделі в повній мірі використовувати навчальні дані.

Цю проблему можна вирішити, знайшовши для кожного слова його основу (word stem). Це має на увазі ідентифікацію або об'єднання (conflating) всіх слів з однієї і тієї ж основою. Якщо цей процес виконується за допомогою евристик на основі правил (наприклад, видалення загальних суфіксів), його зазвичай називають стемінг (stemming). Якщо замість цього використовується словник із заздалегідь заданими формами слів (Явний процес, контрольований людиною) і враховується роль слова в реченні (тобто беремо до уваги, до якої частини мови належить слово), то цей процес називається лематизації (lemmatization), а стандартизована форма слова називається леммой (Lemma). Лематизації і стемінг є способами нормалізації (normalization), які намагаються отримати певну нормальну (тобто початкову) форму слова.

Моделювання тем і кластезизація документів

Ще один метод, який часто застосовується до текстових даними - моделювання тем (topic modeling). Моделювання тим - це зонтичний термін, що описує процедуру присвоєння кожному документу однієї або кількох тем, яка здійснюється, як правило, без учителя. Хорошим прикладом моделювання тим є новинні дані, які можна згрупувати за такими темами, як «політика», «спорт», «фінанси» і так далі.

Якщо кожен документ може мати тільки одну тему, то мова йде про завдання кластеризації документів. Якщо кожен документ може мати кілька тем, ця задача відноситься до Декомпозиційні методам. Кожна отримана компонента відповідає одній темі, а коефіцієнти компонент, які описують документ, дозволяють нам судити про те, наскільки тісно даний документ пов'язаний з конкретною темою. Часто, коли люди говорять про моделювання тим, вони мають на увазі конкретний декомпозиційний метод під назвою латентне розміщення Дирихле (Latent Dirichlet Allocation, LDA).

Завдання до виконання:

Підготувати проект, який реалізує програмно наступні конструкції:

1. Приклад аналізу тональності текстів.
2. Представлення текстових даних у вигляді «мішка слів».
3. Токенізація, стемінг, лематизація.