Извлечение признаков

Ксемидов Борис Сергеевич

Chillers

2 апреля, 2020

Извлечение признаков

Чаще всего данные представлены в сыром виде. Для использования в методах машинного обучения их необходимо привести в соответствующий вид (вид набора данных - матрицы).

Обработка текста

Текст представляется в виде последовательности символов.

Пример:

```
Кто сражается с чудовищами, тому следует
остерегаться, чтобы самому при этом
не стать чудовищем.
И если ты долго смотришь в бездну,
то бездна тоже смотрит в тебя.
```

Этапы обработки текста

1. Токенизация
2. Стемминг/лемматизация
3. Векторизация

Токенизация

Разбиение текста на токены (чаще всего на слова).

Пример текста:

```
И если ты долго смотришь в бездну,
то бездна тоже смотрит в тебя.
```

Примеры токенов:

```
['И', 'если', 'долго', 'смотришь',
 'в', 'бездну', 'то', 'бездна',
 'тоже', 'смотрит', 'в', 'тебя']
```

Проблемы токенизации

- Форма слова
- Опечатки
- Жаргонизмы

Проблемы токенизации

Примеры:

- `Нижний Новгород -> ['Нижний', 'Новгород']`
- `воруй-убивай! -> ['воруй', 'убивай']`

N-граммы

N-граммы — это непрерывные последовательности n-элементов в предложении. N может быть 1, 2 или любым другим положительным целым числом.
Пример текста:

```
И если ты долго смотришь в бездну,
то бездна тоже смотрит в тебя.
```

Пример разбиения (для $N = 2$):

```
['И если', 'если ты', 'ты долго', ...]
```

Стемминг и лемматизация

Стемминг

Цель стемминга заключается в нахождении основы слова, то есть части слова без словоизменительных аффиксов. Стемминг отсекает суффиксы и окончания слов.

Пример:

```
лесной -> лес
походный -> поход
столовый -> стол
```

Стемминг и лемматизация

Лемматизация

Цель лемматизации заключается в приведении слова в нормальную форму.

Принцип преобразования лемматизации:

1. Существительное — единственное число, именительный падеж.
2. Прилагательное — единственное число, мужской род, именительный падеж.
3. Глагол — неопределенная форма (инфинитив).

Стемминг и лемматизация

Пример лемматизации:

```
Бежала -> бежать
Кошку -> кошка
зеленого -> зеленый
```

Векторизация текста

Методы векторизации текста:

- Bag of Words (мешок слов)
- TF-IDF

Bag of Words

Идея

Токенизация всех текстов и подсчёт количества вхождений для каждого токена в каждый текст.

Пример текстов:

```
1  Я есть Грут!
2  Я не есть Грут!
```

Примеры "мешка слов":

```
1  {'Я': 1, 'не': 0, 'есть': 1, 'Грут': 1}
2  {'Я': 1, 'не': 1, 'есть': 1, 'Грут': 1}
```

Недостатки Bag of Words

- Идентичность предложений с разной семантикой;
- объём словаря.

TF-IDF

Определение

Статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов.

TF-IDF состоит из двух идей:

- TF (частота слова в документе)
- IDF (обратная частота документа)

TF

Идея

Частота вхождения слова.

$$TF(t, d) = \frac{n_t}{\sum_k n_k},$$

где

n_t - число вхождений слова t в документ

$\sum_k n_k$ - общее число слов в документе

IDF

Идея

Уменьшение веса широкоупотребительных слов.

$$\mathrm{IDF}(t, D) = \log \frac{|D|}{|D_t|}, D_t = \{ d_i \in D \mid t \in d_i \}$$

где

$|D|$ — число документов в коллекции;

$|D_t|$ - число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$).

TF-IDF

Формула

$$TFIDF(t, d, D) = TF(t, d) * IDF(t, D)$$

Обработка изображений

Способы обработки изображений:

- "пиксельная" векторизация;
- обнаружение границ;
- обнаружение углов.

Обнаружение границ изображения

Рисунок: Выделение границ

Фильтр Собеля

Идея

Оператор Собеля основан на операции свёртки.

$$G_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} * A$$

$$G_y = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} * A$$

$$G = \sqrt{G_x^2 + G_y^2}$$

Обнаружение углов изображения

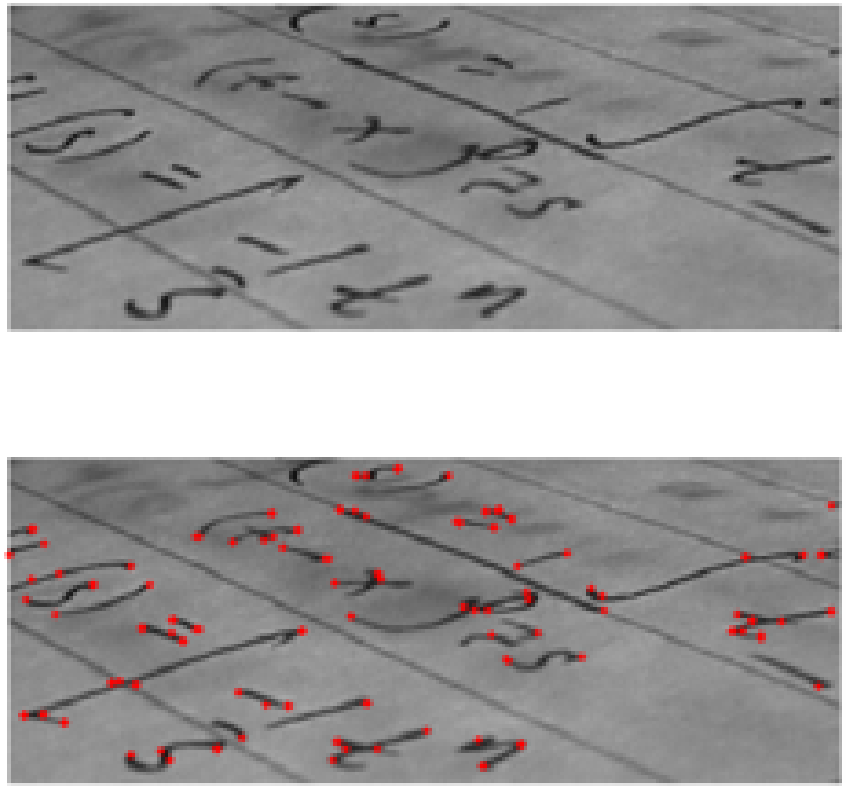

Рисунок: Выделение углов на изображении

Спасибо за внимание!