О языковых моделях

Евгений Борисов

о языке и задачае автоматичекой его обработки

```
обработка текстов на естественом языке (ЕЯ) natural language processing (NLP)
```

- NLU / natural language understathding
- NLG / natural language generation
- SP / speech processing (recognition/generation)

обработка текстов на естественом языке

NLP/ NLU natural language understanding

- natural entity recognition распознавание именованных сущностей
- classification intent классификация намерений
- sentiment analysis оценка тона

схема применения методов ML

объект -> вектор-признак -> модель -> ответ

кодирование текста и извлечение признаков

- кодирование текста
- кодирование слова
- кодирование отдельных символов

кодирование текста - частотный анализ (TF-IDF)

- собираем словарь
- считаем для каждого слова его повторы в тексте
- обрабатываем вектора-признаки (TF-IDF) текстов

кодирование текста - частотный анализ (TF-IDF)

- собираем словарь
- считаем для каждого слова его повторы в тексте
- обрабатываем вектора-признаки (TF-IDF) текстов

недостатки:

- не учитывает порядок слов
- для коротких сообщений не информативен

простое кодирование слов в тексте

- собираем словарь
- заменяем слова в тексте на их номера в словаре
- обрабатываем последовательности номеров (RNN)

простое кодирование слов в тексте

- собираем словарь
- заменяем слова в тексте на их номера в словаре
- обрабатываем последовательности номеров (RNN)

недостатки:

• номер не отражает семантику слова

кодирование символов в тексте

- собираем алфавит из текстов
- заменяем символы в тексте на их номера в алфавите
- обрабатываем последовательности номеров символов

кодирование символов в тексте

- собираем алфавит из текстов
- заменяем символы в тексте на их номера в алфавите
- обрабатываем последовательности номеров символов

недостатки:

- отдельные символы не информативны
- может занимать много памяти

кодирование слов word2vec

- собираем словарь
- заменяем слова в тексте на их номера в словаре
- из текстов собираем пары [номер слова, [номера слов контекста]]
- строим нейросеть *SkipGram* пытаемся по слову определить контекст
- извлекаем матрицу первого слоя [номер слова, код слова]
- заменяем слова в текстах кодами w2v
- обрабатываем последовательности кодов w2v

кодирование слов word2vec

- собираем словарь
- заменяем слова в тексте на их номера в словаре
- из текстов собираем пары [номер слова, [номера слов контекста]]
- строим нейросеть *SkipGram* пытаемся по слову определить контекст
- извлекаем матрицу первого слоя [номер слова, код слова]
- заменяем слова в текстах кодами w2v
- обрабатываем последовательности кодов w2v

недостатки:

• w2v однозначен, слова часто имеют несколько смыслов (лук,коса,кисть)

языковая модель

предсказываем следующее слово на основе предыдущих

последовательность слов

$$w = (w_1, w_2, w_3, \dots, w_k)$$

правило условной вероятности:

$$p(w)=p(w_1)p(w_2|w_1)p(w_3|w_2w_1)...p(w_k|w_{k-1}...w_1)$$

предположение Маркова:

$$p(w_i|w_{i-1}...w_1)=p(w_i|w_{i-1}...w_{i-n})$$

биграммная языковая модель (n = 2):

$$p(w)=p(w_1)p(w_2|w_1)p(w_3|w_2w_1)...p(w_k|w_{k-1}w_{k-2})$$

нейросетевая языковая модель (word based model)

- собираем словарь
- заменяем слова в тексте на их номера в словаре
- из текстов собираем пары [[номера слов контекста], номер слова]
- обучаем RNN по контексту определять слово input -> LSTM -> softmax
- выкидываем выходной слой (softmax), остальное используем как feature extractor

свёрточная нейросетевая языковая модель (charCNN)

- кодируем слово посимвольно в ОНЕ, получаем матрицу {0,1} размера
 [позиция символа в слове, количество символов в алфавите] к ней можно применить двумерную свёртку
- собираем словарь
- собираем алфавит
- разбираем текст на слова (токенизация)
- заменяем символы в словах на их номера в алфавите
- из текстов собираем пары [[матрицы символов слов контекста], номер слова в словаре]
- обучаем RNN по контексту определять следующий символ
 - input -> Conv2D -> MaxPooling2D -> LSTM -> softmax
- выкидываем выходной слой (softmax), остальное используем как feature extractor

Литература

git clone https://github.com/mechanoid5/ml_lectorium.git

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean Distributed Representations of Words and Phrases and their Compositionality

Евгений Борисов О методе кодирования слов word2vec http://mechanoid.su/ml-w2v.html

A.A.Потапенко Языковое моделирование http://www.machinelearning.ru

Анатолий Востряков Языковые модели на все случаи жизни, ODS Data Fest 2018 https://www.youtube.com/watch?v=TaCbj1kaDQY



Вопросы?