Лекция 17:
Автоматическая обработка текстов на естественном языке.
Метод кодирования слов word2vec.

Евгений Борисов

NLP word2vec

Способ описание текста

частотный анализ

- нужен достаточный размер текста
- не учитывает последовательность

NLP word2vec

Способ описание текста

частотный анализ
- нужен достаточный размер текста
- не учитывает последовательность

кодирование отдельных слов
- можно использовать для коротких сообщений
- можно учитывать последовательность

NLP word2vec

способ кодирования слов

тривиальный способ

составить словарь, отсортировать и занумеровать

NLP word2vec

способ кодирования слов

тривиальный способ

составить словарь, отсортировать и занумеровать

Недостатки: номер не отражает смысла

NLP word2vec

способ кодирования слов

Word2Vec

из текста извлекаем словарь W
каждому слову из W ставим в соответствие точку из V

$$w2v: W \rightarrow V\ ; V \subset \mathbb{R}^n$$

NLP word2vec

способ кодирования слов

Word2Vec

из текста извлекаем словарь W
каждому слову из W ставим в соответствие точку из V

$$w2v : W \rightarrow V \; ; V \subset \mathbb{R}^n$$

совместно употребляемые в тексте слова из W
отображаются в близкие точки пространства V

w2v[king] − w2v[man] + w2v[woman] ≈ w2v[queen]

NLP word2vec

Как это работатет?

подготовка данных Word2Vec – учитываем контекст слов.

- очищаем текст T от лишних символов

- из очищенного текста T собираем словарь W

- для каждого слова w собираем контекст (окрестность)
  т. е. слова удалённые от w не более чем на s позиций в T

- выполняем унитарное кодирование(one-hot encoding) W

Pi:

0 0 1 0 0

Qi:

0 1 0 0 0
0 0 0 0 1
0 0 0 1 0
1 0 0 0 0
0 0 0 1 0

NLP word2vec

Как это работатет?

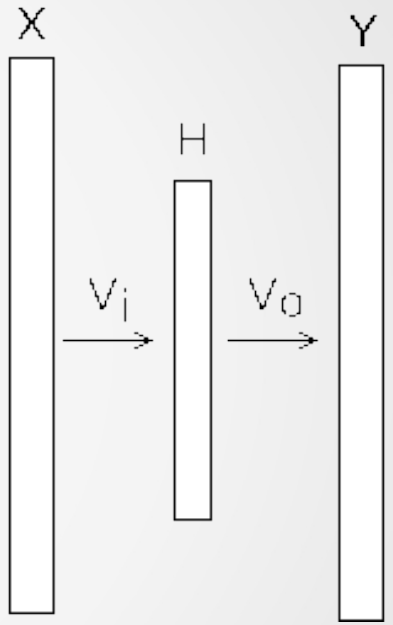
нейросеть Word2Vec

размер входного слоя X = размеру словаря W
= размеру выходного слоя Y

скрытый слой H - линейная активация

выходной слой Y – активация softmax

$$Y=softmax((X \cdot Vi) \cdot Vo)$$



конечный результат - матрица внутренних представлений *Vi*

NLP word2vec

обучение сети word2vec

метод градиентного спуска

одна из двух стратегии

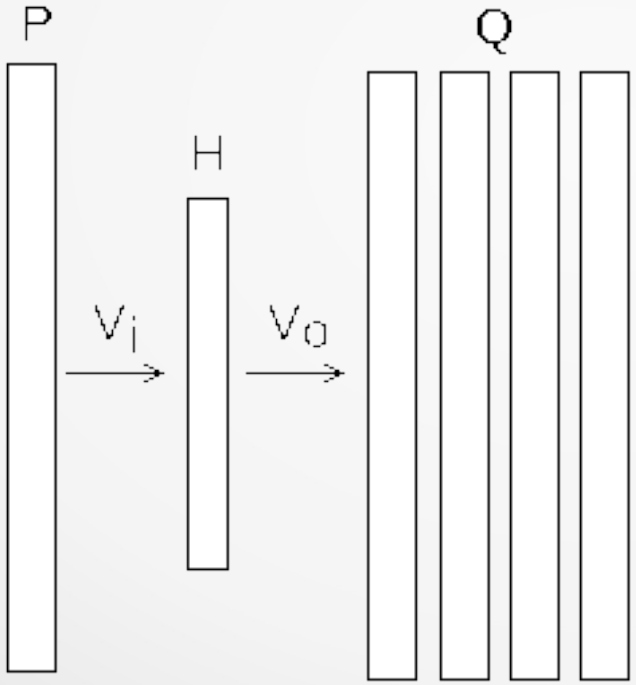
- Skip-Gram - по слову восстанавливаем контекст.

- CBOW(Continuous Bag of Words) по контексту восстанавливаем слово

NLP word2vec

обучение сети word2vec

- Skip-Gram - по слову восстанавливаем контекст.



NLP word2vec

обучение сети word2vec - Skip-Gram - по слову восстанавливаем контекст.

1.на вход сети подаётся код слова P,
вычисляем состояние скрытого слоя H
вычисляем выход сети O

2. вычисляем значение функции потери

$$E_i = \left| \log \sum \exp(U_i) - \sum \sum_j (U_i * Q_{ij}) \right|$$

если значение потери увеличилость
то конец работы

3. для каждого слова контекста Q_j и входа P:

вычисляем ошибку D на выходе сети O
и изменение весов сети ΔV_o, ΔV_i

$$D = O - Q_j$$
$$\Delta Vo_j = H^T \cdot D$$
$$\Delta Vi_j = D^T \cdot P \cdot Vo^T$$

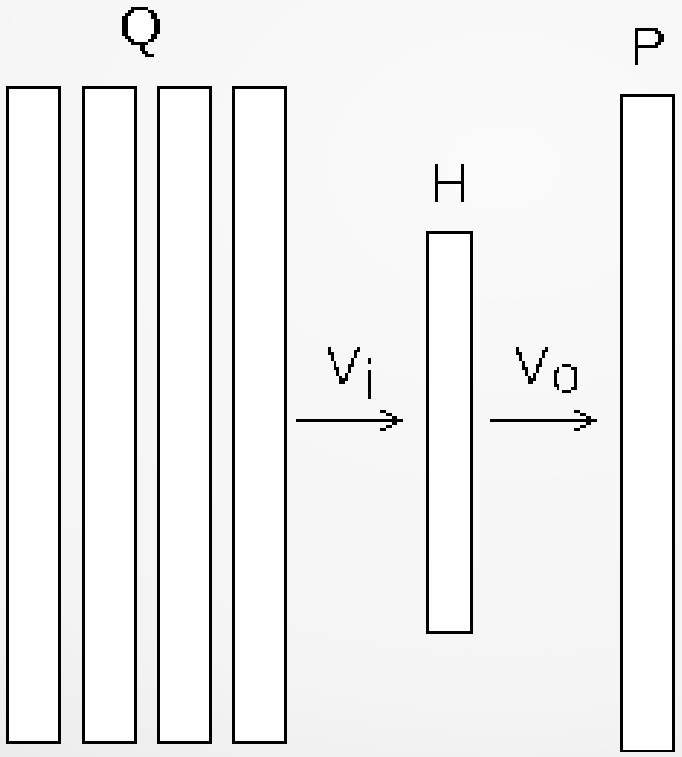
4. вычисляем суммарное изменение
весов сети ΔV_o, ΔV_i
корректируем веса
и повторяем цикл для другого слова P

$$\Delta Vo = \sum_j \Delta Vo_j$$
$$\Delta Vi = \sum_j \Delta Vi_j$$

NLP word2vec

обучение сети word2vec

- CBOW(Continuous Bag of Words) по контексту восстанавливаем слово



NLP word2vec

обучение сети word2vec - CBOW, по контексту восстанавливаем слово

1.на вход сети подаётся усреднённое значение контекста Q,
вычисляем состояние скрытого слоя H
вычисляем выход сети O

$$H = \frac{1}{c}\sum_{j=1}^{c} Q_j \cdot Vi$$

$$U = H \cdot V_o$$
$$O = softmax(U)$$

2.вычисляем значение функции потери

$$E_i = \left|\log \sum \exp(U_i) - \sum (U_i * P_i)\right|$$

если значение потери увеличилость
то конец работы

3. для каждого слова контекста Qj и кода слова P,
вычисляем ошибку D на выходе сети O
и изменение весов сети ΔVo, ΔVi.

$$D = O - P$$
$$\Delta Vo = H^T \cdot D$$
$$\Delta Vi = \sum_j D^T \cdot Q_j \cdot Vo^T$$

4. корректируем веса
и повторяем цикл для другого слова P

NLP word2vec

Результат работы

слово	близкие по w2v
смотрит	подозрительно, кровати
при	приворовываешь, чём
она	семья, разваливается
ещё	важно, поучительно
самого	конца, последней
алкоголик	покуриваешь, травку
способности	определённые, солнца
ответственность	странице, авторской
портал	произведения, читателей
разваливается	знаю, семья
рецензию	написать, рукой
подобию	господа, образу

NLP word2vec

Литература

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

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

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

NLP word2vec

Вопросы ?

NLP word2vec

источники данных для экспериментов

sklearn.datasets
UCI Repository
kaggle
www.ruscorpora.ru