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

PMI pointwise mutual information

оценка совместного использования слов *u v*

$$PMI(u,v)=\log\left(\frac{p(u,v)}{p(v)\,p(u)}\right)$$

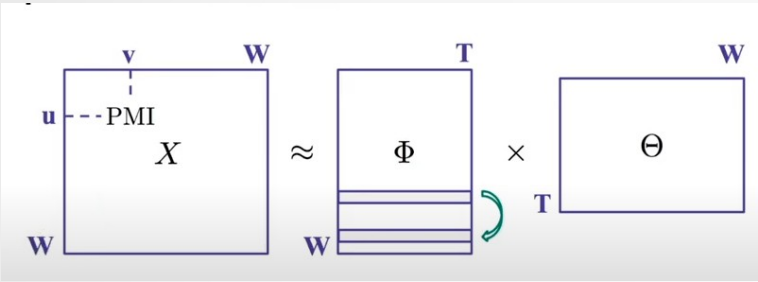
p(u,v) – частота использования словосочетания

p(u) и *p(v)* - частота использования слов

NLP word2vec

используем контекст для кодирования слов

применим матричное разложение к квадратной матрице PMI



для кодирования слов используем матрицу Φ

NLP word2vec

способ кодирования слов Word2Vec

построим модель ML и обучим её кодировать слова по контексту

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

построим модель ML и обучим её кодировать слова по контексту

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?

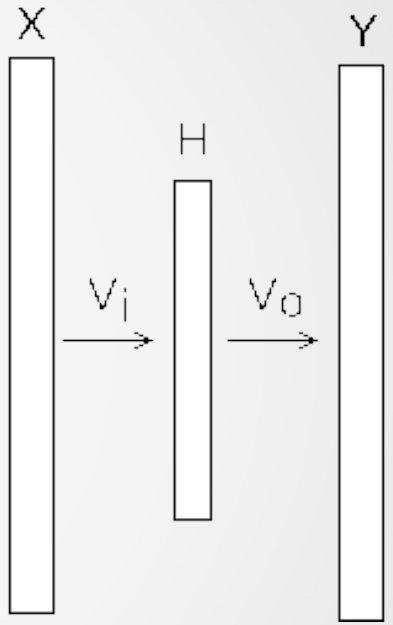
нейросеть Word2Vec

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

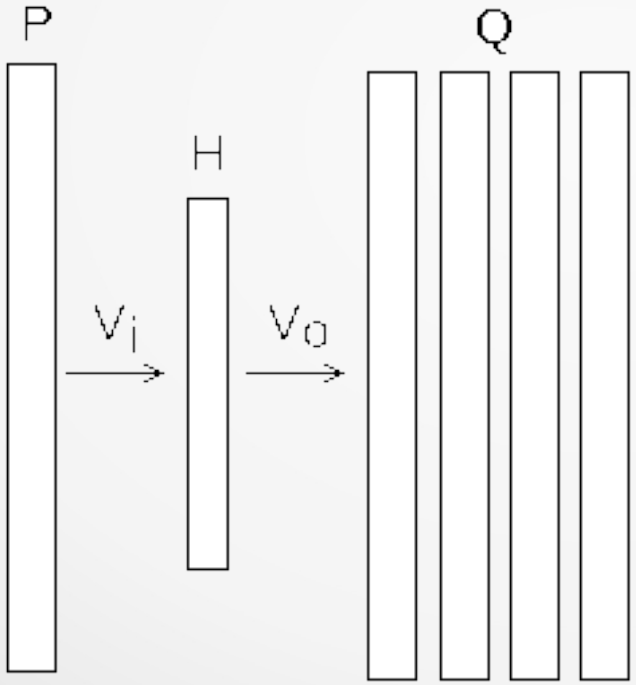
одна из двух стратегии

- 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

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

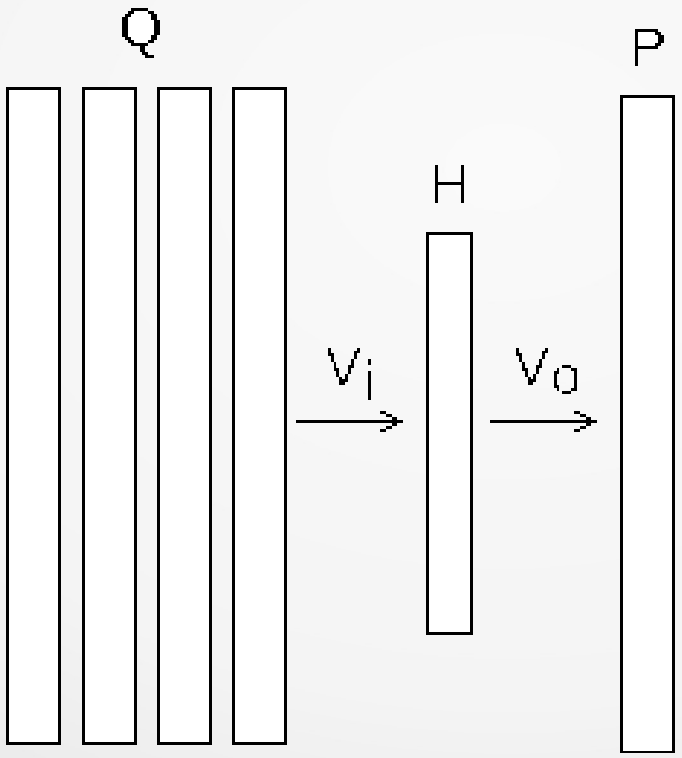
4. вычисляем суммарное изменение
весов сети ΔV_o, ΔV_i
корректируем веса
и повторяем цикл для другого слова P

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

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

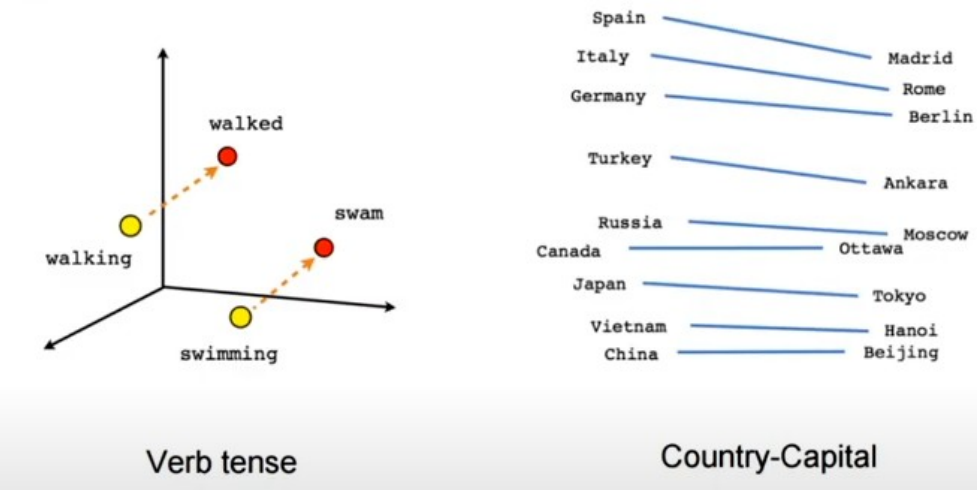
NLP word2vec

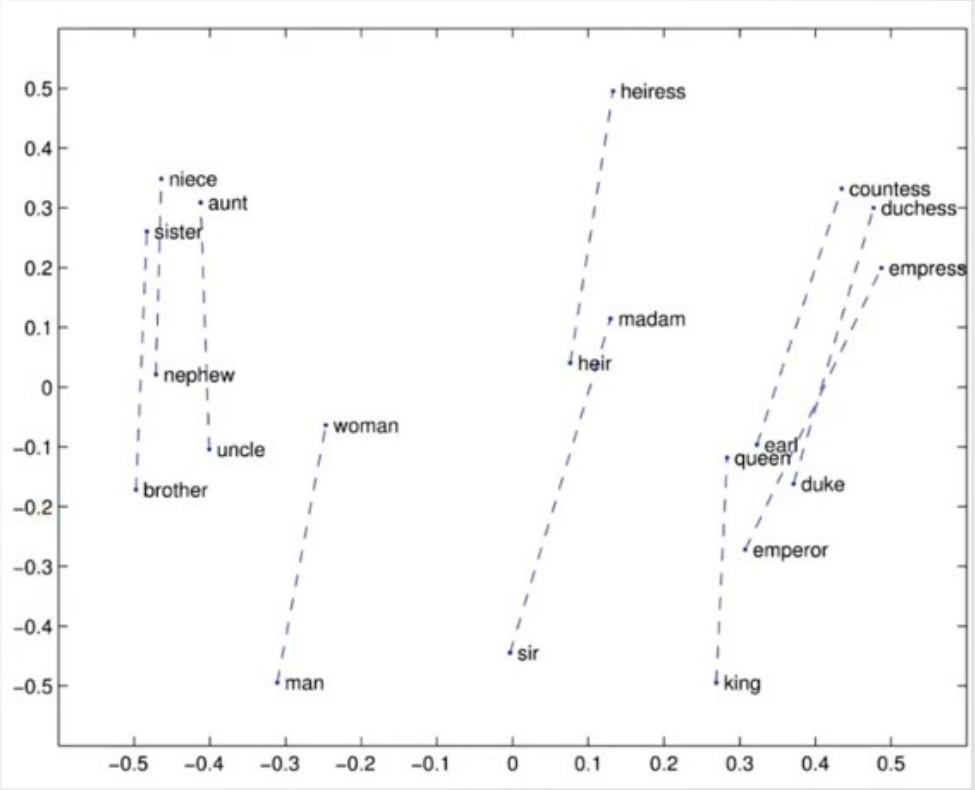
Результат работы

```
{'состоится': ['конечно', 'бал'],
 'волнения': ['описывать', 'виргинский'],
 'печатных': ['листа', 'липутину'],
 'лямшин': ['вечер', 'чуткости'],
 'или': ['отчасти', 'нет'],
 'крышу': ['под', 'надлежащего'],
 'общее': ['пригорюнясь', 'пострадать'],
 'степанович': ['осилил', 'новых'],
 'часов': ['кучке', 'часу'],
 'вас': ['для', 'завтра'],
 'слишком': ['милости', 'была'],
 'крикнул': ['попа', 'прочтет'],
 'видеть': ['свое', 'будто'],
 'наготове': ['кармазинове', 'беспокойством'],
 'помещался': ['этою', 'подыматься'],
 'выполнено': ['планете', 'несмотря'],
 'тебя': ['люблю', 'взрывов'],
 'началом': ['фунтов', 'центру'],
 'страдал': ['заявите', 'тесно'],
 'отчаянная': ['новая', 'кончив']}
```

NLP word2vec

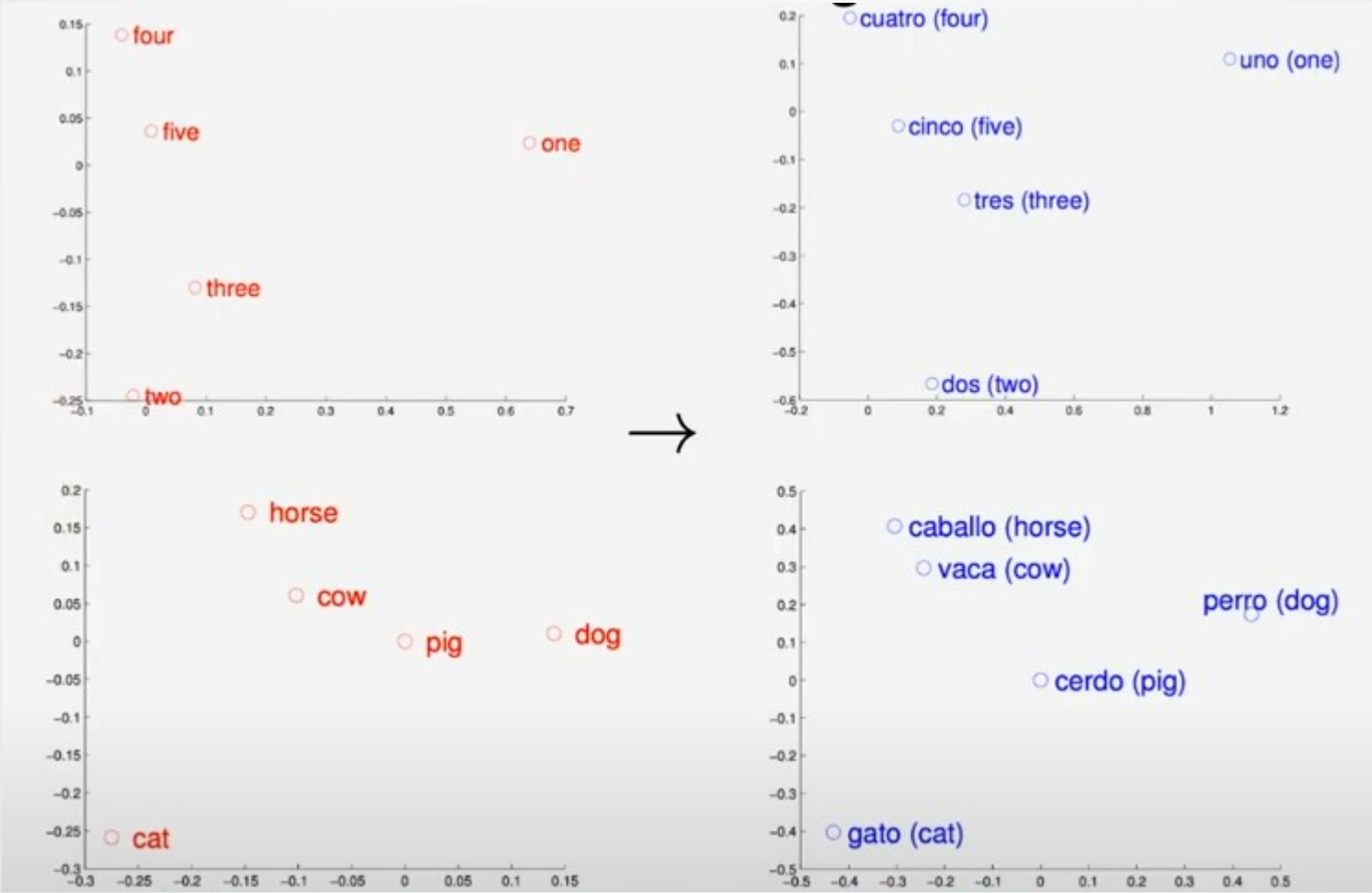
близкие по контексту слова отображаются в близкие точки w2v





NLP word2vec

взаимное расположения w2v в разных языках похожи

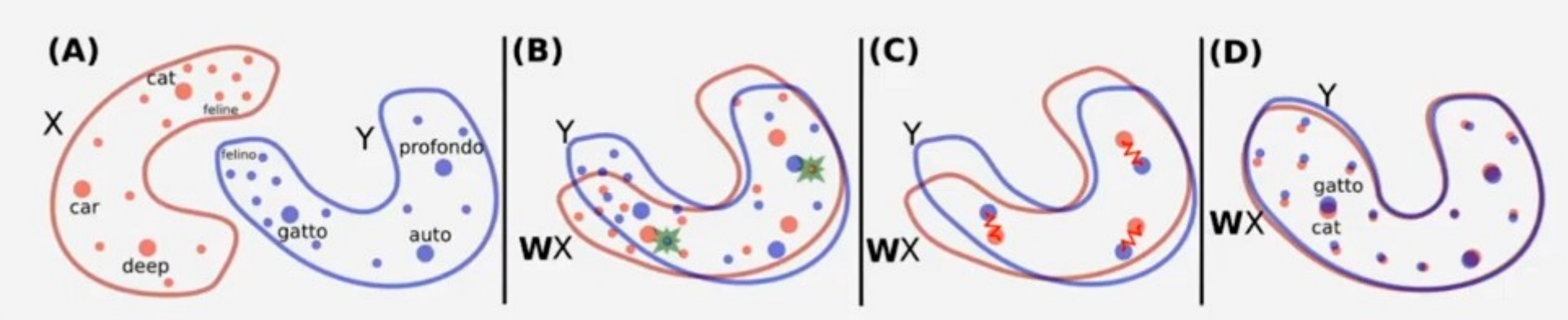


NLP word2vec

взаимное расположения w2v в разных языках похожи

зная перевод некоторых слов и на основе этого построив отображение из w2v пространства одного языка в другое,

мы получаем перевод всех остальных слов на основе контекста



NLP word2vec

косинусная мера близости – оценка сонаправленности w2v

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

NLP word2vec

Литература

```
git clone https://github.com/mechanoid5/ml_nlp
```

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

Радослав Нейчев Прикладное машинное обучение 1.Intro to NLP. Word embeddings - Лекторий ФПМИ

Николай Карпачев Прикладное машинное обучение. Семинар 1.Word embeddings - Лекторий ФПМИ

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