Автоматическая обработка текстов на естественном языке. Методы кодирования слов.

Евгений Борисов

NLP : Методы кодирования слов

Уровни сложности при автоматической обработке текстов

Прагматика (Дискурс) - смысловые контексты
Семантика - смыслы последовательностей слов
Синтаксис - правила формирования последовательностей слов
Лексика - отдельные слова и устойчивые словосочетания

NLP : Методы кодирования слов

Частотный анализ (TF-IDF) не учитывает порядок слов

Будем кодировать слова отдельно и обрабатывать цепочки кодов

Способы кодирования слов:

- Составить словарь и занумеровать слова
- CharCNN - посимвольное кодирование в матричное представление
- PMI - оценка совместного использования слов
- Word2Vec - кодирование слова по контексту (Word Embeddings)

NLP : Методы кодирования слов

Посимвольное кодирование слов charCNN

матрица [позиция символа в слове, номер символа в алфавите]

CAB :

	A	B	C
1	0	0	1
2	1	0	0
3	0	1	0

текст представляем как упорядоченый набор матриц слов (тензор)

[позиция символа в слове, номер символа в алфавите, номер слова]

к тензору уже можно применять свёрточную сеть вида

input → Conv2D → MaxPooling2D → MLP → softmax

NLP : Методы кодирования слов

Статистическая оценка семантической близости

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

квадратная матрица контекстов

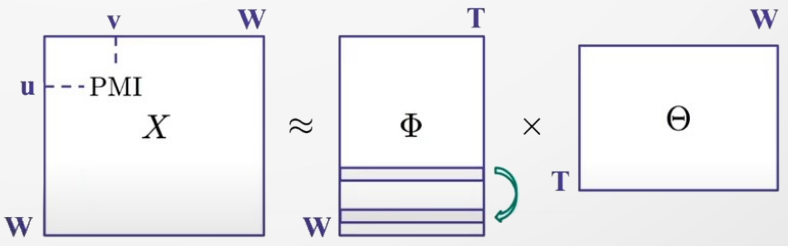
Pointwise Mutual Information (PMI)

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



применим матричное разложение к квадратной матрице PMI
для кодирования слов используем матрицу Φ

NLP : Методы кодирования слов

Оценка семантической близости в семантических пространствах

Word Embeddings - кодирование слова по контексту

Word2Vec - совместно употребляемые в тексте слова
отображаются в близкие точки пространства

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

Gensim – реализация на Python

построим ML-модель и обучим её кодировать слова по контексту

NLP : Методы кодирования слов

подготовка данных Word2Vec – учитываем контекст слов.

- из 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 : Методы кодирования слов

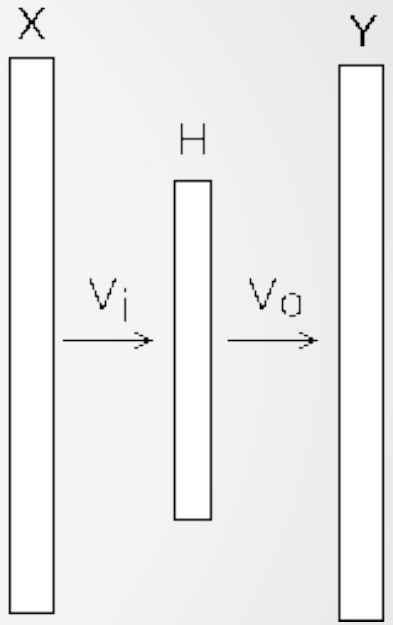
нейросеть Word2Vec

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 : Методы кодирования слов

обучение сети word2vec

метод градиентного спуска

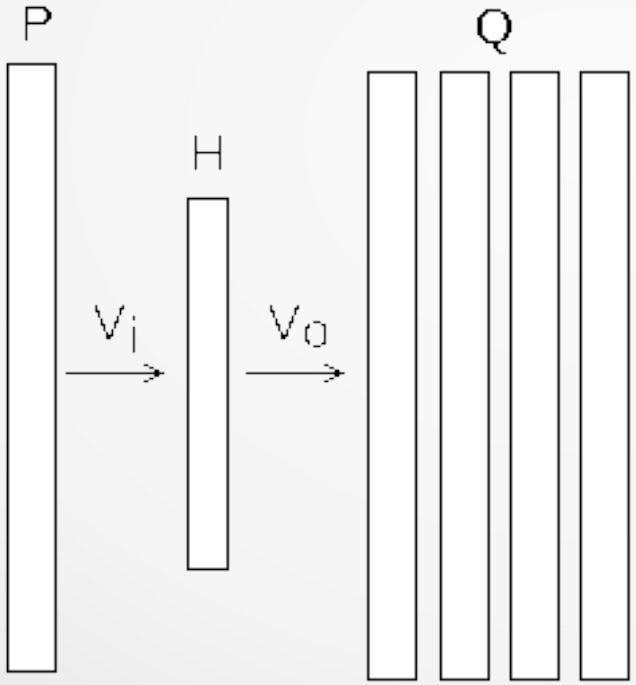
одна из двух стратегии

- Skip-Gram - по слову восстанавливаем контекст.
- CBOW(Continuous Bag of Words) по контексту восстанавливаем слово

NLP : Методы кодирования слов

обучение сети word2vec

- Skip-Gram - по слову восстанавливаем контекст.



NLP : Методы кодирования слов

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

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

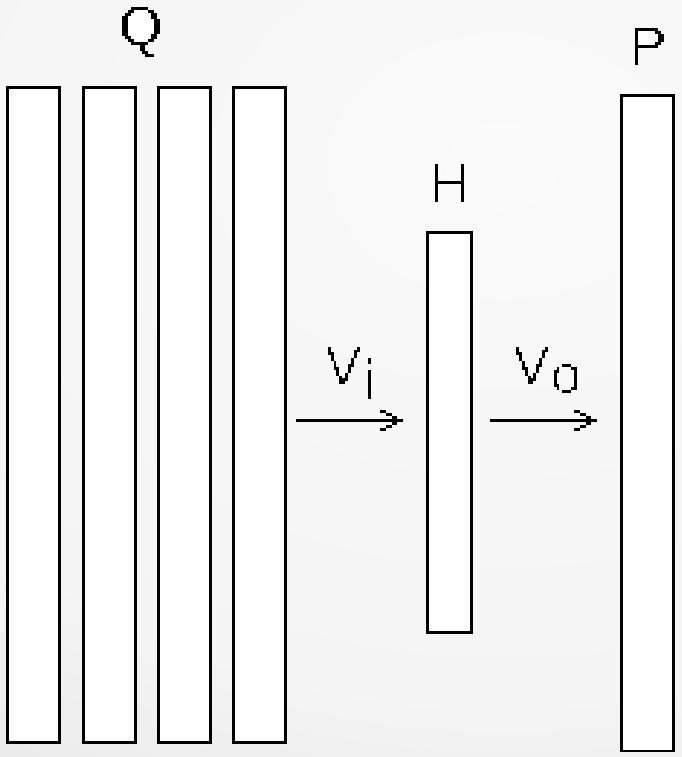
4. вычисляем суммарное изменение
весов сети ΔV_o, ΔV_i
корректируем веса
и повторяем цикл для другого слова P

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

NLP : Методы кодирования слов

обучение сети word2vec

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



NLP : Методы кодирования слов

обучение сети word2vec - CBOW, по контексту восстанавливаем слово

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

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

вычисляем выход сети O

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

2.вычисляем значение функции потери

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

если значение потери увеличилось
то конец работы

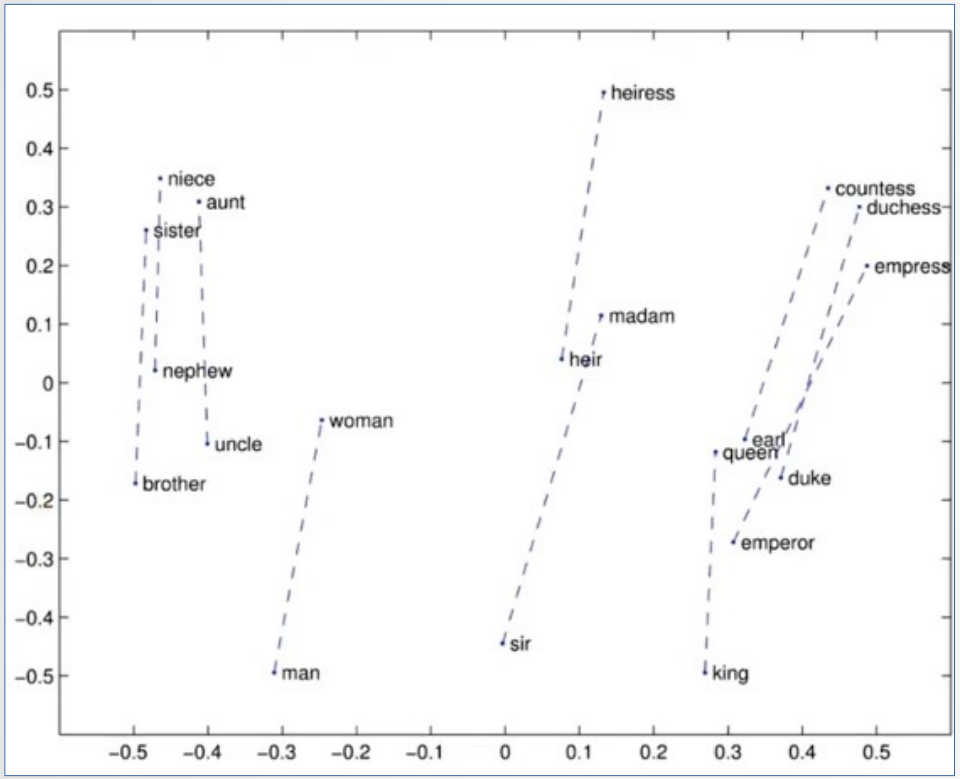
3. для каждого слова контекста Qj и кода слова P,
вычисляем ошибку D на выходе сети O
и изменение весов сети ΔVo, ΔVi.

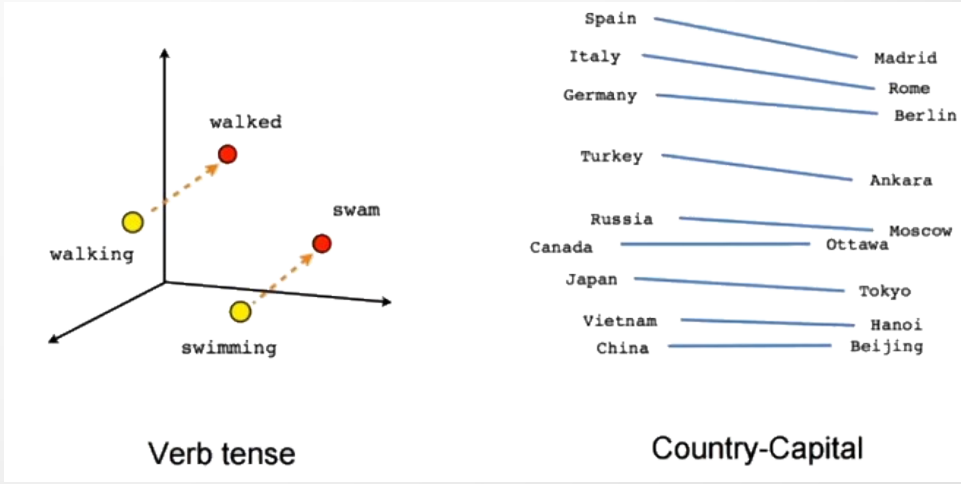
$$D = O - P$$
$$\Delta Vo = H^T \cdot D$$
$$\Delta Vi = \sum_j D^T \cdot Q_j \cdot Vo^T$$

4. корректируем веса
и повторяем цикл для другого слова P

NLP : Методы кодирования слов

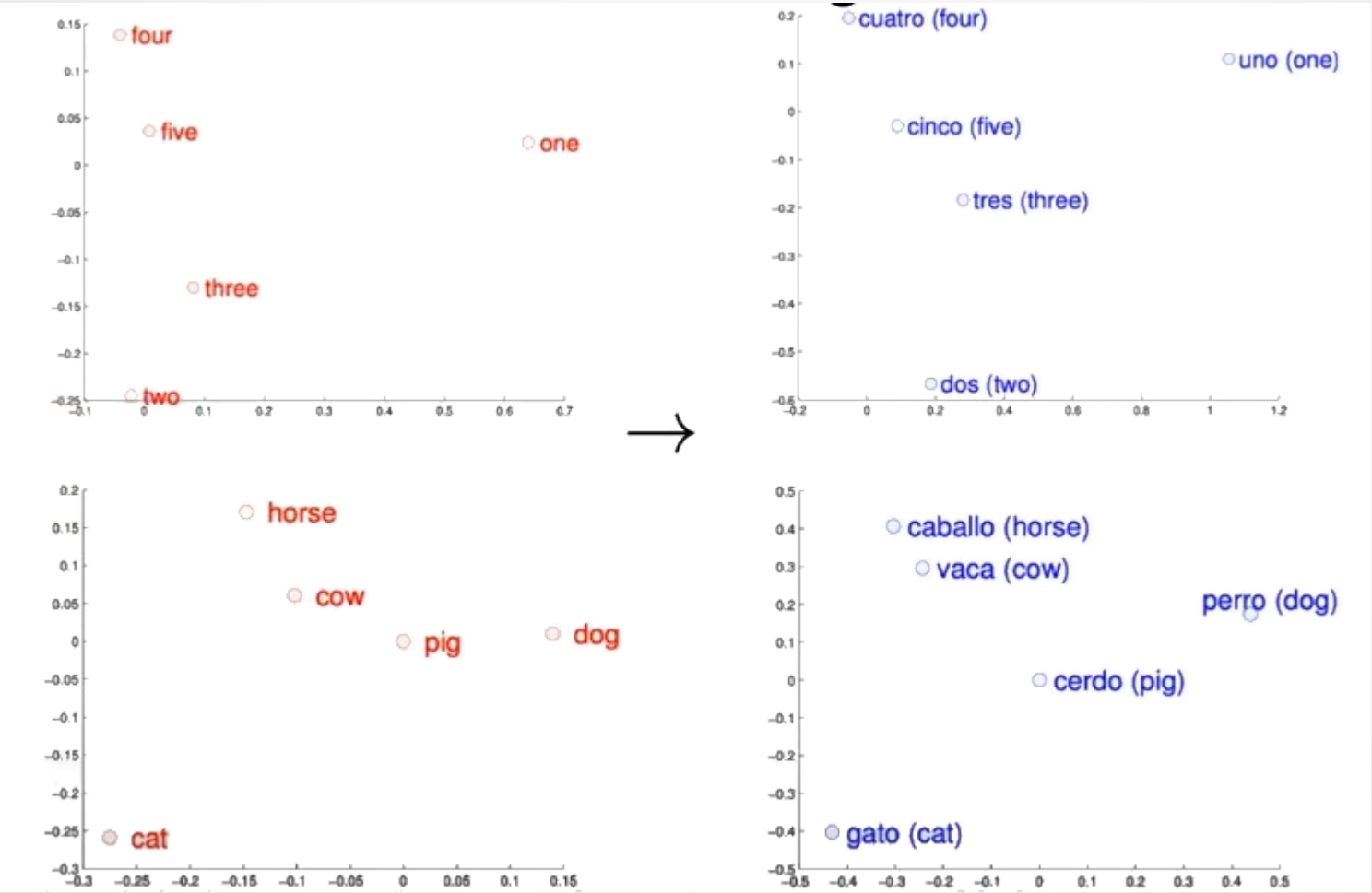
близкие по контексту слова отображаются в близкие точки w2v





NLP : Методы кодирования слов

взаимное расположение w2v в разных языках схожи

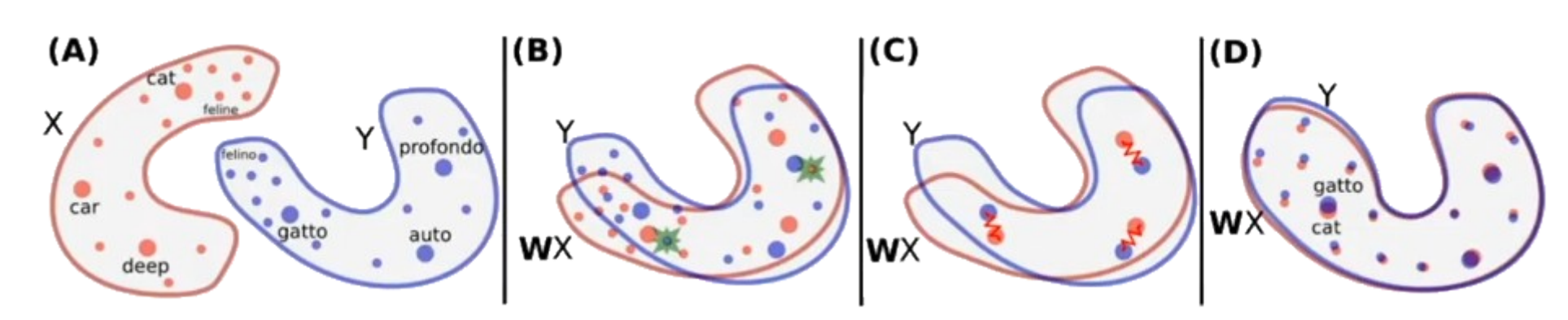


NLP : Методы кодирования слов

взаимное расположения w2v в разных языках схожи

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 : Методы кодирования слов

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

Радослав Нейчев Прикладное машинное обучение 1.Intro to NLP. Word embeddings - Лекторий ФПМИ
https://www.youtube.com/watch?v=aZ5se_SW81c

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