

Машинное обучение

Дополнительная лекция

Введение в обработку естественных языков

Власов Кирилл Вячеславович



2018

Обработка естественного языка

Обработка естественного языка (*Natural Language Processing, NLP*) — общее направление искусственного интеллекта и математической лингвистики.

Распознавание речи

Анализ текста

- Извлечение информации
- Информационный поиск
- Анализ высказываний
- Анализ тональности текста
- Вопросно-ответные системы
- Генерирование текста

Синтез речи

Задачи анализа и синтеза в комплексе:

- Машинный перевод
- Автоматическое рефериование, аннотирование или упрощение текста



Яндекс



Google



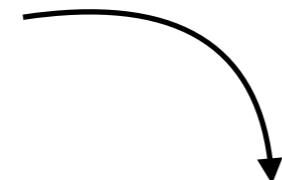
Почему задача не триивиальна?

Перед нами стол. На столе стакан и вилка. Что они делают? Стакан стоит, а вилка лежит. Если мы воткнем вилку в столешницу, вилка будет стоять. Т.е. стоят вертикальные предметы, а лежат горизонтальные? Добавляем на стол тарелку и сковороду. Они вроде как горизонтальные, но на столе стоят. Теперь положим тарелку в сковородку. Там она лежит, а ведь на столе стояла. Может быть, стоят предметы готовые к использованию? Нет, вилка-то готова была, когда лежала. Теперь на стол залезает кошка. Она может стоять, сидеть и лежать. Если в плане стояния и лежания она как-то лезет в логику «вертикальный-горизонтальный» , то сидение - это новое свойство. Сидит она на попе. Теперь на стол села птичка. Она на столе сидит, но сидит на ногах, а не на попе. Хотя вроде бы должна стоять. Но стоять она не может вовсе. Но если мы убьём бедную птичку и сделаем чучело, оно будет на столе стоять. Может показаться, что сидение - атрибут живого, но сапог на ноге тоже сидит, хотя он не живой и не имеет попы. Так что, поди ж пойми, что стоит, что лежит, а что сидит.



Препроцессинг

Исходный документ из корпуса текстов



На практике очень часто возникают задачи для решения которых используются методы оптимизации в обычной жизни при множественном выборе например подарков к новому году мы интуитивно решаем задачу минимальных затрат при заданном качестве покупок

Препроцессинг

Исходный документ из корпуса текстов

На практике очень часто возникают задачи для решения которых используются методы оптимизации в обычной жизни при множественном выборе например подарков к новому году мы интуитивно решаем задачу минимальных затрат при заданном качестве покупок

Лемматизация

на практика очень часто возникать задача для решение который использоваться метод оптимизация в обычный жизнь при множественный выбор например подарок к новый год мы интуитивно решать задача минимальный затрата при задавать качество покупка

Препроцессинг

**Исходный документ из
корпуса текстов**

На практике очень часто возникают задачи для решения которых используются методы оптимизации в обычной жизни при множественном выборе например подарков к новому году мы интуитивно решаем задачу минимальных затрат при заданном качестве покупок

Стеминг

на практик очен част возника задачи, для решен котор использ метод оптимизац в обычн жизн при множествен выборе, например, подарк к нов год мы интуитивн реше задач минимальн затрат при зада качеств покупок

Лемматизация

на практика очень часто возникать задача для решение который использоваться метод оптимизация в обычный жизнь при множественный выбор например подарок к новый год мы интуитивно решать задача минимальный затрата при задавать качество покупка

Feature engineering

Использовать ли короткие слова? (предлоги)

Использовать ли очень частые и/или очень редкие?

Какой длины использовать Ngrams? Нужны ли они?

Feature engineering

Использовать ли короткие слова? (предлоги, а что с «не»)
Использовать ли очень частые и/или очень редкие?
Какой длины использовать Ngrams? Нужны ли они?

Ngrams

практика очень часто возникать задача

2grams



практика очень
очень часто
часто возникать
возникать задача

Feature engineering

Использовать ли короткие слова? (предлоги)
Использовать ли очень частые и/или очень редкие?
Какой длины использовать Ngrams? Нужны ли они?

Шум в данных

**Огромный объем
данных**

Ngrams

практика очень часто возникать задача

2grams



практика очень
очень часто
часто возникать
возникать задача

Feature engineering

Bag of words

Три документа:

- Артур любит смотреть фильмы
- Ольга тоже любит фильмы
- Денис любит смотреть футбол

Feature engineering

Bag of words

Три документа:

- Артур любит смотреть фильмы
- Ольга тоже любит фильмы
- Денис любит смотреть футбол



Артур любит смотреть фильмы
Ольга тоже Денис футбол



Feature engineering

Bag of words

Три документа:

- Артур любит смотреть фильмы
- Ольга тоже любит фильмы
- Денис любит смотреть футбол



Артур любит смотреть фильмы
Ольга тоже Денис футбол



id	Артур	Ольга	Денис	Любит	смотреть	Фильмы	тоже	Футбол
1	1	0	0	1	1	1	0	0
2	0	1	0	1	0	1	1	0
3	0	0	1	1	1	0	0	1

Feature engineering

TF-IDF (от англ. *TF* – *term frequency*, *IDF* – *inverse document frequency*) – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

Feature engineering

TF (term frequency — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

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

где n_t есть число вхождений слова t в документ, а в знаменателе — общее число слов в данном документе.

Feature engineering

TF (term frequency — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

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

где n_t есть число вхождений слова t в документ, а в знаменателе — общее число слов в данном документе.

IDF (inverse document frequency — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}, \text{[2]}$$

где

- $|D|$ — число документов в коллекции;
- $|\{d_i \in D \mid t \in d_i\}|$ — число документов из коллекции D , в которых встречается t (когда $n_t \neq 0$).

Feature engineering

TF (term frequency — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

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

где n_t есть число вхождений слова t в документ, а в знаменателе — общее число слов в данном документе.

IDF (inverse document frequency — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}, \text{[2]}$$

где

- $|D|$ — число документов в коллекции;
- $|\{d_i \in D \mid t \in d_i\}|$ — число документов из коллекции D , в которых встречается t (когда $n_t \neq 0$).

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Проблемы

**Огромные и сильно
разряженные матрицы**

Word2Vec

Source Text

The quick brown fox jumps over the lazy dog. →

The quick brown fox jumps over the lazy dog. →

The quick brown fox jumps over the lazy dog. →

The quick brown fox jumps over the lazy dog. →

Training Samples

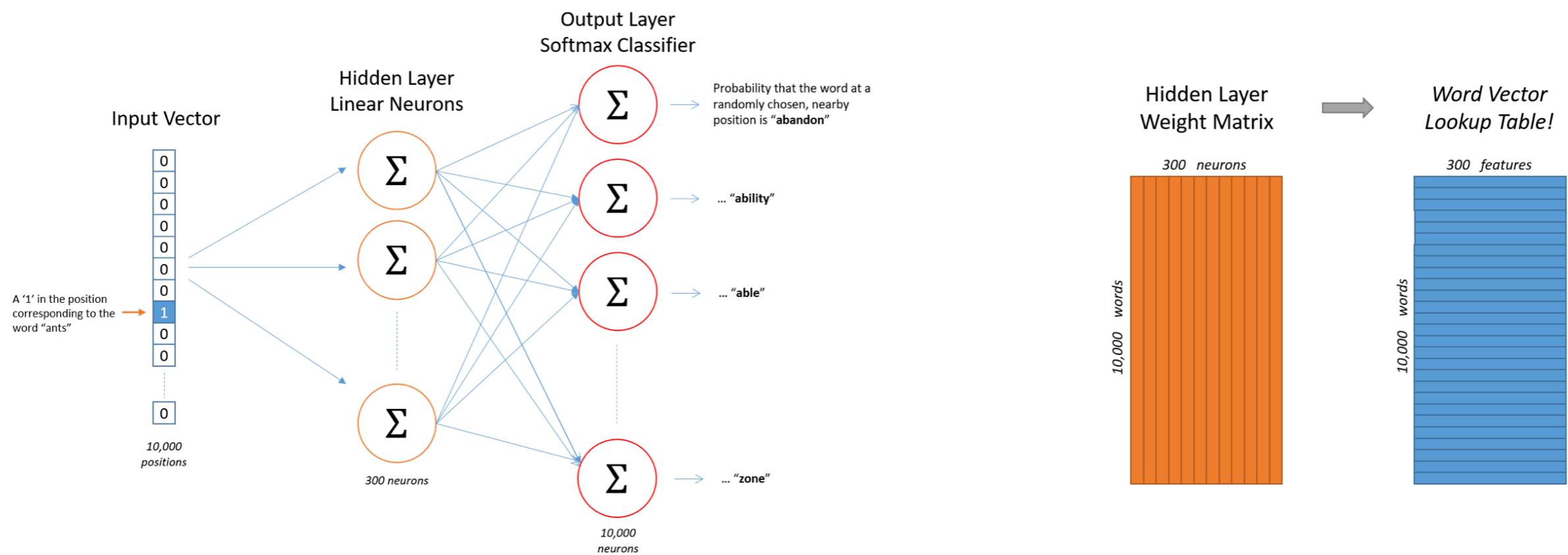
(the, quick)
(the, brown)

(quick, the)
(quick, brown)
(quick, fox)

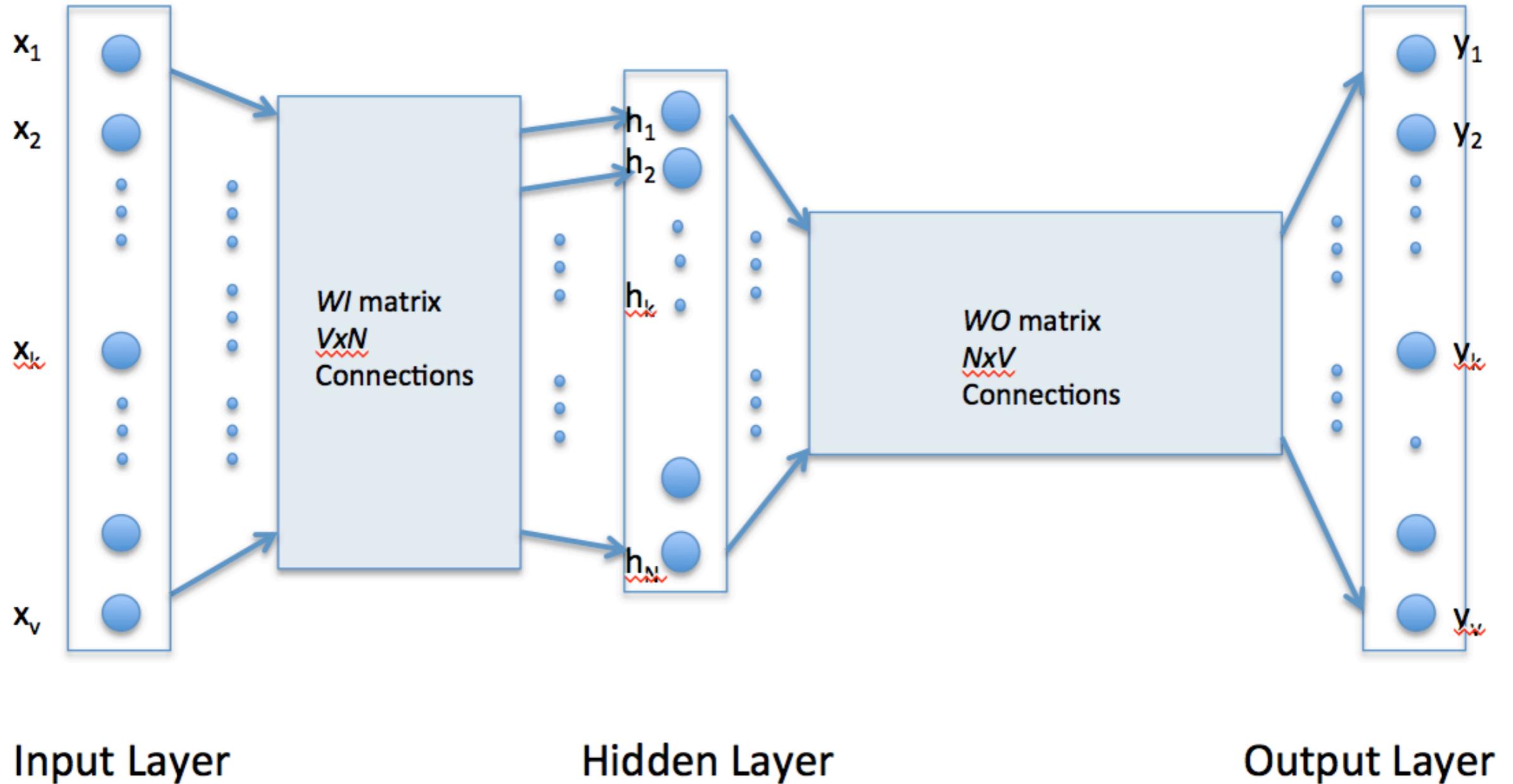
(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)

(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

Word2Vec

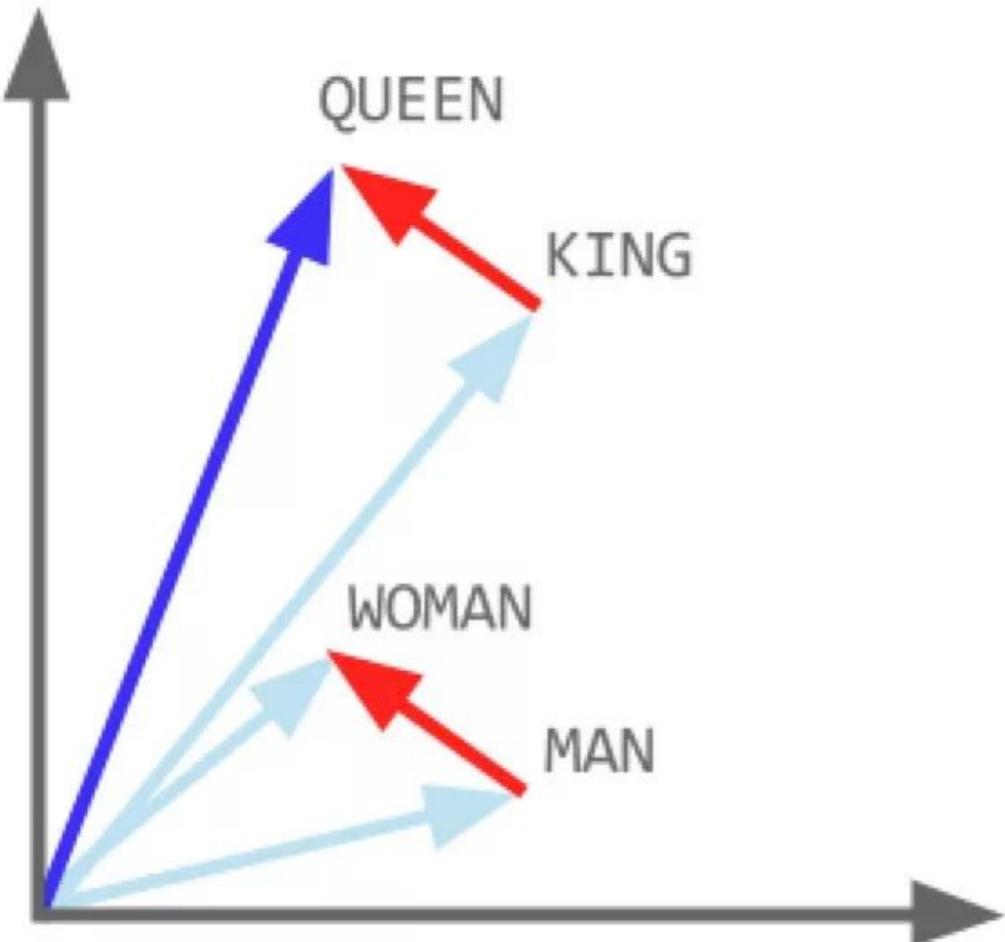


Word2Vec



Word2Vec

So $\text{king} + \text{man} - \text{woman} = \text{queen}$!



- Geopolitics: Iraq - Violence = Jordan
- Distinction: Human - Animal = Ethics
- President - Power = Prime Minister
- Library - Books = Hall
- (Moscow - Russia) + France = ?

И это все?

FastText
Glove

Что не обсуждаем?
Тематическое моделирование