Анализ неструктурированных данных

Лекция

Базовые векторные представления слов и документов

Мурат Апишев (mel-lain@yandex.ru)

Векторные представления

- *Векторное представление* (embedding) — сопоставление произвольному объекту некоторого числового вектора в пространстве фиксированной размерности
- Наиболее известный вид – векторные представления слов (word embedding)
- Векторы могут обладать разнообразными полезными свойствами, отражать близость объектов в разных смыслах
- Для слов это может быть семантическая близость

Зачем нужны векторные представления

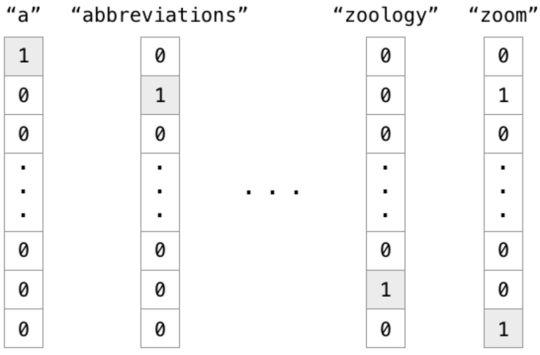
В современных подходах эмбеддинги используются в качестве признаков для решения почти любых задач машинного обучения

В текстовой аналитике это:

1. выделение именованных сущностей (NER)
2. выделение частей речи (POS-tagging)
3. машинный перевод
4. кластеризация документов
5. классификация документов, анализа тональности (sentiment)
6. ранжирование документов
7. генерация текста

One-hot encoding

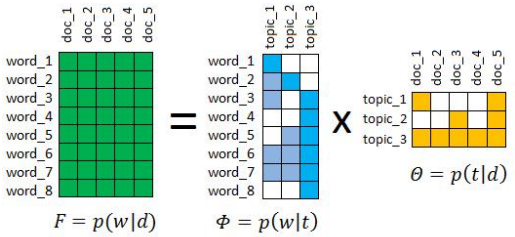
Самый простой способ кодирования категориальных признаков:



Полученные векторы огромные и ортогональные

Ссылка на источник картинки

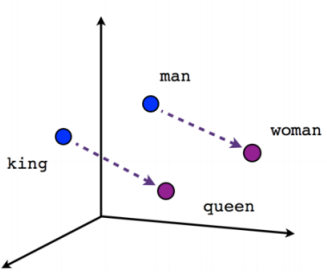
Тематическое моделирование

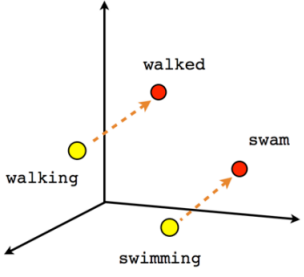


- Классические тематические модели получают на вход матрицу «мешка слов» или tf-idf и строят два типа распределений:
  - слов в кластерах-темах
  - тем в документах
- По факту получается стохастическое матричное разложение
- Строки матрицы «слова-темы» можно использовать в качестве эмбеддингов
- Современные реализации инкрементально обучаются на больших данных

word2vec

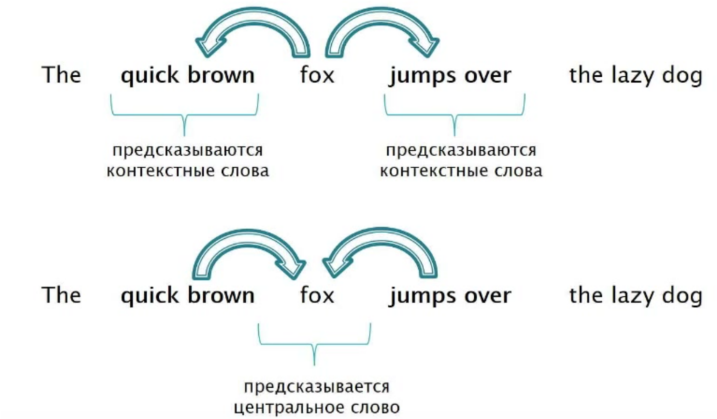
- **word2vec** — группа алгоритмов, предназначенных для получения вещественных векторных представлений слов
- **Идея:** «Слова со схожими значениями разделяют схожий контекст»
- Как правило, в векторном представлении семантически близкие слова оказываются рядом





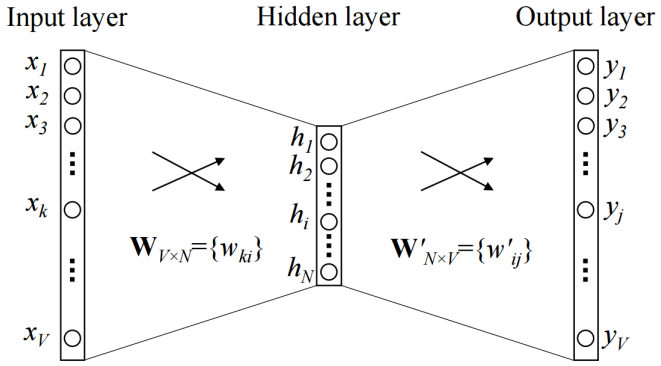
Don't count, predict!

Две модели: **Skip-gram** и **Continuous BOW**



Ссылка на источник картинки

Модель CBOW (единичный контекст)



Модель Skip-gram