Представление текста для анализа нейросетью

Нейросети для анализа текстов

Входные данные для нейросетей

Формат входных данных в нейронную сеть

- Числа

Изображения

- Интенсивность пикселей от 0 до 255

Структурированные данные

- Числовые признаки – без изменений
- Категориальные признаки – one hot encoding (dummy variables)

Текст

- Векторизация – представление текста в виде чисел

Токенизация

Символы

- Буквы, цифры, знаки препинания и т.п.

Слова

Предложения

Методы извлечения признаков из текста

- N-граммы (последовательности слов длиной от 1 до N)
- Мешок слов (bag of words, множество всех слов)

Векторизация

Числовое кодирование

- Для каждого токена свой код (частота, ASCII, UTF-8 и т.п.)

One hot encoding

- Вектор по одному символу на каждый возможный токен
- Все элементы вектора 0, кроме того, который соответствует токену

Плотные векторные представления (embedding)

- Каждому токену сопоставляется вектор
- Размерность вектора ниже, чем у one hot encoding

One hot encoding

Ограничиваем максимальное количество используемых слов

- Oxford 3000
- Merriam Webster 3,000 Core Vocabulary Words

Длина вектора для каждого слова равна максимальному количеству слов

Значения элементов вектора

- 0 – если слово не встречается в тексте
- 1 – если встречается

Разряженные vs плотные векторы

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

0.56	0.93	2.34	1.99

Обучение плотных векторных представлений

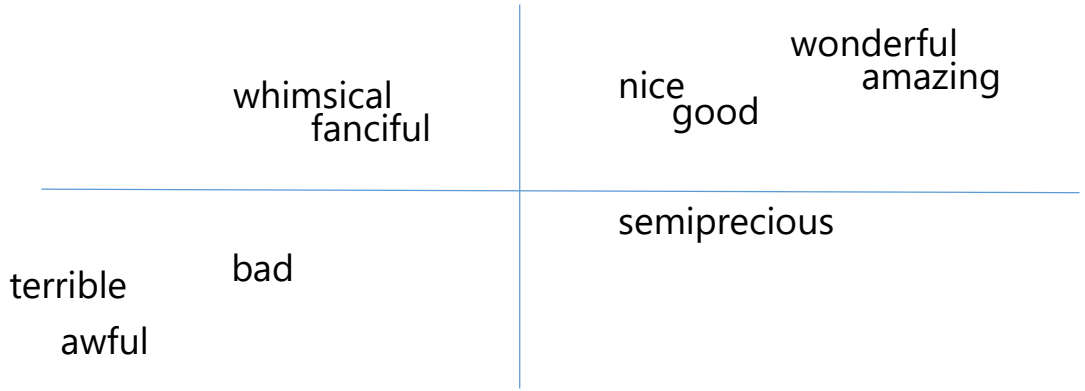
В нейронных сетях плотное векторное представление слов определяется в процессе обучения

- На первом этапе элементы векторов инициализируются случайными числами
- Изменение значений векторов с помощью метода обратного распространения ошибки

Обучение плотных векторных представлений

В нейронных сетях плотное векторное представление слов определяется в процессе обучения

- На первом этапе элементы векторов инициализируются случайными числами
- Изменение значений векторов с помощью метода обратного распространения ошибки



Предварительно обученные векторные представления слов

GloVe (Global Vectors)

- Стэнфордский университет
- https://nlp.stanford.edu/projects/glove/

Word2Vec

- Google
- https://code.google.com/archive/p/word2vec/

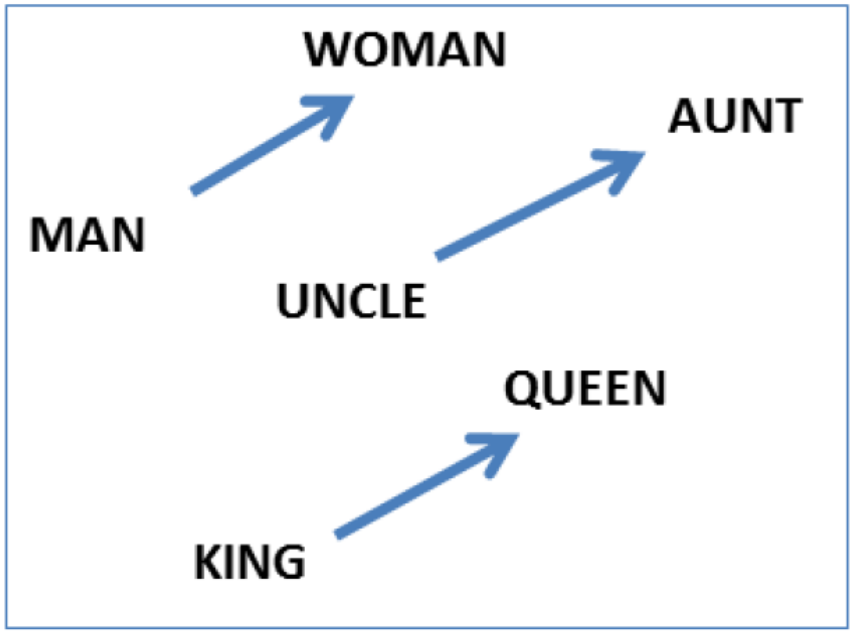
FastText

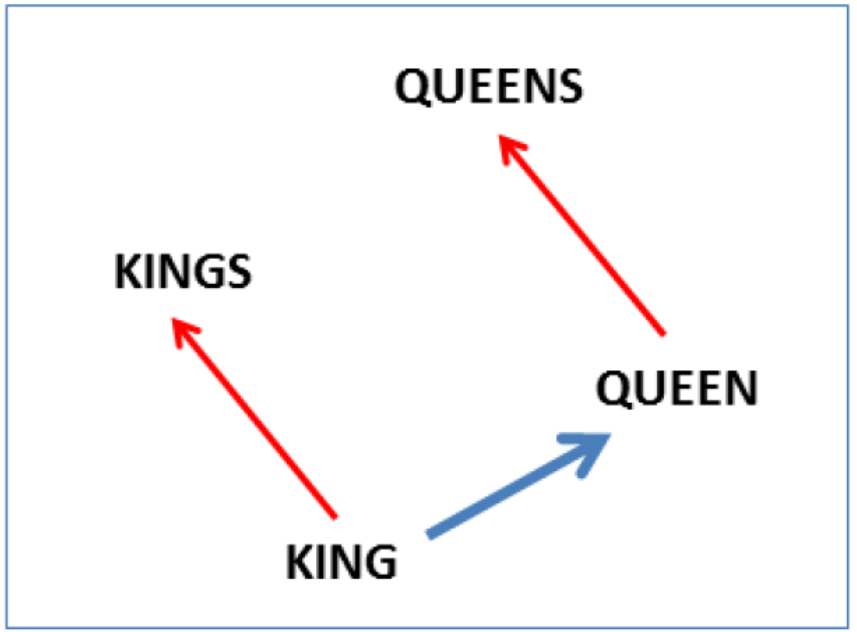
- Facebook
- https://fasttext.cc

Векторные представления слов для русского языка

- RusVectōrēs – https://rusvectores.org
- RUSSE (Russian Semantic Evaluation) – https://russe.nlpub.org/downloads/

word2vec





Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations

Итоги

Формат входных данных в нейронную сеть

- Числа

Тонизация текста

- Символы, слова, предложения

Векторизация текста

- Коды токенов, one hot encoding, плотные векторные представления

Преобразование текста в набор чисел

- Комбинация методов токенизации и векторизации
- Самый сложный этап при анализе текстов нейросетями
- Разные методы подходят для разных типов задач