



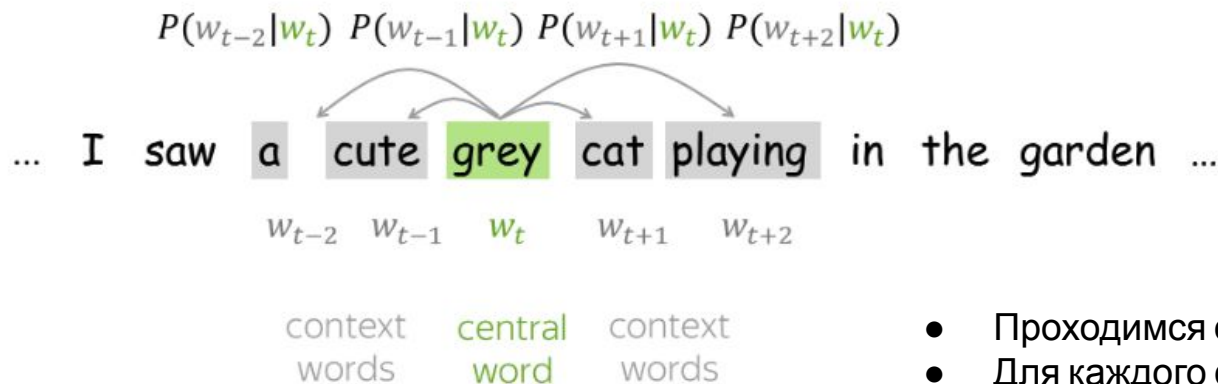
NLP

Word2Vec, FastText



Мария Макарова
Senior Data Analyst, Playrix

Word2Vec. Вспоминаем



- Проходимся окном по тексту
- Для каждого слова – два вектора
- Считаем вероятность встретить контекст возле центрального слова
- Максимизируем эту вероятность

Word2Vec. Обобщаем



- Хотим представить каждое слово в виде вектора, например, размерности 300 (это гиперпараметр)
- Накладываем условия на эти вектора – арифметика, близость векторов слов со схожим контекстом
- Для удобства обучения для каждого слова обучаем дополнительный вектор (когда слово является контекстом другого слова)
- Основная задача – максимизировать правдоподобие (вероятность получить именно такой корпус текстов, который есть у нас по факту)
- Функция потерь = минус логарифм правдоподобия

Word2Vec. Обобщаем



- Чтобы посчитать вероятности для правдоподобия, используем softmax
- Берем весь корпус текстов и проходимся по нему скользящим окном (например, центральное слово, 2 слова слева и 2 слова справа)
- На каждом шаге (одно окно) делаем шаг градиентного спуска для основного вектора центрального слова и для контекстных векторов всех слов. По факту, для окна $2+1+2$ у нас будет 4 шага на окно.
- На выходе модели получаем по два вектора на каждое слово – берем один основной (контекстные нужны только для обучения)

Помним!

- Обучаем и центральные вектора, и контекстные

Если всё-таки что-то не поняли

- Читаем [тут](#)

Word2Vec. Гиперпараметры



- Размер вектора (обычно 300, но варианты 100 и 500 тоже могут подойти)
- Для Negative Sampling
 - Для маленьких датасетов 15-20 наблюдений
 - Для больших датасетов 2-5 наблюдений
 - Окно контекста обычно 5-10 слов

Word2Vec. Советы

- Предобработка текстов
 - Если достаточно большой корпус текстов, то можно не делать лемматизацию/стемминг
 - У нас достаточно контекста, чтобы обучить вектора для слова “цветок” и для слова “цветы” отдельно
- Если у вас специфическая задача с особой лексикой – можно взять предобученную модель и дообучить под свои нужды
- Как обобщить word2vec до текста
 - Усреднить все вектора по тексту
 - Взять максимум
 - Склеить в матрицу и передать нейронной сети
 - Doc2Vec (обучаем вектора для текста целиком – особо не используется, но существует)

Word2Vec. Проблемы

- Не умеем понимать новые слова
- Не понимаем разные контексты одного и того же слова – деньги в банке
- Мы все еще работаем в парадигме мешка слов – не учитываем порядок слов
- Никак не можем учесть опечатки
 - Слова “естественный” и “естественый” будут встречаться в схожих контекстах, иметь схожие вектора, но все равно **разные**
- **Word2Vec может построить вектора только для тех слов, что были в обучающей выборке**
- **Есть FastText**

Word2Vec. Где и как использовать



- Можно строить вектора для токенов из любой последовательности
 - Банковские транзакции (чек за такси рядом с чеком за ресторан)
 - Веб-сессии (страница одного товара рядом со страницей другого товара)
 - Поездки в такси
 - Порядок, в котором пользователь отранжировал фильмы
 - Чеки в супермаркете (молоко рядом с яйцами)

RusVectors

[RusVectōrēs](#)
[Похожие слова](#)
[Визуализации](#)
[Калькулятор](#)
[2D-текст](#)
[Различные операции](#)
[Модели](#)
[О проекте](#)
[EN/RU](#)

Модели

Все модели можно скачать и свободно использовать на условиях лицензии [CC-BY](#) (**жирным** выделены модели, доступные для использования в веб-интерфейсе и API).

Контекстуализированные модели

Постоянный идентификатор	Скачать	Алгоритм	Корпус	Размер корпуса	Таргет	Размерность вектора	RUSSE'18	ParaPhraser	Дата создания
araneum_lemmas_elmo_2048_2020	1.5 Гбайт	ELMo	Araneum (леммы)	около 10 миллиардов слов	Нет	2048	0.91	0.56	Октябрь 2020
tayga_lemmas_elmo_2048_2019	1.7 Гбайт	ELMo	Тайга (леммы)	почти 5 миллиардов слов	Нет	2048	0.93	0.54	Декабрь 2019
ruwikiruscorpora_tokens_elmo_1024_2019	197 Мбайт	ELMo	НКРЯ и Википедия за декабрь 2018 (токены)	989 миллионов слов	Нет	1024	0.88	0.55	Август 2019
ruwikiruscorpora_lemmas_elmo_1024_2019	197 Мбайт	ELMo	НКРЯ и Википедия за декабрь 2018 (леммы)	989 миллионов слов	Нет	1024	0.91	0.57	Август 2019

Word2Vec. Русская литература

```
model.most_similar(u'интеллектуал')
[('моралист', 0.7139864563941956),
 ('теоретик', 0.6941959857940674),
 ('литератор', 0.6819325089454651),
 ('фанатик', 0.6814083456993103),
 ('эрудит', 0.6789889335632324),
 ('демагог', 0.6755205988883972),
 ('марксист', 0.6714329719543457),
 ('рационалист', 0.6712930202484131),
 ('авантюрист', 0.6707291603088379),
 ('революционер', 0.6677388548851013)]
```

```
model.most_similar(u'интеллектуалка')
[('бездельница', 0.6617184281349182),
 ('бунтарка', 0.6578608751296997),
 ('дилетантка', 0.6419748663902283),
 ('скромница', 0.6378872990608215),
 ('вертихвостка', 0.6353027820587158),
 ('профурсетка', 0.6342650055885315),
 ('карьеристка', 0.6335839629173279),
 ('сумасбродка', 0.6256570816040039),
 ('провинциалка', 0.6232886910438538),
 ('пуританка', 0.621334433555603)]
```


FastText



- Улучшение модели Word2Vec
 - Добавляем в модель n-граммы
 - Решаем проблему OOV (Out Of Vocabulary)
- Например, биграммы для слова <eating> – <e, ea, at, ti, in, ng, g>
 - “<” – символ начала слова
 - “>” – символ конца слова
- Итоговый вектор слова – усредняем вектора слова и его n-грамм

FastText

... I saw a cute grey cat playing in the garden ...

$$J_{t,j}(\theta) = -\log P(\text{cute}|\text{cat}) = -\log \frac{\exp u_{\text{cute}}^T v_{\text{cat}}}{\sum_{w \in \text{Voc}} \exp u_w^T v_{\text{cat}}} = -u_{\text{cute}}^T v_{\text{cat}} + \log \sum_{w \in \text{Voc}} \exp u_w^T v_{\text{cat}}$$


- Вот здесь теперь вместо **зеленого вектора cat** будет сумма векторов <c, ca, at, t>, cat
- Серые вектора (контекст) берутся без n-грамм

Natasha



- Библиотека для распознавания именованных сущностей на русском языке
 - Имена
 - Адреса
 - Контакты
 - Даты
 - И многое другое
- Разработана в Яндексе
- Зачем нам? Можем строить качественные эмбединги для задач обработки русского языка
- Что почитать?
 - <https://habr.com/ru/articles/516098/>
 - <https://habr.com/ru/articles/349864/>
 - <https://vc.ru/newtechaudit/358200-instrumenty-dlya-resheniya-ner-zadach-dlya-russkogo-yazyka>