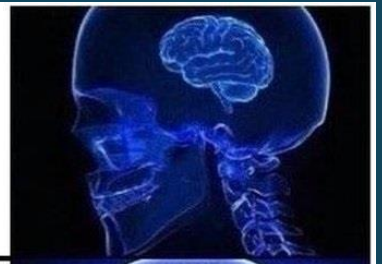
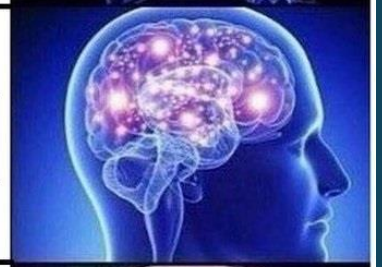
Расширяя границы препроцессинга

@akv17

```
mystem.lemmatize(text)
```



```
re.sub(r'^[\W]|[\W]$', '', text)
```



```
text.strip(punctuation + '«…»')
```

```
text.lower()
```

```
взять_корень(слово)
```

Панч

Панч

- Удачно подобранный препроцессинг сильно упрощает жизнь

Кейс

Кейс

- Нечеткий поиск по базе заказчика

Кейс

- Нечеткий поиск по базе заказчика
- Нужно матчить пары типа «*рекламировать* – *реклама*»

Кейс

- Нечеткий поиск по базе заказчика
- Нужно матчить пары типа «*рекламировать* – *реклама*»
- И не матчить сюда же, например, «*ламинировать*»

Нечеткий поиск

Нечеткий поиск

- Бустим полноту – смягчаем метрику

Нечеткий поиск

- Бустим полноту – смягчаем метрику
- Смягчаем метрику – теряем точность

Нечеткий поиск

- Бустим полноту – смягчаем метрику
- Смягчаем метрику – теряем точность
- Бустим точность – закручиваем гайки метрике

Нечеткий поиск

- Бустим полноту – смягчаем метрику
- Смягчаем метрику – теряем точность
- Бустим точность – закручиваем гайки метрике
- Закручиваем гайки метрике – теряем полноту

Нечеткий поиск

- Бустим полноту – смягчаем метрику
- Смягчаем метрику – теряем точность
- Бустим точность – закручиваем гайки метрике
- Закручиваем гайки метрике – теряем полноту
- It’s a trap!

Где баланс?

Где баланс?

- А что, если взглянуть на препроцессинг иначе?

Где баланс?

- А что, если взглянуть на препроцессинг иначе?
- Например, давайте брать только корни слов

Где баланс?

- А что, если взглянуть на препроцессинг иначе?
- Например, давайте брать только корни слов
- Есть только корни – почти нет шума

Где баланс?

- А что, если взглянуть на препроцессинг иначе?
- Например, давайте брать только корни слов
- Есть только корни – почти нет шума
- Нет шума – подкручиваем гайки нечеткой метрике

Где баланс?

- А что, если взглянуть на препроцессинг иначе?
- Например, давайте брать только корни слов
- Есть только корни – почти нет шума
- Нет шума – подкручиваем гайки нечеткой метрике
- Подкручиваем гайки, не теряя полноту

Где баланс?

- А что, если взглянуть на препроцессинг иначе?
- Например, давайте брать только корни слов
- Есть только корни – почти нет шума
- Нет шума – подкручиваем гайки нечеткой метрике
- Подкручиваем гайки, не теряя полноту
- ggez

Как реализовать?

Как реализовать?

- Грузим в память гигантский морфо словарь?

Как реализовать?

- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно

Как реализовать?

- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно

Как реализовать?

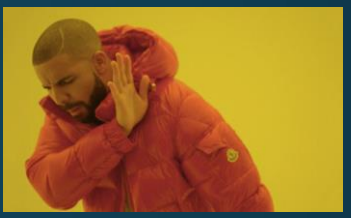
- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно
- Может, стеммер?

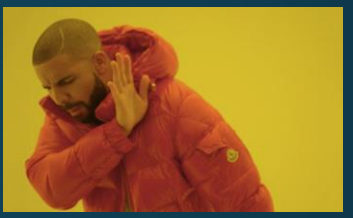
Как реализовать?

- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно
- Может, стеммер?
  - Sounds good, doesn’t work – слишком шумно

Как реализовать?

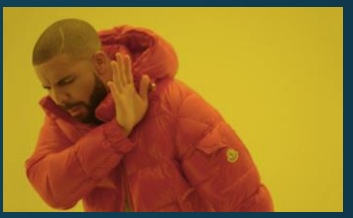
- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно

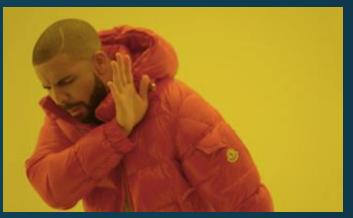


- Может, стеммер?
  - Sounds good, doesn't work – слишком шумно
  - ```
    stemmer = SnowballStemmer('russian')
    stemmer.stem('рекламировать') -> 'рекламирова'
    ```
    ```

Как реализовать?

- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно



- Может, стеммер?
  - Sounds good, doesn't work – слишком шумно
  - ```
    stemmer = SnowballStemmer('russian')
    stemmer.stem('рекламировать') -> 'рекламирова'
    ```


    ```

Как реализовать?

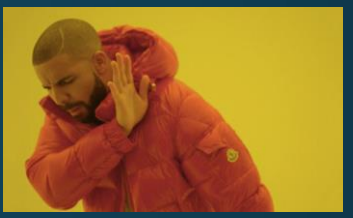
- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно
- Может, стеммер?
  - Sounds good, doesn't work – слишком шумно
  - ```
    stemmer = SnowballStemmer('russian')
    stemmer.stem('рекламировать') -> 'рекламирова'
    ```
- ML!



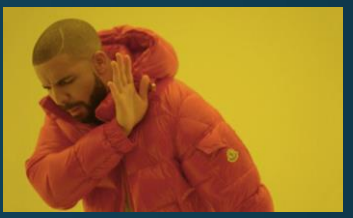

    ```

Как реализовать?

- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно



- Может, стеммер?
  - Sounds good, doesn't work – слишком шумно
  - ```
    stemmer = SnowballStemmer('russian')
    stemmer.stem('рекламировать') -> 'рекламирова'
    ```

- ML!
  - Пусть все сделает моделька
    ```

Как реализовать?

- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно



- Может, стеммер?
  - Sounds good, doesn't work – слишком шумно
  - ```
    stemmer = SnowballStemmer('russian')
    stemmer.stem('рекламировать') -> 'рекламирова'
    ```

- ML!
  - Пусть все сделает моделька
  - `model.predict('рекламировать') -> 'реклам'`
    ```

Как реализовать?

- Грузим в память гигантский морфо словарь?
  - Минус OOV и в целом мало элегантно



- Может, стеммер?
  - Sounds good, doesn't work – слишком шумно
  - ```
    stemmer = SnowballStemmer('russian')
    stemmer.stem('рекламировать') -> 'рекламирова'
    ```

- ML!
  - Пусть все сделает моделька
  - `model.predict('рекламировать') -> 'реклам'`


    ```

Как поставить задачу?

Как поставить задачу?

- Все уже придумано

Как поставить задачу?

- Все уже придумано

Обри	Дрейк	продает	мерч	OVO	только	на	западе
B-PER	I-PER	O	O	B-ORG	O	O	B-LOC

З	А	К	А	С	Т	О	М	И	Т	Ь
PRE	PRE	ROOT	ROOT	ROOT	ROOT	ROOT	ROOT	SUF	SUF	SUF

Чем решать?

Чем решать?

- Любой архитектурой под последовательности

Чем решать?

- Любой архитектурой под последовательности (RNN, CNN, Transformer)

Чем решать?

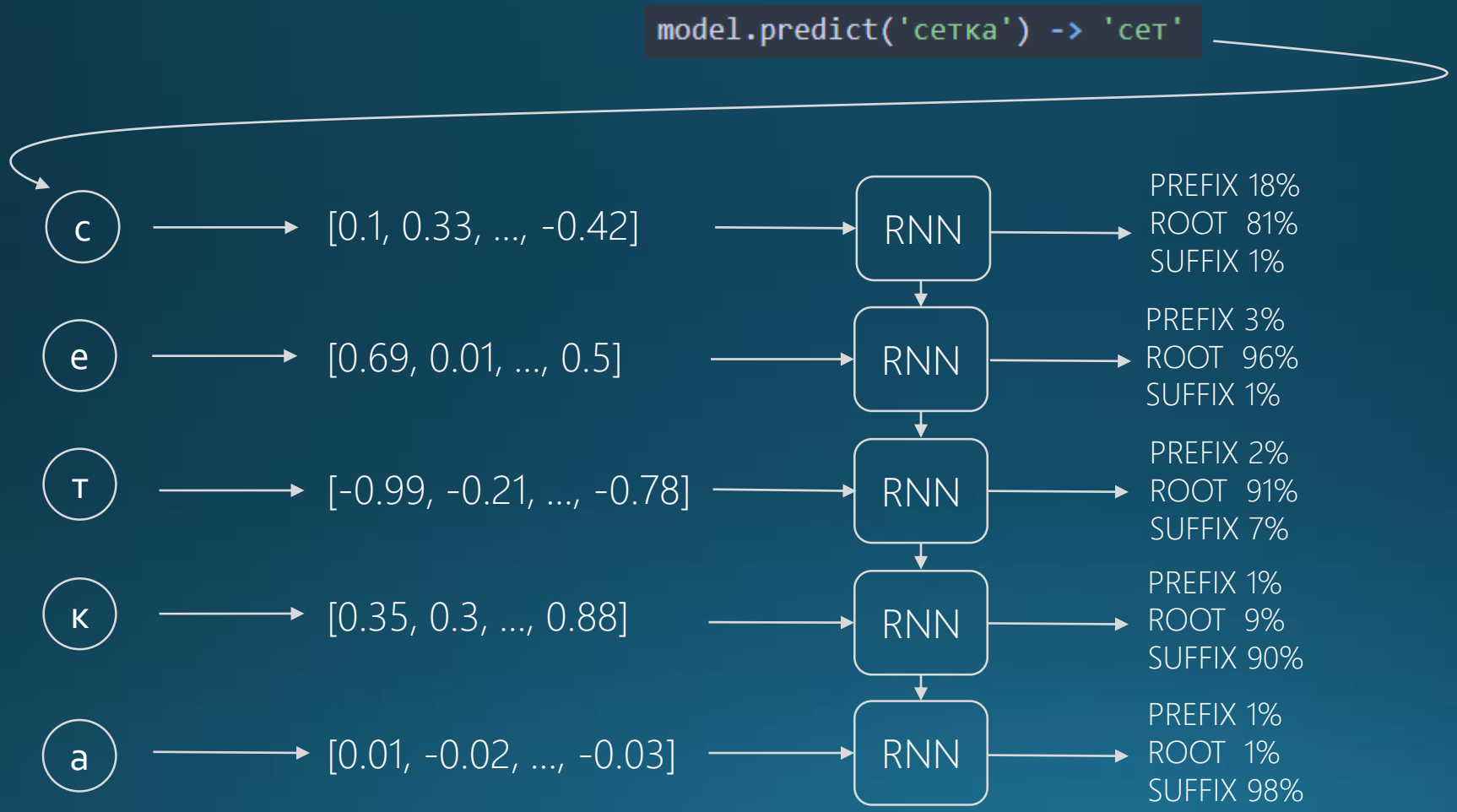
- Любой архитектурой под последовательности (RNN, CNN, Transformer)
- Можно и классическим ML поверх признаков

Чем решать?

- Любой архитектурой под последовательности (RNN, CNN, Transformer)
- Можно и классическим ML поверх признаков (наверное)

Чем решать? RNN

Чем решать? RNN



Пасиб!