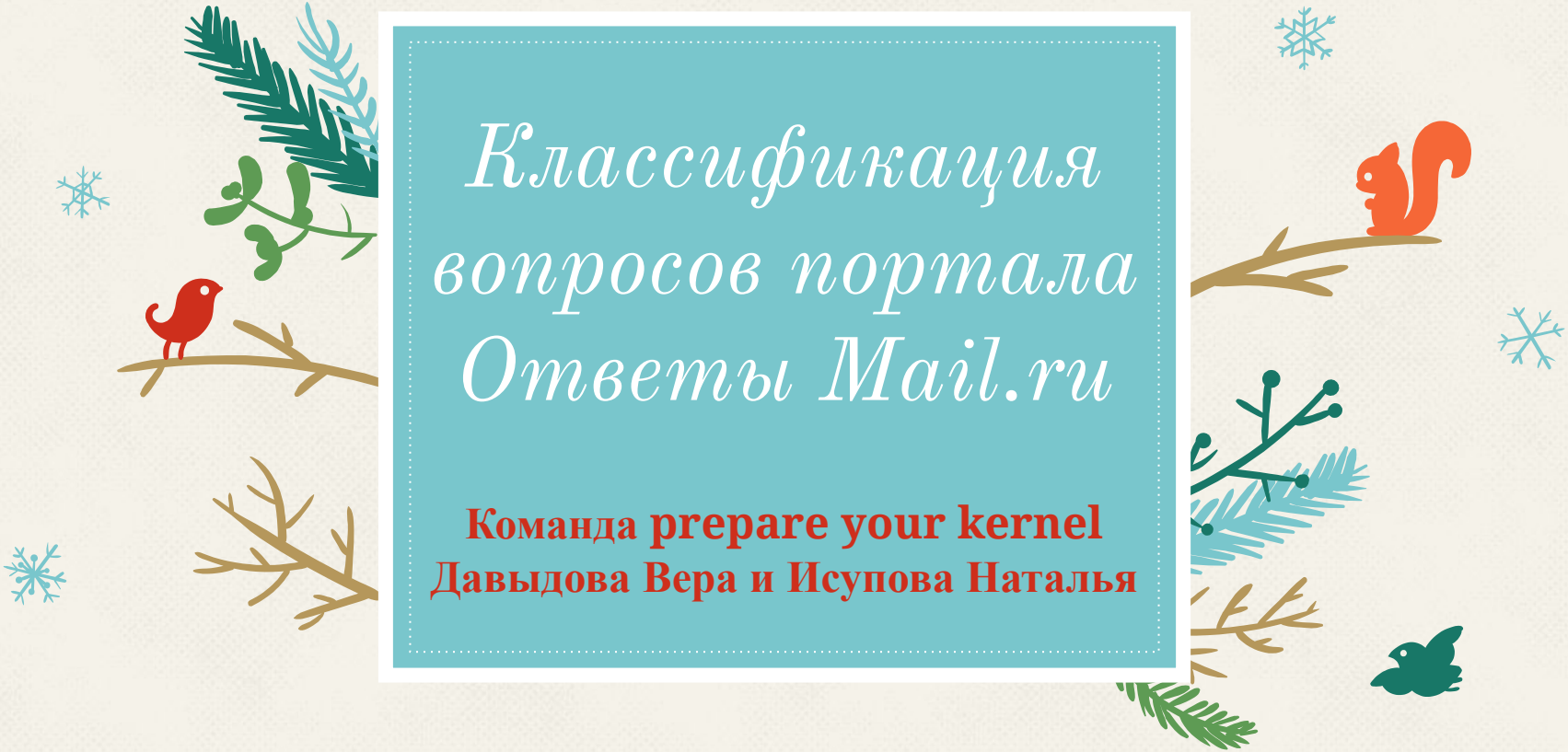


Классификация вопросов портала Ответы Mail.ru

Команда prepare your kernel
Давыдова Вера и Исупова Наталья

Задача

Предсказать классы (28) вопросов для того, чтобы пользователь не ошибался в выставлении категории вопроса.

Как мы хотели это решать?

Домашка № 3 + навороты

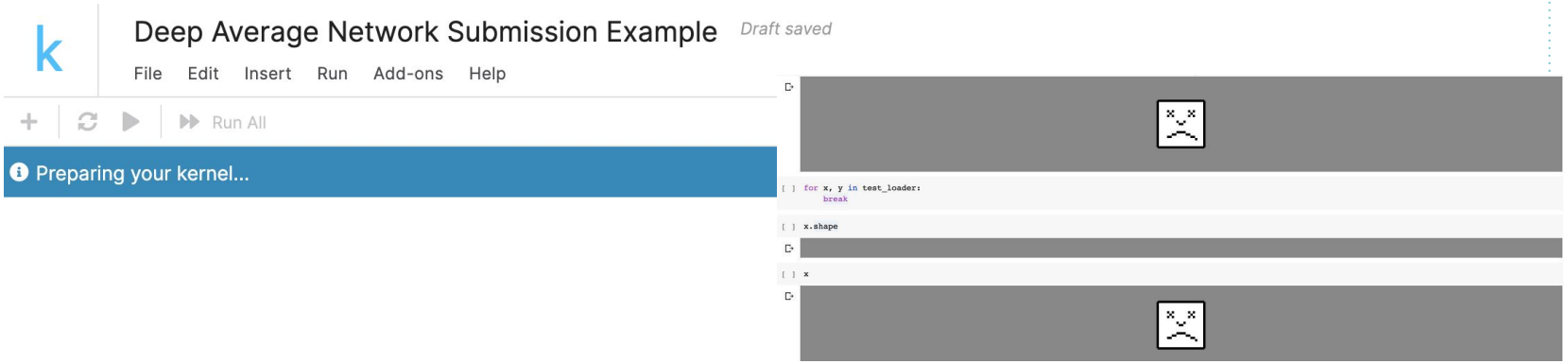
1. pretrained эмбеддинги (прячем часть, запаковываем, подаем в LSTM, распаковываем
2. транспонируем, чтобы передать в cnn
3. добавляем 4 параллельные свёртки и пулинги к ним, конкатенируем результаты
4. два линейных слоя на выходе

Получилось ли у нас?

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Sem Sorokin			0.64554	33	4d
2	memy_pro_kotow			0.64453	18	14h
3	DEEP_кусь			0.62301	4	15h
4	Marina			0.61823	14	11h
5	prepare your kernel			0.61714	3	13d
6	Vera Davydova			0.57535	1	9h

Проблемы!

Конечно же, все оказалось не так просто.




```
AssertionError:
The NVIDIA driver on your system is too old (found version 10000).
Please update your GPU driver by downloading and installing a new
version from the URL: http://www.nvidia.com/Download/index.aspx
Alternatively, go to: https://pytorch.org to install
a PyTorch version that has been compiled with your version
of the CUDA driver.
```

```
Мы не знаем 3.02 % слов в датасете
Количество неизвестных слов 107889 из 328896, то есть 32.80 % уникальных слов
в словаре
В среднем каждое встречается 2.08 раз

Топ 5 невошедших слов:
??? с количеством вхождениий - 8604
!!! с количеством вхождениий - 6976
?) с количеством вхождениий - 6613
?? с количеством вхождениий - 6327
"? с количеством вхождениий - 4581
```

Обучим свой фасттекст!

Данные: таблица с неразмеченными данными (2535443 ответов мэйл.ру)

```
from gensim.models import FastText
```

```
%%time
fast_text = FastText([text.split(' ') for text in df['norm']], size=300,window=8, min_n=2,
                                    max_n=8, sorted_vocab=1,negative=10, iter=10)
```

```
CPU times: user 1h 35min 19s, sys: 29.6 s, total: 1h 35min 49s
Wall time: 33min 14s
```

Что получилось?

Неизвестных слов не осталось.

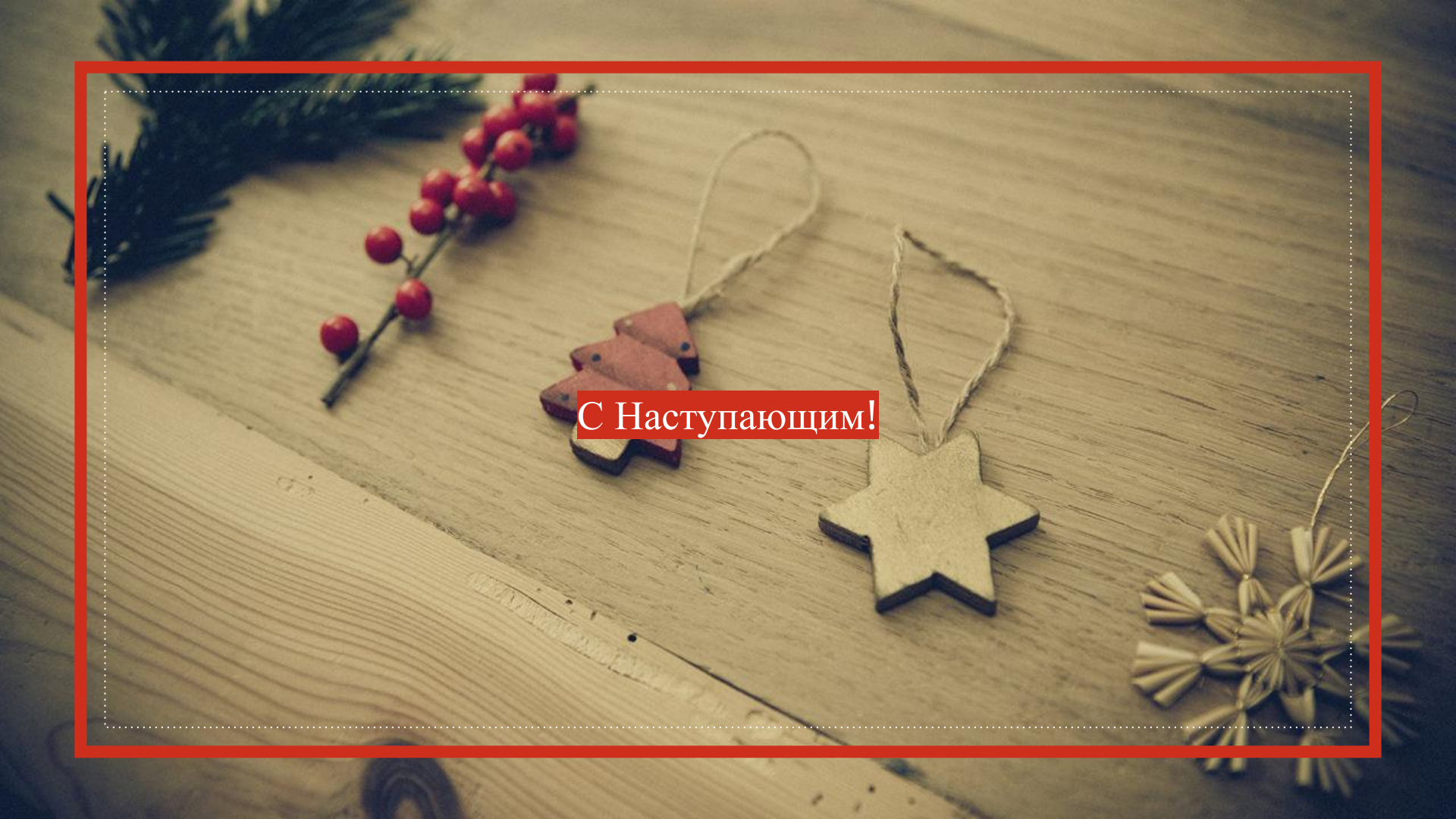
Стало хуже.

61 -> 57

```
unknown_words = []
for word in tqdm(train_list):
    vector = model.wv[word]
    comparison = np.array_equal(vector, zero_vector)
    if comparison:
        unknown_words.append(word)
```

```
100%|██████████| 135229/135229 [00:11<00:00, 11509.84it/s]
```

Видимо, дело не в эмбеддингах.


С Наступающим!