



InClass Prediction Competition

## **Анализ веб-документов**

Сможете ли Вы найти в группе документов те, которые связаны друг с другом?

25 teams · 4 hours to go
















# `\varEpsilon`



Буллат Валиахметов



Андрей Коновалов

#	Team Name	Notebook	Team Members	Score ?	Entries	Last
1	[ML1-sphere] Yangle rocks!		 	0.74544	11	1h
2	[ML1-sphere] SuperPuperKara...		   	0.74464	130	14m
3	[ML1-sphere]\varEpsilon		 	0.73223	30	2m
<p>Your Best Entry </p> <p>Your submission scored 0.73223, which is an improvement of your previous score of 0.72288. Great job!</p> <p> <a href="#">Tweet this!</a></p>						
4	[ML1-sphere] zladimir			0.73054	13	1d
5	[ML1-sphere] GMG		   	0.72748	26	1h

По состоянию на 17:57 МСК 21.12.2020

# Основная идея

Измеряем расстояния между документами своей группы следующими метриками:

Bag-of-words 
$$\frac{|file_i \cap file_j|}{1 + |file_i \cup file_j|}$$

text  $\rightarrow$  Bag-of-words  $\rightarrow$  cosine similarity

TF-IDF

text  $\rightarrow$  TF-IDF  $\rightarrow$  cosine similarity

Doc2vec

text  $\rightarrow$  doc2vec  $\rightarrow$  cosine similarity

Берём топ- $k$  наименьших расстояний до соседей в качестве признаков

# Формирование документов

Заголовки + h1-h6 tags

SVM: val\_score = 0,68

LightGBM: val\_score = 0,73

Заголовки + h1-h6 tags + paragraphs

SVM: val\_score = 0,71

LightGBM: val\_score = 0,78

Еще было: LogReg, RandomForest (сразу не зашло, дальше не смотрели, тк оставалось меньше 20 часов)

# Предобработка

- Оставим слова длины хотя бы 4 и лемматизируем (pymorphy2)
- Числа

# Вопрос лемматизации

```
import pymorphy2
```

Light GBM

SVM

val\_score = 0,77

public score = ##

val\_score = 0,7937

public score = 0,7228

```
# import pymorphy2
```

Light GBM

SVM

val\_score = 0,787

public score = 0,70716

val\_score = 0,791

public score = 0.70868

# Итог итогов

Лемматизация заголовков ухудшила результат (около 0.62 на public)

Решение: не лемматизировать заголовки и h-теги

Итог:

SVM:

val\_score = 0,7844

public score = 0,73223

Это и есть наша финальная модель :)



# За нехваткой времени не успелось

- Выделить ядро группы –  $k$  документов, у которых сумма расстояний друг до друга минимальна среди всех
- Посчитать для каждого документа расстояние до ядерных документов