Skoltech

Skolkovo Institute of Science and Technology

Multilabel text classification

Sofia Medvedeva
Tatiana Smirnova
Almir Dzhumaev
Adel Galimov

Moscow 2014

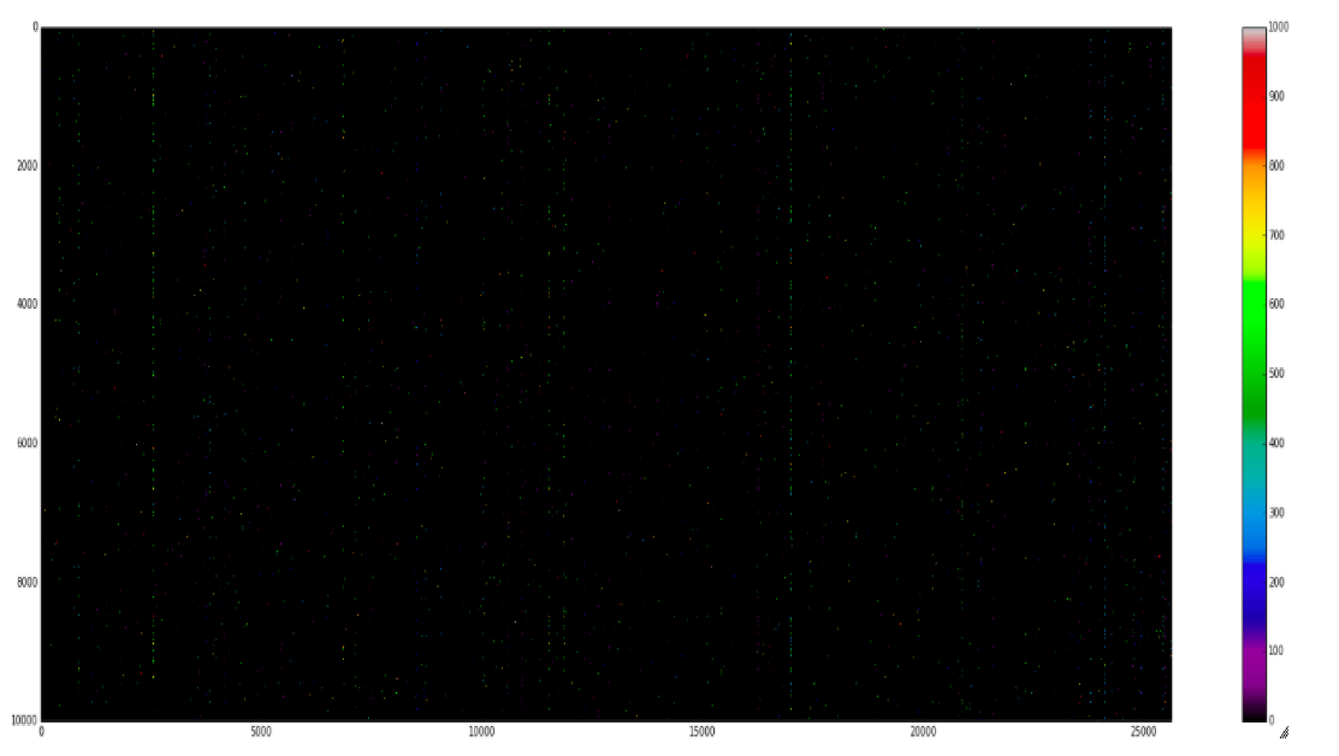
Task and goal

- We have data about scientific terms importances in different articles
- Each article may belong to several topics
- Goal: to build a classifier for topic prediction (and win Kaggle competition!)

Roles in the team

- Tatiana, Adel - data visualizations, preprocessing
- Sofia, Almir - classificators’ training

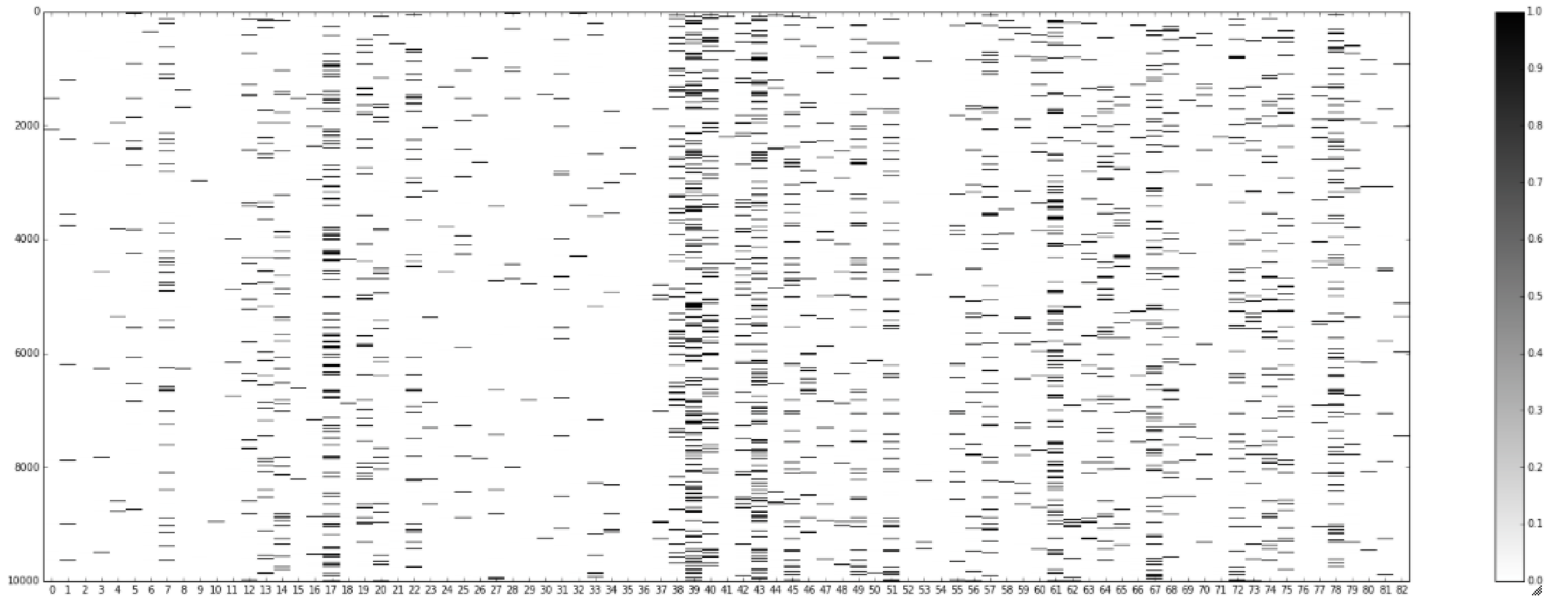
Data Structure. X_train matrix. 10,000x25,000 sparse matrix



N of the document

N of the term

Labels Structure. y_train. 10,000x83 sparse matrix

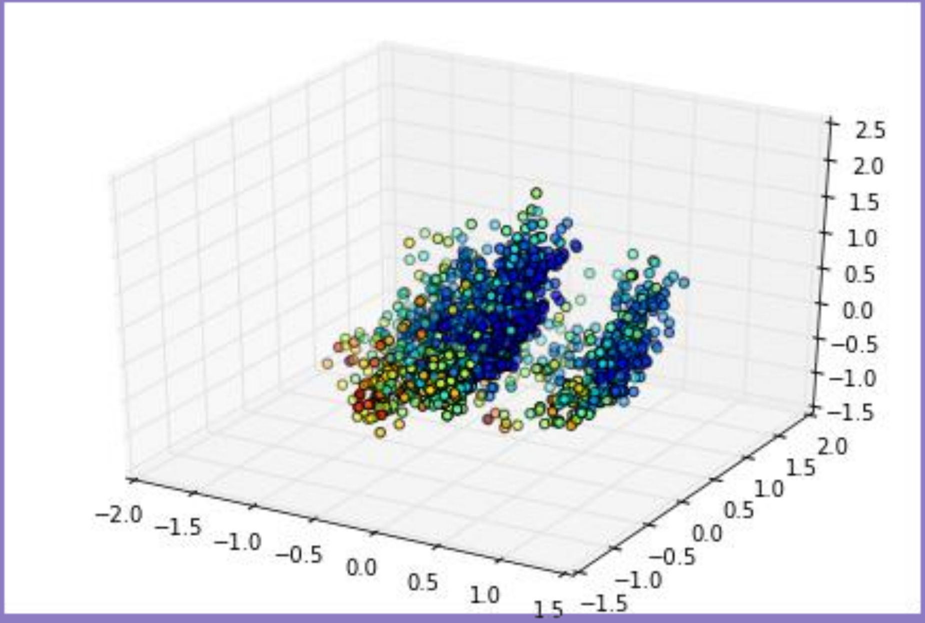


N of the document

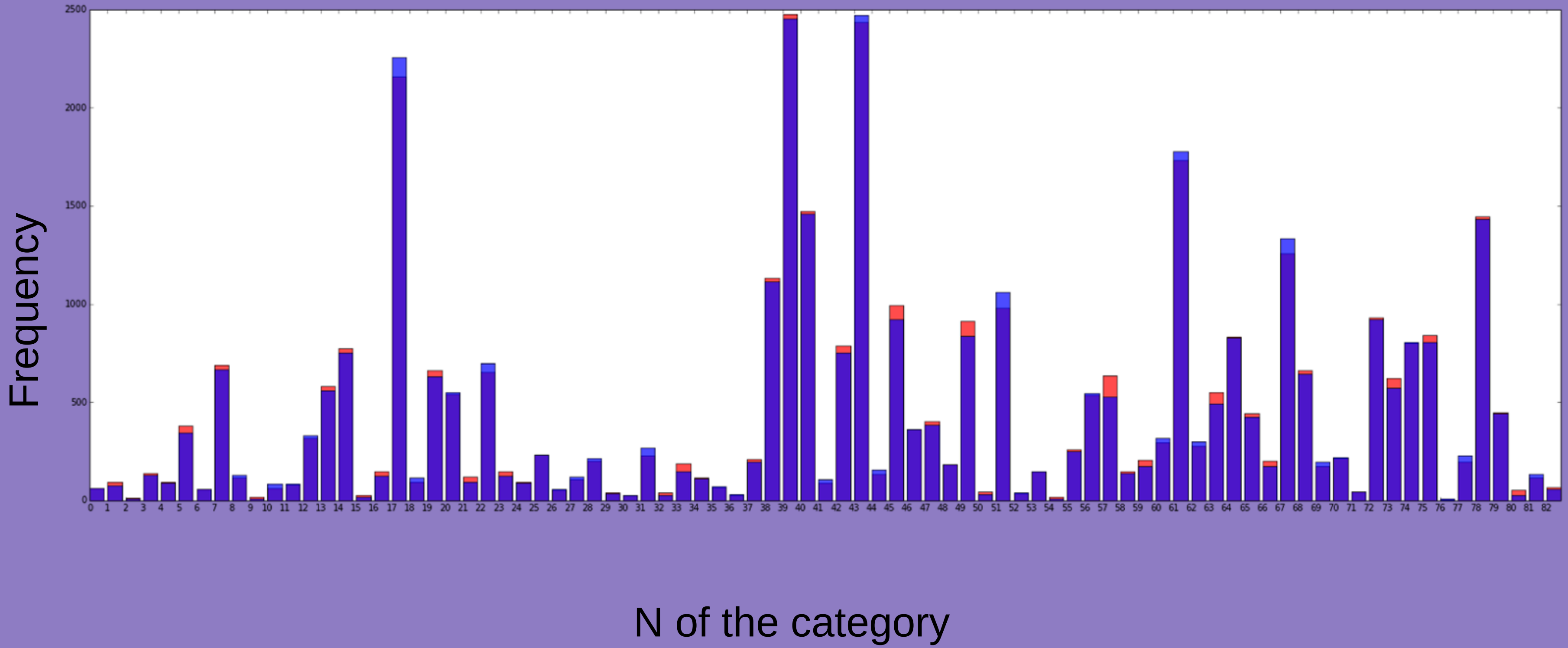
N of the category

Data Structure. Labels visualization

PCA with 4 dimensions



Train and test labels structure.
Categories frequencies.



ML methods used

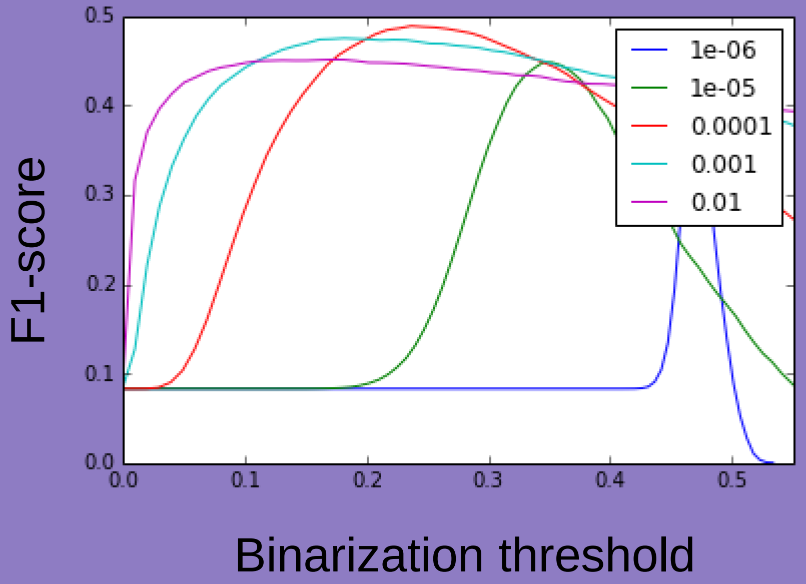
- Random Forest
- SVM
- Naive Bayes
- Logistic regression

Strategies used to improve the result

- Combination of methods (blending, pipeline)
- Binarization
- Feature selection

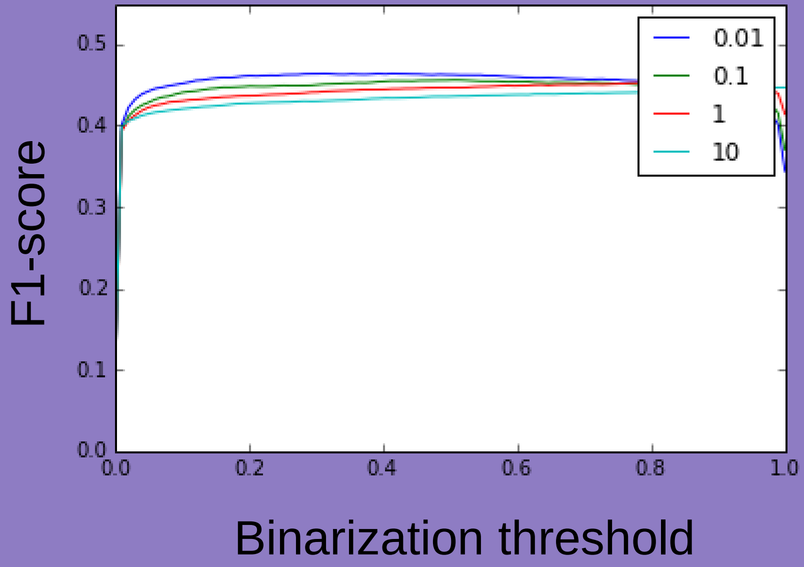
Binarization

Binarization can help!



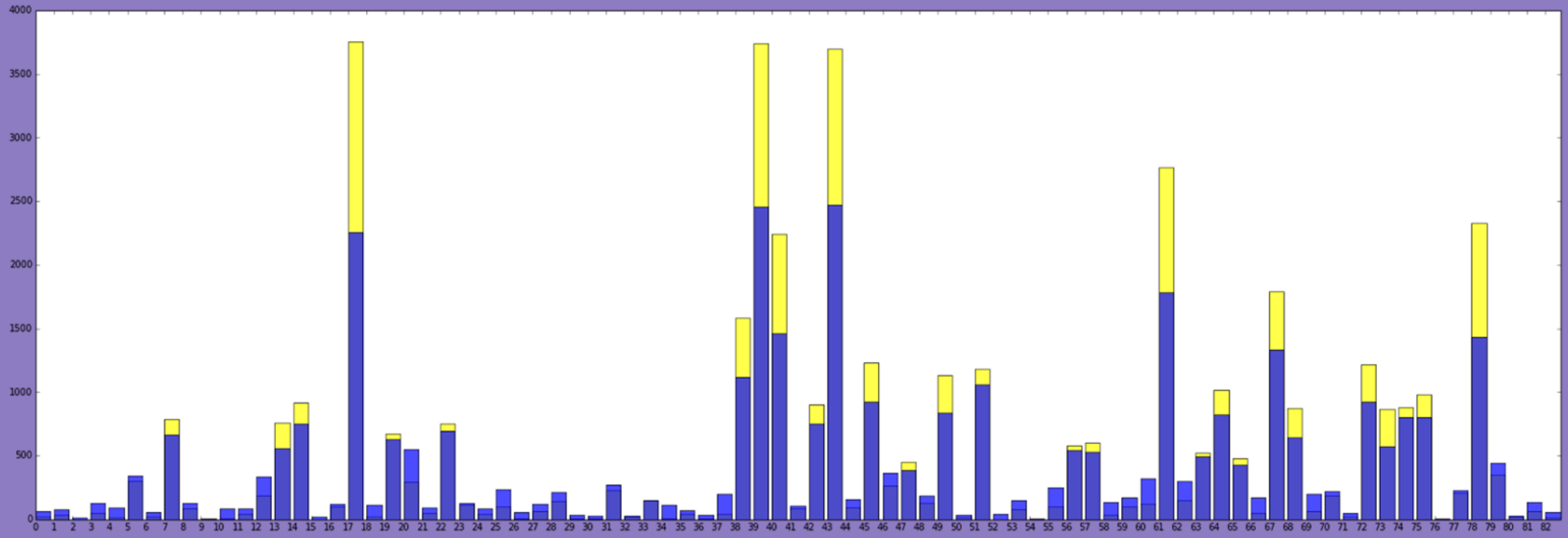
Feature selection

- l1 regularization
- K-best



Final result - 1st place

Best result achieved by logistic regression with optimized parameters’ values - 0.51 f1-score


4000
3500
3000
2500
2000
1500
1000
500
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82