Natural Language Processing course (stream 6, spring 2024)

Финальный проект

май 2024 **Козырев Дмитрий**

Тема

Классификация тем научных диссертаций по названиям

Оглавление:

1. **Постановка задачи:** создать модель классификации научных диссертаций по областям науки основываясь на названии темы диссертации.

Для экспериментов были выбраны четыре специальности ВАК:

- **математика ВАК 01.01.00** (Математический анализ, Дифференциальные уравнения, Математическая физика, Геометрия и топология, Теория вероятностей и математическая статистика, Математическая логика, алгебра и теория чисел, Вычислительная математика, Математическая кибернетика, Системный анализ и автоматическое управление)
- **физика ВАК 01.04.00** (Техника физического эксперимента, физика приборов, автоматизация физических исследований, Теоретическая физика, Радиофизика, Физическая электроника, Оптика, Акустика, Физика твердого тела, Физика и химия плазмы, Физика низких температур, Физика полупроводников и диэлектриков, Физика магнитных явлений, Электрофизика, Теплофизика и молекулярная физика, Физика ядра и элементарных частиц, Химическая физика, в том числе физика горения и взрыва, Кристаллография, физика кристаллов, Физика полимеров, Физика пучков заряженных частиц и ускорительная техника, Лазерная физика, Сверхпроводимость, Физика высоких энергий)
- **химические науки ВАК 02.00.00** (Неорганическая химия, Аналитическая химия, Органическая химия, Физическая химия, Электрохимия, Химия высокомолекулярных соединений, Химия элементоорганических соединений, Радиационная химия, Биоорганическая химия, химия природных и физиологически активных веществ, Коллоидная и мембранная химия, Нефтехимия, Радиохимия, Химическая кинетика и катализ, Химия и технология композиционных материалов, Квантовая химия, Химия, физика и технология поверхности, Химия высокочистых веществ, Хроматография, Химия твердого тела)
- **биологические науки ВАК 03.00.00** (Радиобиология, Биофизика, Молекулярная биология, Биохимия, Ботаника, Вирусология, Микробиология, Зоология, Энтомология, Ихтиология, Эмбриология, гистология и цитология, Физиология растений, Физиология человека и животных, Антропология, Генетика, Экология, Гидробиология, Паразитология, Гельминтология, Криобиология, Биотехнология, Микология, Клеточная биология, Молекулярная генетика, Почвоведение)

Такие специальности были выбраны сознательно, поскольку существует большое количество тем находящихся на стыке этих наук:

- физика-математика (например, ВАК РФ 01.01.03 математическая физика, 229 диссертаций);
- физика-химия (например, ВАК РФ 01.04.17 химическая физика, 628 диссертаций; ВАК РФ 02.00.04 физическая химия, 4280 диссертаций; физика и химия плазмы ВАК РФ 01.04.08, 688 диссертаций; физика полимеров ВАК 01.04.19, 57 диссертаций);
- физика-биология (например, ВАК РФ 03.00.02 биофизика, 902 диссертации);
- биология-химия (например, ВАК РФ 03.00.04 биохимия, 1763 диссертации; ВАК РФ 02.00.10 биоорганическая химия, 596 диссертаций);
- биология-математика (например, ВАК РФ 03.00.28 биоинформатика, 36 диссертаций)

Интересно было посмотреть качество метрик для тем находящиеся на стыке разных наук и как они влияют на общее качество метрик.

Модель предполагает возможность неограниченного расширения количества классов при необходимости.

2. **Ограничения налагаемые на модель:**

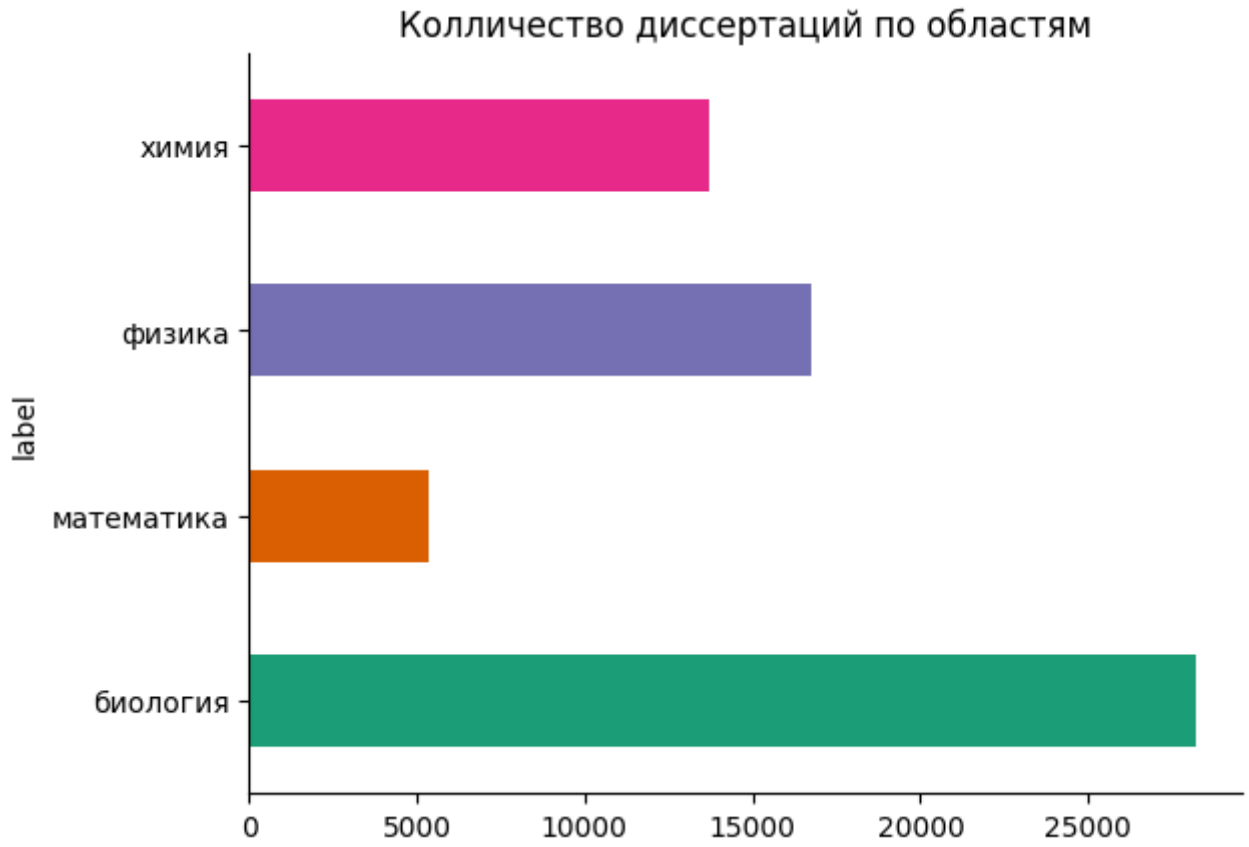
- Брались только русскоязычные темы диссертаций
- Диссертации периода до 1991 года оцифрованы в наименьшей степени, поэтому нельзя делать выводы о количестве диссертаций за разные периоды.
- Брались только четыре специализации ВАК (указаны выше) из 24
- Отсутствие пред обученных моделей word2vec для естественно научных специализированных текстов. Поэтому предполагается обучить word2vec модель на специально подготовленных текстах
- Большое количество уникальных токенов свойственных только выбранным дисциплинам и не встречающихся в классических текстах

3. **Сбор данных:**

- Данные были собраны методом веб-скрейпинга с помощью библиотеки beautifulsoup из источников, находящихся в свободном доступе:
www.dissercat.com – электронная библиотека диссертаций
- Собирались только наименования диссертаций, автор и год по темам ВАК 01.01.00, 01.04.00, 02.00.00, 03.00.00 включая все под темы указанные ранее.
Оглавление и содержание диссертаций не скачивались, хотя такая возможность предусматривалась, но привадила с частым банам, борьба с которыми занимала слишком много времени.
- Итого было скачано тем диссертаций за период с 1973 по 2023 годы:

- математика ВАК 01.01.00, 4379 диссертаций (файл mathematic.csv)
- физика ВАК 01.04.00, 13411 диссертаций (файл physics.csv)
- химические науки ВАК 02.00.00, 11057 диссертаций (файл chemistry.csv)
- биологические науки ВАК 03.00.00, 22655 диссертаций (файл biology.csv)

всего



- таким образом, классы являются частично разбалансированным
- **качество полученных данных очень высокое**: практически отсутствуют не распознанные символы и полностью отсутствуют лишние тестовые данные.

4. **Обучающая выборка (train dataset):**

   a. Для борьбы с разбалансировкой классов и для обучения Word2vec модели в обучающую выборку были добавлены следующие данные, которые тоже соответствующим образом были размечены по классам (файл textbook.csv):

учебные пособия высшей школы: «Курс общей химии», «Курс общей биологии», «Курс общей физики», «Курс теоретической физики», «Курс высшей математики», «Названия химических элементов», «Физические постоянные», «Основные термины по физике», «Основные термины по высшей математике», «Основные термины по химии», «Основные термины по биологии»
- Поскольку учебные пособия содержат большое количество формул и переводились из формата PDF в формат TXT, то **качество этих данных достаточно низкое**

   b. В обучающую выборку также добавлены названия и выборочные содержания научных статей по соответствующим специальностям ВАК взятые с сайта: https://cyberleninka.ru/article (файл train.csv)

- математика ВАК 01.01.00, 51 статья
- физика ВАК 01.04.00, 50 статей
- химические науки ВАК 02.00.00, 50 статей
- биологические науки ВАК 03.00.00, 49 статей

отбор статей производился в ручную
- **качество данных достаточно высокое**, но содержит значительное количество не распознанных символов (формул)

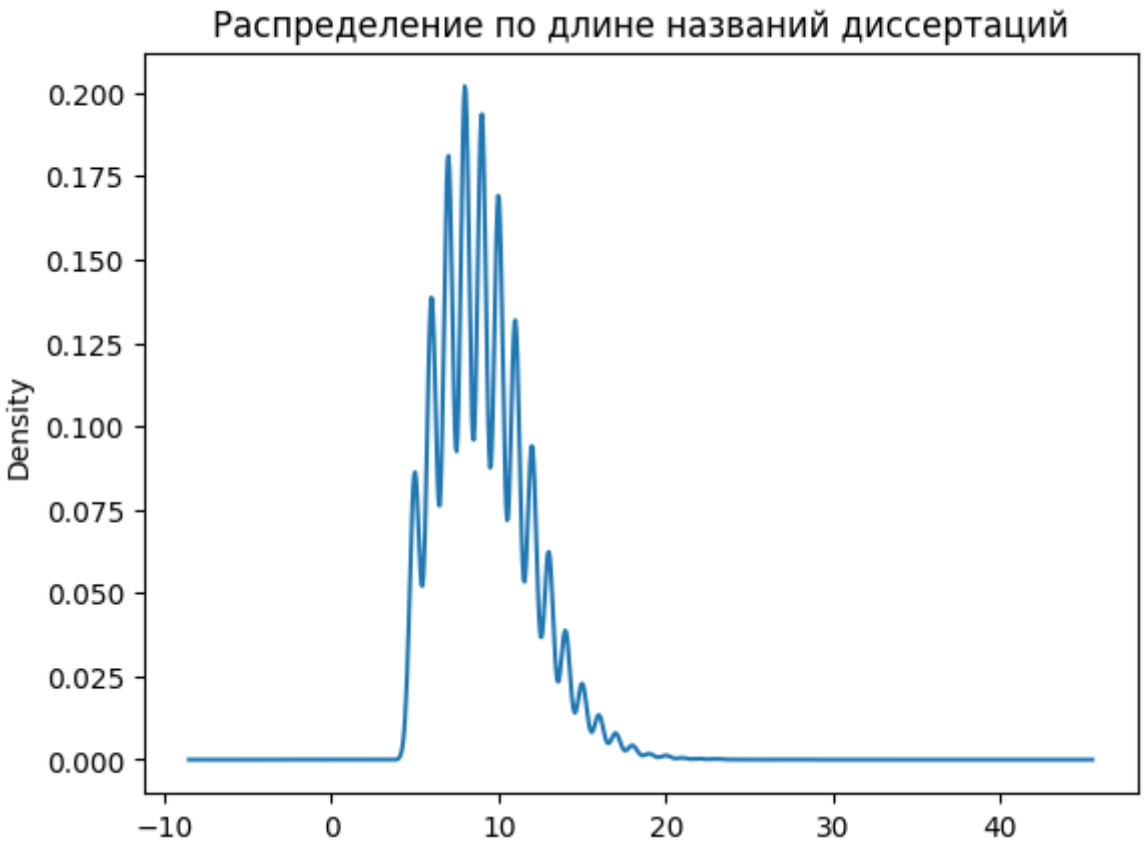
- Для увеличения количества данных, все учебные пособия и статьи разделены на токены длиной 600 знаков (гиперпараметр sent_length в программе).

4.3. В обучающую выборку также добавлены 70-80% датасета с названиями диссертаций (data dataset)

5. **Тестовая выборка (test dataset):**
Тестирование проводилось на 30-20% датасета с названиями диссертаций.

6. **Предобработка данных:**
Были опробованы разные способы предобработки данных. Наилучшие результаты метрик качества были получены при следующей предобработке данных:
- перевод в нижний регистр
- удаление стоп слов с помощью библиотеки nltk.stopwords. Список стандартных стоп слов был расширен за счет анализа наиболее часто встречающихся в корпусе слов. (Например, ‘вак’, ‘iii’, ‘xxx’)
- удаление всех небуквенных символов, в том числе цифр, с целью удаления формул.
- выделение основы слов с помощью библиотеки SnowballStemmer(“Russian”)
- удаление всех токенов короче 3 символов
- удаление всех множественных пробелов
- удаление всех строк с количеством токенов (слов) менее 4

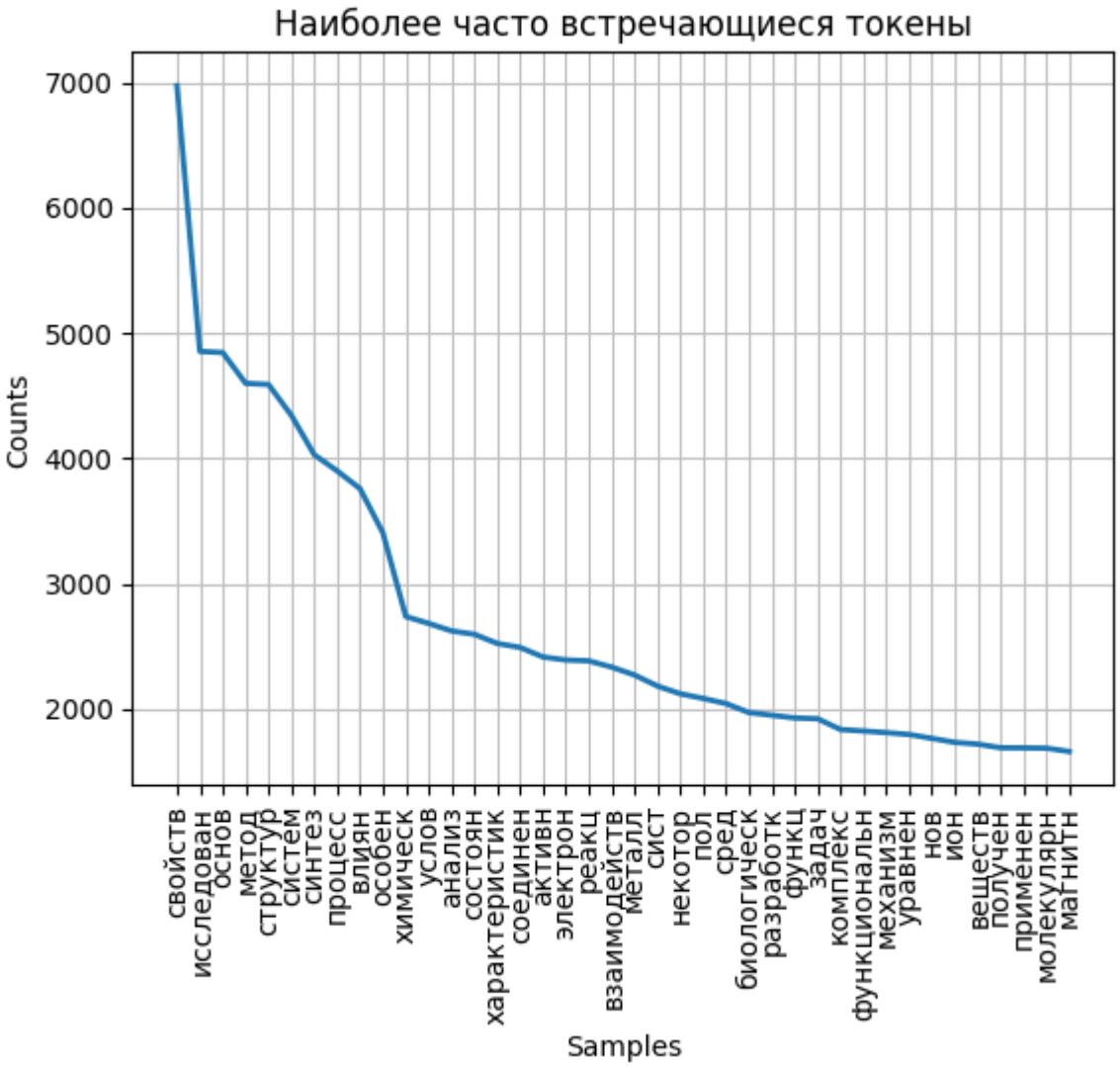
7. **Анализ данных**:



Анализ данных названий диссертаций:
максимальная длина строки, слов: 32
минимальная длина строки, слов: 5
средняя длина строки, слов: 9.16
всего слов: 586109
Всего строк: 63983

Анализ данных обучающих текстов (учебных пособий и научных статей, после разбиения на фрагменты по 600 знаков):
максимальная длина строки, слов: 2666
миниимальная длина строки, слов: 7
средняя длина строки, слов: 411.98
всего слов: 130188
Всего строк: 316

Всего строк в двух датафреймах: 64299
Всего слов (после очистки): 716297
Количество уникальных токенов (слов): 56905 (8%)



100 наиболее часто встречающихся токенов (слова, частота):

[('свойств', 6975), ('исследован', 4854), ('основ', 4845), ('метод', 4599), ('структур', 4590), ('систем', 4343), ('синтез', 4029), ('процесс', 3900), ('влиян', 3758), ('особен', 3405), ('химическ', 2737), ('услов', 2682), ('анализ', 2622), ('состоян', 2595), ('характеристик', 2522), ('соединен', 2490), ('активн', 2414), ('электрон', 2390), ('реакц', 2383), ('взаимодейств', 2333), ('металл', 2270), ('сист', 2182), ('некотор', 2120), ('пол', 2083), ('сред', 2042), ('биологическ', 1970), ('разработк', 1949), ('функц', 1927), ('задач', 1921), ('комплекс', 1835), ('функциональн', 1823), ('механизм', 1810), ('уравнен', 1795), ('нов', 1764), ('ион', 1732), ('веществ', 1718), ('получен', 1689), ('применен', 1688), ('молекулярн', 1686), ('магнитн', 1659), ('оценк', 1640), ('динамик', 1615), ('физик', 1597), ('структурн', 1578), ('формирован', 1540), ('раствор', 1523), ('кислот', 1512), ('использован', 1512), ('област', 1508), ('действ', 1493), ('различн', 1480), ('производн', 1460), ('сследован', 1454), ('рол', 1443), ('моделирован', 1439), ('теор', 1432), ('растен', 1416), ('вид', 1365), ('экологическ', 1351), ('воздейств', 1342), ('излучен', 1341), ('материал', 1340), ('пример', 1336), ('эффект', 1329), ('ген', 1327), ('модел', 1297), ('человек', 1274), ('развит', 1265), ('элемент', 1248), ('вод', 1241), ('белк', 1238), ('решен', 1226), ('тип', 1223), ('определен', 1211), ('групп', 1189), ('генетическ', 1188), ('котор', 1178), ('почв', 1157), ('фактор', 1155), ('оптическ', 1144), ('поверхн', 1125), ('энерг', 1124), ('связ', 1123), ('кристалл', 1117), ('изменен', 1114), ('экспериментальн', 1112), ('органическ', 1111), ('строен', 1101), ('организм', 1076), ('оксид', 1076), ('частиц', 1053), ('физиологическ', 1005), ('фазов', 978), ('регуляц', 978), ('изучен', 976), ('квантов', 966), ('тверд', 957), ('водн', 945), ('электрическ', 944), ('род', 937)]

8. Векторизация и эмбединг данных

Были опробованы несколько типов векторизации:

a. Bag of Word при помощи CountVectorizer из пакета Sklearn.

При этом пробовалась векторизация как по униграммам, так и биграммам (гиперпараметр ngram_range=(1,2).

Первые 50 значений вектора слов:

array(['aabb', 'aabb aabb', 'aabb аав', 'aabb аавв', 'aad',
'aad иммортализова', 'aaptos', 'abam', 'abc', 'abc abc',
'abc рецессивн', 'abc собствен', 'abc такж', 'abca',
'abca атеросклероз', 'abcc', 'abcc транспортер', 'abcg',
'abcg аполипопротеин', 'abcl', 'abcl bcl', 'abcosc',
'abcosc absinc', 'abcosc acosc', 'abcosc абсолютн', 'abcosc найд',
'abd', 'abd drosophila', 'abdominal', 'abdominal drosophila',
'abdr', 'abh', 'abies', 'abies hill', 'abies holophylla',
'abies karst', 'abies mill', 'abies nephrolepis', 'abies pungens',
'abies sibirica', 'abietinus', 'abietinus tristis', 'abietoideae',
'abietoideae rich', 'abl', 'abl развит', 'abo', 'abo тверд',
'abortus', 'abortus francisella'], dtype=object)

b. Bag of Words с использованием хэш функции при помощи HashingVectorizer (ngram_range=(1, 2), n_features=8388608) из пакета Sklearn

c. TD/IDF трансформация Bag of Words с помощью TfidfTransformer (sublinear_tf=True) из пакета Sklearn

d. Word2Vec модель обученная на текущем корпусе с использованием w2v_model из пакета Gensim.

Для этого весь корпус был разбит на маркированные предложения длинной не более 600 знаков:

Количество предложений: 51970

Общее колличестdo слов: 627066

Средняя длина предложения, слов: 12.06

Параметры обучения модели:

Word2Vec(sentences, window=10, min_count=1, workers=4, epochs=30, sg=0, negative=5)

Общее количество слов в обученном словаре: 52744

Обученная модель сохранялась в файл для возможности косинусной близости слов:

Например, косинусная близость слов к слову "клетка":

[('клеток', 0.5930703282356262), ('органоид', 0.4957215189933777),
('цитоплазм', 0.49116936326026917), ('клеточн', 0.47668787837028503),
('пластид', 0.47269803285598755), ('вакуол', 0.46297383308410645),
('митоз', 0.45710527896881104), ('нуклеопротеидн', 0.4554140269756317),
('выполня', 0.4524175524711609), ('животнойклетк', 0.45162448287010193)]

e. Word2Vec модель предобученная ruscorpora_upos_cbow_300_20_2019 (модель для русского языка, созданную на основе Национального корпуса русского языка НКРЯ)

f.

9. Выбор модели

9.1. Были обучены следующие классические модели с параметрами по умолчанию на векторизации Bag of Word с биграммами:

DecisionTreeClassifier:
F1 weighted: 0.8332714310529961
F1 macro: 0.8184812696687775
accuracy: 0.8334765960772056
CPU times: total: 1min 46s
Wall time: 1min 49s

KNeighborsClassifier
F1 weighted: 0.6058405057232046
F1 macro: 0.6213907345665199
accuracy: 0.5813862624052513
CPU times: total: 15.6 s
Wall time: 19.1 s

RandomForestClassifier
F1 weighted: 0.907069903731588
F1 macro: 0.8973503216469157
accuracy: 0.9074001719152927
CPU times: total: 20min 42s
Wall time: 23min 11s

LogisticRegression
F1 weighted: **0.9367904682213456**
F1 macro: 0.9310538528697619
accuracy: 0.9367039149800734
CPU times: total: 2.77 s
Wall time: 18.2 s

SVM
F1 weighted: 0.9273693623785011
F1 macro: 0.9195039569989137
accuracy: 0.9275611471438618
CPU times: total: 8min 50s
Wall time: 10min 9s

GradientBoostingClassifier
F1 weighted: 0.7949188984542157
F1 macro: 0.7801000728545825
accuracy: 0.7997186840665781
CPU times: total: 5min 31s
Wall time: 6min 5s

SGDClassifier
F1 weighted: **0.9388774996737284**
F1 macro: 0.9340437427395161

accuracy: 0.9389700711104165
CPU times: total: 984 ms
Wall time: 1.19 s

AdaBoostClassifier
F1 weighted: 0.6868389221618258
F1 macro: 0.6816768462683029
accuracy: 0.6969602250527467
CPU times: total: 15.1 s
Wall time: 17.8 s

LabelPropagation
F1 weighted: 0.5991446480143101
F1 macro: 0.6178806328592685
accuracy: 0.576072516996171
CPU times: total: 1min 26s
Wall time: 1min 48s

CatBoostClassifier

со следующими параметрами:
leaf_estimation_method= 'Gradient',
iterations = 9000,
learning_rate = 0.035,
max_depth = 6,
task_type="GPU",
bootstrap_type = 'Bernoulli',
objective= 'MultiClass',
subsample= 0.9,
eval_metric ='TotalF1'
F1 weighted: 0.9103925067861248
F1 macro: 0.9195039569989137
accuracy: 0.9111388260971237
CPU times: total: 53min 14s
Classification Report for CatBoost:

	precision	recall	f1-score	support
0	0.90	0.84	0.86	2840
1	0.90	0.90	0.90	3338
2	0.93	0.86	0.89	1008
3	0.92	0.97	0.94	5643
accuracy			0.91	12829
macro avg	0.91	0.89	0.90	12829
weighted avg	0.91	0.91	0.91	12829

9.2. Нейросеть BERTModel

Токенизация на основе предобученной модели
BertTokenizer.from_pretrained('sentence-transformers/LaBSE)
Гиперпараметры:
MAX_LEN = 430, batch_size = 8,
n_input = 768, n_hidden = 50, n_output = 4,
EPOCH = 2
optimizer = AdamW(bert_classifier.parameters(),
lr=5e-5, eps=1e-8

F1 weighted: 0.9360674848353339
F1 macro: 0.9293015827491917
accuracy: 0.9363161587029386
CPU times: total: 8815s
Classification Report for BERT :

	precision	recall	f1-score	support
0	0.91	0.87	0.89	2840
1	0.91	0.92	0.92	3338
2	0.94	0.94	0.94	1008
3	0.96	0.97	0.97	5643
accuracy			0.94	12829
macro avg	0.93	0.93	0.93	12829
weighted avg	0.94	0.94	0.94	12829

10. Тонкая настройка параметров модели (fine tuning)

По результатам предварительного обучения моделей (п.9) были выбраны две модели для тонкой настройки параметров:

- LogisticRegression
- SGDClassifier

10.1. Настройка гиперпараметров классических моделей производилась с помощью оптимизатора BayesSearchCV из пакета SciKit-optimize

Были подобраны следующие наилучшие гиперпараметры:

- LogisticRegression(C= 3.36, class_weight= 'balanced', solver='liblinear', random_state=42, max_iter=100)
- SGDClassifier(max_iter=5000, penalty='l2', loss='hinge')

10.2. Для каждой из моделей были опробованы все четыре метода векторизации и эмбединга (п. 8):

CountVectorizer;

LogisticRegression(C= 3.36, class_weight= balanced, solver=liblinear, max_iter=100)
CountVectorizer(ngram_range=(1,2))
F1 weighted: 0.9445867509592065
F1 macro: 0.93746860412653
accuracy: 0.944518246464015
Classification Report for Logistic regresion:

	precision	recall	f1-score	support
0	0.90	0.90	0.90	2718
1	0.92	0.94	0.93	3302
2	0.94	0.95	0.95	1042
3	0.98	0.97	0.97	5735
accuracy			0.94	12797

HashingVectorizer;

метрики для LogisticRegression(C= 3.36, class_weight= balanced, solver=liblinear, max_iter=100), HashingVectorizer(ngram_range=(1, 2), n_features=8388608)
F1 weighted: 0.937090112083063
F1 macro: 0.9296366634701323
accuracy: 0.9370164882394311

TfidfTransformer;

метрики для LogisticRegression(C= 3.36, class_weight= balanced, solver=liblinear, max_iter=100), TfidfTransformer(use_idf=True, sublinear_tf=True)
F1 weighted: 0.9398193295811853
F1 macro: 0.931735495530067
accuracy: 0.9395952176291318

Word2Vec обученный на текущем корпусе;

метрики для LogisticRegression(C= 3.36, class_weight= balanced, solver=liblinear, max_iter=100), Word2Vec(sentences, window=10, min_count=1, workers=4, epochs=30, sg=0, negative=5)
F1 weighted: 0.9122908149851168
F1 macro: 0.901382413033948
accuracy: 0.9120106274908182

Word2Vec предобученный

метрики для LogisticRegression(C= 3.36, class_weight= balanced, solver=liblinear, max_iter=100), Word2Vec model -- ruscorpora_upos_cbow_300_20_2019
F1 weighted: 0.27737440182540624
F1 macro: 0.15473235484567235
accuracy: 0.4481519106040478

10.3. С целью уменьшения влияния дисбаланса классов были опробованы следующие методы аугментации:

- ADASYN из пакета imblearn
- SMOTE из пакета imblearn
- RandomUnderSampler из пакета imblearn
- RandomOverSample из пакета imblearn
- взвешивание классов с помощью comput_class_weight из пакета sclearn

Ни одна из перечисленных методик не дала увеличения значений метрик

11. Наилучшие значения метрик

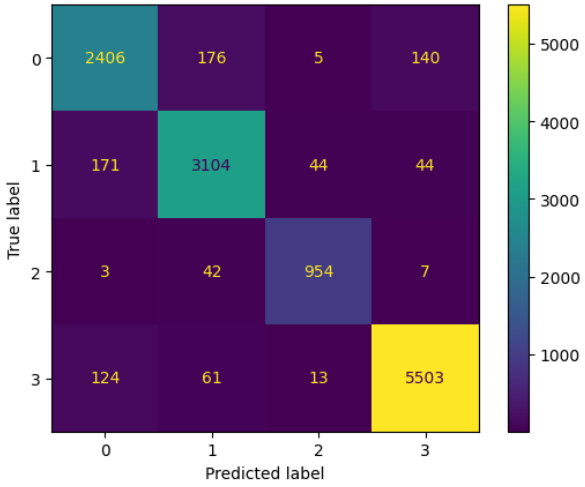
- В качестве основной метрики была выбрана F1 weighted,поскольку классы частично разбалансированы

- наилучшие значения метрик получились для модели:

LogisticRegression(C= 3.36, class_weight= balanced, solver=liblinear, max_iter=100),
CountVectorizer(ngram_range=(1,2)),
без аугментации классов

F1 weighted: 0.9445867509592065
F1 macro: 0.93746860412653
accuracy: 0.944518246464015
Classification Report for Logistic regresion:

	precision	recall	f1-score	support
химия	0.90	0.90	0.90	2718
физика	0.92	0.94	0.93	330
математика	0.94	0.95	0.95	1042
биология	0.98	0.97	0.97	5735
accuracy			0.94	12797



12. Результаты предсказаний для тестовой выборки

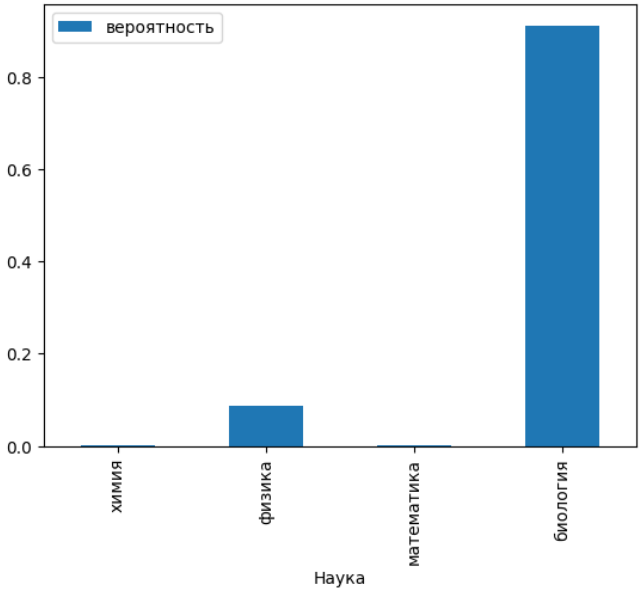
- для тестовой выборки были посчитаны вероятности отношения к каждому из классов методом predict_proba из пакета sklearn
- для более точного предсказания вероятностей модель была откалибрована методом CalibratedClassifierCV(logreg, cv=5, method=sigmoid)
- результаты предсказаний выводились в виде таблицы:

	test_text_original	test_label	химия	физика	математика	биология
0	некоторые задачи теории асимптотического интег...	математика	0.000294	0.000078	0.999575	0.000053
1	влияние легирования поведение гелия развитие г...	физика	0.023554	0.961606	0.000746	0.014094
2	трехкомпонентная конденсация гетероциклических...	химия	0.963356	0.029912	0.001376	0.005356
3	магнитные электрические тепловые свойства инте...	физика	0.033345	0.960209	0.002007	0.004439
4	исследование процессов электронно-волнового вз...	физика	0.002485	0.996595	0.000270	0.000649
...	...	...	...	...	...	...
12792	синтез коллоидно-химические свойства гидрозоле...	химия	0.995738	0.002457	0.001160	0.000645
12793	новые топологические нетривиальные решения стр...	физика	0.102920	0.882465	0.006284	0.008330
12794	механизмы действия ряда перспективных лекарств...	биология	0.002693	0.000106	0.002242	0.994958
12795	редкие исчезнувшие находящиеся угрозой исчезно...	биология	0.006659	0.002133	0.007570	0.983638
12796	эколого-фаунистический обзор чернотелок (coleo...	биология	0.007815	0.006660	0.003485	0.982040

12797 rows × 6 columns

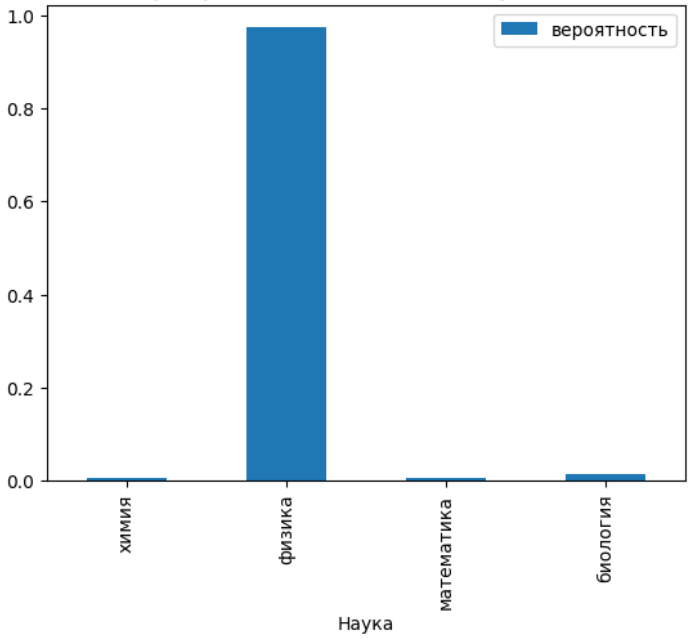
а также в виде гистограмм:

электрическая активность моторной сенсорной областей спинного мозга время спонтанных движений новорожденных крыс

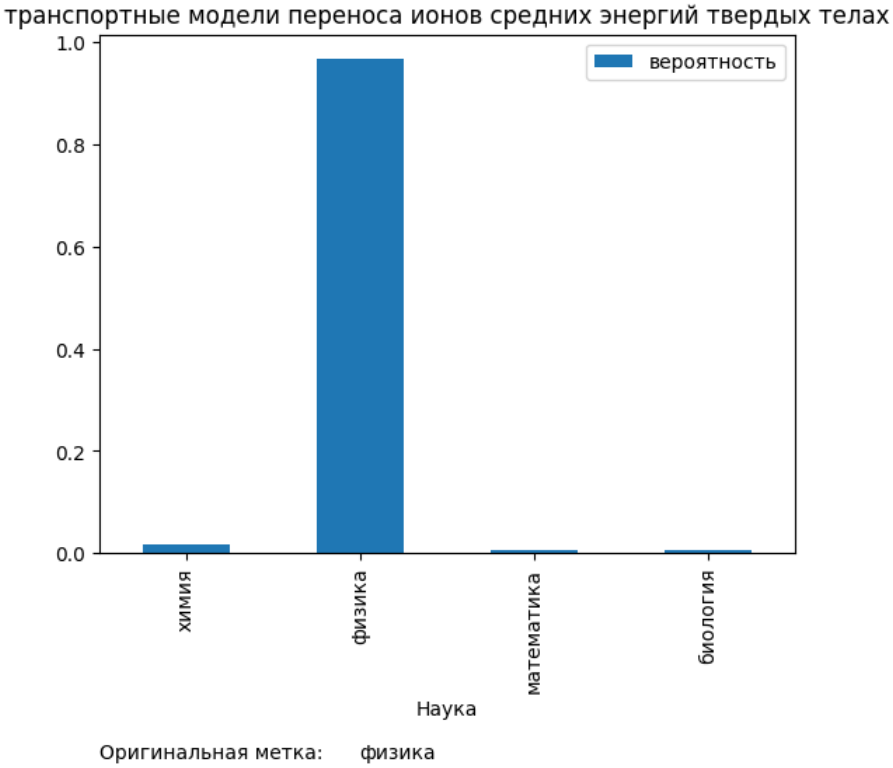


Оригинальная метка: биология

электрические свойства земной коры крайненизкочастотном-сверхнизкочастотном диапазонах радиоволн



Оригинальная метка: физика



13. Выводы

- Классические модели показали очень хорошие показатели метрики F1 weighted = 0.945, несмотря на то, что много диссертаций находятся на стыке наук.

Полагаю, по причине что данные очень чистые и с минимумом второстепенных слов в названиях, что характерно для научного языка.

- Данные метрик для Word2Vec векторизации ниже чем для Bag of Word. Думаю по причине ограниченности корпуса (всего 716297 слов) из которых всего 8% вошли в словарь (являются уникальными).

- Нейронная сеть с большим количеством параметров (Bert) не привела к улучшению метрик, по видимому потому что количество данных не достаточно (количество токеном значительно меньше количества параметров модели)

14. Ссылки

www.dissercat.com – электронная библиотека диссертаций

www.cyberleninka.ru/article - Каталог научных статей

Общая биология. Маонтгов С.Г, Захаров В.Б

Общая химия Глинка Н.Л.

Общая физика Савельев И.В. [4 тома]

Основы математического анализа [2 тома] Ильин, Позняк

Линейная алгебра. Ильин, Позняк

Лекции по теоретической физике, Белавин А.А., Кулаков А. Г., Усманов Р.А.