SeverusPT automated text clasification

JG, APP

1 Text mining and classification of papers

The main objective for this task is to use examples of papers that are selected (or not) for each of the phases of wildfire research, namely:

- Pre-fire
- During-fire
- Post-fire

Then a Random Forest model will be used to learn with these binary examples and classify all the remaining papers as selected or non-selected.

1.1 Count functions

Load libs and define ancillary functions

1.2 Data

Read the data and perform some basic preparation

Check colnames in the dataset

```
 [1] "...1"                       "label"
 [3] "label_1"                    "type"
 [5] "author"                     "title"
 [7] "year"                       "journal"
 [9] "volume"                     "number"
[11] "pages"                      "doi"
[13] "url"                        "affiliations"
[15] "abstract"                   "author_keywords"
[17] "correspondence_address"     "publisher"
```

```
[19] "issn"                        "language"
[21] "abbrev_source_title"         "publication_stage"
[23] "source"                      "note"
[25] "keywords"                    "coden"
[27] "pmid"                        "isbn"
[29] "editor"                      "address"
[31] "affiliation"                 "earlyaccessdate"
[33] "eissn"                       "keywords_plus"
[35] "research_areas"              "web_of_science_categories"
[37] "author_email"                "orcid_numbers"
[39] "cited_references"            "number_of_cited_references"
[41] "times_cited"                 "usage_count_last_180_days"
[43] "usage_count_since_2013"      "journal_iso"
[45] "doc_delivery_number"         "web_of_science_index"
[47] "unique_id"                   "month"
[49] "researcherid_numbers"        "article_number"
[51] "funding_acknowledgement"     "funding_text"
[53] "booktitle"                   "series"
[55] "book_author"                 "filename"
[57] "n_duplicates"                "oa"
[59] "da"                          "citation"
[61] "screened_titles"             "notes"
[63] "included"
```

Print class percentages:

Var1	Freq
excluded	48.32
selected	51.68

1.3 Training data

The training dataset is made by combining the following binary features,

- Top-30 most common words/terms from the **title** in *selected papers*;
- Top-30 most common words/terms from the **title** in *non-selected papers*;
- Top-30 most common words/terms from the **keywords** in *selected papers*;
- Top-30 most common words/terms from the **keywords** in *non-selected papers*;

1.3.1 Step 1 - count the most frequent terms in the title and keywords

Title top-30 terms

Count words in titles (after removing common/stop words) and select those terms that are most common in selected/accepted papers and non-selected/non-accepted papers.

These terms will be used as binary features in RF classification.

List of common terms in *titles* for *selected papers*:

included	word	n
1	fire	1039
1	forest	292
1	management	276
1	prescribed	178
1	risk	168
1	regimes	159
1	wildfire	159
1	weather	113
1	effects	106
1	forests	96
1	regime	95
1	fires	92
1	climate	79
1	fuel	66
1	model	55
1	change	54
1	prediction	54
1	vegetation	54
1	pine	52
1	carbon	51
1	soil	50
1	hazard	49
1	landscape	49
1	post	48
1	australia	47
1	implications	47
1	assessment	44
1	usa	44
1	mediterranean	41
1	response	41

List of common terms in *titles* for *non-selected papers*:

included	word	n
0	protection	296
0	prevention	91
0	hazards	83
0	study	66
0	treatment	61
0	safety	58
0	concrete	56
0	analysis	54
0	intumescent	53
0	steel	53
0	blight	50
0	coatings	50
0	strategies	42
0	composites	41
0	passive	40
0	performance	40
0	epoxy	37
0	evaluation	36
0	thermal	36
0	power	35
0	systems	35
0	properties	34
0	resistance	34
0	reducing	32
0	system	32
0	coal	30
0	coating	30
0	nuclear	30
0	storage	30
0	plants	27

Keywords top-30 terms

Now repeat the same process but now using keywords:

List of common terms in *keywords* for *selected papers*:

included	keyword	n
1	fire	1421
1	forest	896
1	management	712
1	fires	504
1	climate	405
1	wildfire	376
1	ecosystem	295
1	change	290
1	united	273
1	analysis	263
1	risk	253
1	environmental	233
1	burning	217
1	prescribed	217
1	forestry	213
1	assessment	191
1	carbon	181
1	vegetation	171
1	species	162
1	model	160
1	soil	155
1	weather	155
1	conservation	144
1	land	143
1	australia	142
1	modeling	139
1	population	137
1	pinus	126
1	ecology	122
1	north	119

List of common terms in *keywords* for *non-selected papers*:

included	keyword	n
0	protection	420
0	safety	201
0	heat	186
0	human	181
0	article	176
0	thermal	163

included	keyword	n
0	health	162
0	resistance	160
0	flame	152
0	coatings	143
0	hazards	142
0	combustion	138
0	temperature	137
0	smoke	132
0	concrete	129
0	humans	116
0	study	115
0	extinguishers	109
0	coal	107
0	materials	104
0	steel	102
0	male	101
0	systems	100
0	adult	99
0	female	94
0	hazard	92
0	gas	85
0	water	84
0	aged	83
0	performance	83

Abstract top-30 terms

Now repeat the same process but now using the abstract:

List of common terms in *abstracts* for *selected papers*:

included	abstract_	n
1	fire	6983
1	forest	1621
1	fires	1264
1	management	1215
1	species	1013
1	model	798

included	abstract_	n
1	risk	781
1	wildfire	764
1	vegetation	753
1	fuel	735
1	prescribed	680
1	climate	656
1	study	645
1	results	618
1	burned	613
1	data	601
1	forests	565
1	burning	563
1	weather	558
1	soil	554
1	effects	534
1	regimes	508
1	treatments	477
1	severity	474
1	change	452
1	land	445
1	conditions	431
1	increased	420
1	time	395
1	landscape	386
1	increase	372
1	regime	369
1	sites	359
1	cover	356
1	low	353
1	frequency	351
1	burn	350
1	models	345
1	ecological	341
1	season	330
1	carbon	328
1	spatial	327
1	post	322
1	wildfires	317
1	tree	316
1	scale	313
1	plant	300

included	abstract_	n
1	potential	300
1	structure	294
1	future	288

List of common terms in *abstracts* for *non-selected papers*:

included	abstract_	n
0	protection	534
0	temperature	470
0	safety	396
0	heat	384
0	thermal	377
0	analysis	340
0	flame	309
0	paper	292
0	rate	268
0	performance	250
0	system	246
0	method	239
0	resistance	233
0	steel	231
0	design	217
0	materials	217
0	properties	217
0	coatings	215
0	effect	215
0	smoke	214
0	elsevier	213
0	test	213
0	compared	212
0	research	210
0	concrete	202
0	combustion	198
0	hazard	198
0	release	195
0	water	192
0	blight	188
0	rights	187
0	assessment	186
0	methods	186

included	abstract_	n
0	control	185
0	developed	182
0	reserved	181
0	coating	178
0	systems	174
0	significant	169
0	coal	166
0	process	165
0	oxygen	160
0	surface	159
0	effective	158
0	intumescent	158
0	retardant	158
0	approach	156
0	gas	154
0	treatment	153
0	experimental	148

1.3.2 Step 2 - make features

Count the selected words in the title and keywords and arrange them as a binary grid:

Title-based features

Keyword-based features

Abstract-based features

1.3.3 Step3 - assemble all features and labels

Make the training dataset by combining everything:

1.4 Prediction dataset

Step 1 - Evaluate the same features as before but now for the entire dataset of papers

Step 2 - Assemble the full prediction dataset:

1.5 Random Forest model development

Make the classification model based on Random Forests:

```
  |
  |                                                                      |   0%
  |
  |==========                                                            |  14%
  |
  |====================                                                  |  29%
  |
  |==============================                                        |  43%
  |
  |========================================                              |  57%
  |
  |==================================================                    |  71%
  |
  |============================================================          |  86%
  |
  |======================================================================| 100%
```

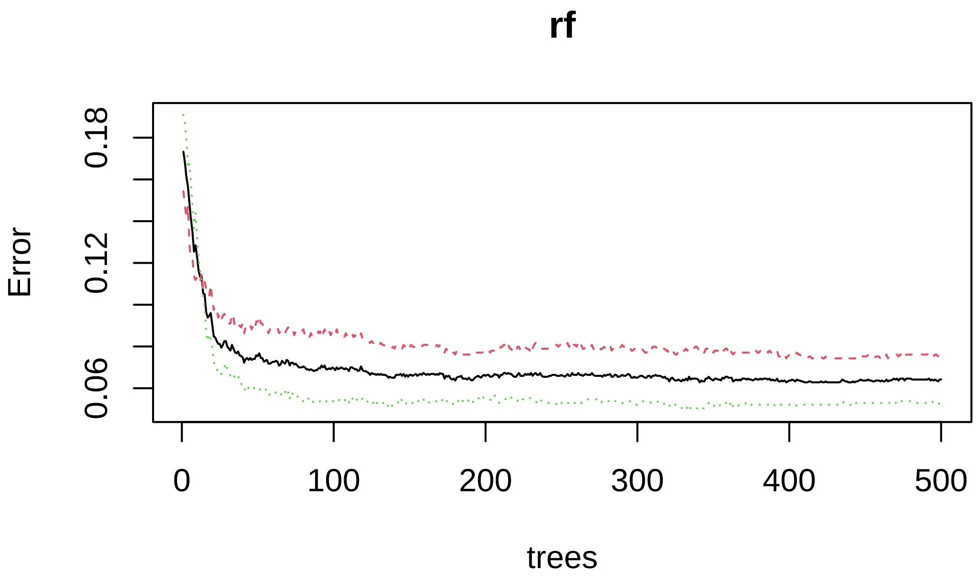
```
best mtry value:
```

```
[1] 7
```

```

Call:
 randomForest(x = train_tb %>% select(-label, -included), y = train_tb %>%      pull(included
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 7

        OOB estimate of  error rate: 6.42%
Confusion matrix:
    0    1 class.error
0 995   82  0.07613742
1  61 1091  0.05295139
```



List of the top-50 features by decreasing order of Mean Decrease in Gini Index:

MeanDecreaseGini	var_name
60.63	abstract_sel_forest
38.63	keywords_sel_forest
31.01	abstract_sel_management
28.25	keywords_sel_wildfire
26.07	abstract_sel_vegetation
25.27	abstract_sel_species
21.85	words_sel_forest
20.32	abstract_sel_burned
20.04	abstract_sel_prescribed
19.69	abstract_sel_regimes
19.59	abstract_sel_wildfire
18.73	keywords_sel_management
17.58	words_nsel_protection
17.28	abstract_sel_climate
17.10	keywords_sel_fire
16.47	abstract_sel_fires
14.90	abstract_sel_spatial
13.94	abstract_nsel_safety
13.92	words_sel_prescribed
13.83	abstract_sel_forests

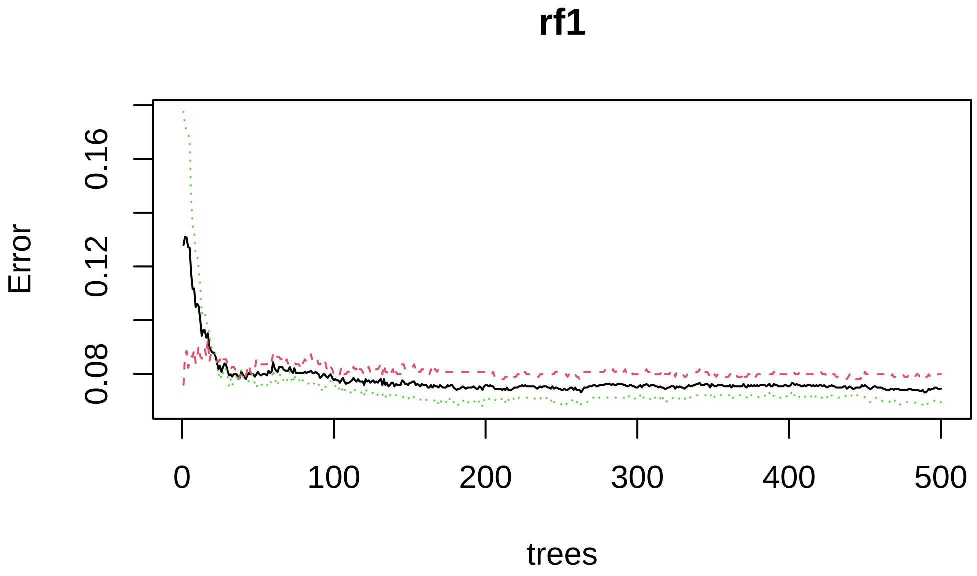
MeanDecreaseGini	var_name
13.59	abstract_sel_land
13.33	keywords_sel_climate
12.94	abstract_sel_fuel
12.51	keywords_sel_burning
10.93	keywords_sel_prescribed
10.85	abstract_sel_weather
10.44	keywords_nsel_protection
10.09	abstract_sel_landscape
10.05	abstract_sel_soil
9.95	keywords_sel_ecosystem
9.84	abstract_nsel_thermal
9.78	abstract_nsel_protection
9.32	words_sel_weather
9.31	keywords_sel_change
7.84	abstract_sel_ecological
7.82	abstract_sel_effects
7.45	abstract_nsel_heat
7.43	words_sel_regimes
7.00	words_sel_management
7.00	abstract_sel_regime
6.91	abstract_sel_fire
6.62	abstract_sel_model
6.59	abstract_sel_burning
6.20	keywords_nsel_human
5.99	abstract_sel_wildfires
5.53	abstract_sel_data
5.38	abstract_sel_frequency
5.17	abstract_sel_tree
4.99	abstract_sel_cover
4.82	keywords_sel_weather

Re-train the model but now with the top-50 best set of features based on the importance rank:

```

Call:
 randomForest(x = train_tb %>% select(all_of(imp_vars)), y = train_tb %>%      pull(included
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 7
```

```
        OOB estimate of  error rate: 7.45%
Confusion matrix:
    0    1 class.error
0 991   86  0.07985144
1  80 1072  0.06944444
```



Optimized cut-off to binarize the results:

thresh	auc	recall	prec
0.526	0.938	0.942	0.938

Predict class labels for the entire dataset using the optimized cut-off.

```
Predicted class percentages:
```

```
pred_class
 0  1
37 63
```