Classification of biomedical texts

Data description

Mario Sänger (WBI, HU Berlin)

saengema@informatik.hu-berlin.de

Classification tasks

- Your are given a data set of 15301 abstracts of biomedical articles

- Goal: Implement a classification pipeline to perform the following tasks
  1) Is the article concerned with cancer? (yes / no)
  2) About which specific type of cancer is the article about? (multi-class)

Training data format

- The training data set is given as tab-separated ("\t") file (*train.tsv*)
  - Each line represents one biomedical article (~ training instance)

```
pmid cancer_type doid is_cancer text
27581830 no cancer -1 0 The management of benign non-infective pleur
11213020 cancer 162 1 Bulky lymphadenopathy with poor clinical outco
22374460 endocrine gland cancer 170 1 Multicentre phase II trial of
27586118 no cancer -1 0 How safe are children when transported by bi
25868852 cancer 162 1 Loss of INPP4B causes a DNA repair defect thro
27599572 no cancer -1 0 TAK1 inhibitor NG25 enhances doxorubicin-med
27578867 no cancer -1 0 Cd47-Sirpα interaction and IL-10 constrain i
27581024 no cancer -1 0 Visualizing the Tumor Microenvironment of Li
27576518 no cancer -1 0 Brain damage resembling acute necrotizing en
25199829 cancer 162 1 RAF suppression synergizes with MEK inhibition
24962318 melanoma 1909 1 Targeting TBK1 inhibits migration and resis
22496619 cancer 162 1 Oncogenic KRAS impairs EGFR antibodies' effici
27597568 no cancer -1 0 Multiwavelength metasurfaces through spatial
18003960 gastrointestinal system cancer 3119 1 Cetuximab for the tre
27576507 no cancer -1 0 Expression of Helios in gastric tumor cells
```

Training data format

- The training data set is given as tab-separated ("\t") file (*train.tsv*)
  - Each line represents one biomedical article (~ training instance)

- Data columns:
  - *pmid:* unique (PubMed) identifier of the article
  - *cancer_type:* name of the specific cancer type the article is about or "no cancer", if the article isn't concerned with cancer
  - *doid:* unique identifier of the specific cancer type or -1, if the article isn't concerned with cancer
  - *is_cancer:* indicates whether the article is about cancer or not
  - *text:* abstract of the article

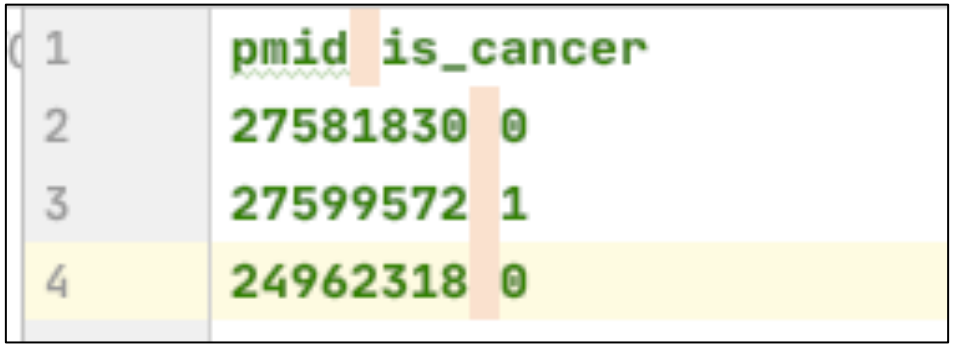
Evaluation

- Your classification model(s) will be evaluated on a hold-out evaluation data set
- For evaluation only the articles will be given (in the same tab-separated format)
  - I.e. the columns *is_cancer*, *cancer_type* and *doid* are missing!
  - Your classification model(s) should predict *is_cancer* and *doid*
  - See train_blind.tsv as an example!

```
1  pmid text
2  27581830 The management of benign non-infective pleural effusions.
3  17213020 Bulky lymphadenopathy with poor clinical outcome is assoc
4  22374460 Multicentre phase II trial of trastuzumab and capecitabin
5  27586118 How safe are children when transported by bicycle? With t
6  25868852 Loss of INPP4B causes a DNA repair defect through loss of
7  27599572 TAK1 inhibitor NG25 enhances doxorubicin-mediated apoptos
```

Submission format

- For evaluation you have to submit two files containing the predictions of your classification model
  - *task1.tsv*: Containing the predictions of your model for task 1 (about cancer?)
    - Two columns: *pmid* and *is_cancer* (0/1)
    - Use "\t" as column separator



Submission format

- For evaluation you have to submit two files containing the predictions of your classification model
  - *task2.tsv*: Containing the predictions of your model for task 2 (cancer-type)
    - Two columns: *pmid* and *doid*
    - Use "\t" as column separator
    - Include all articles from the test set, i.e. also articles that aren't about cancer at all have to be included!

```
1   pmid doid
2   27581830 -1
3   27599572 162
4   24962318 1909
5   24962318 -1
6   ...
```