INLS 690-270 Data Mining: Methods and Applications

Assignment 3

Please submit as a zip file attachment via Canvas.

This is a data ranking task through the online data competition platform Kaggle. You may use the following invitation link to join the competition but please refrain from posting the invitation link elsewhere: https://www.kaggle.com/t/951e88341bf8439ebeba1556ee16db45.

Background

The United States Environmental Protection Agency (EPA) periodically releases large-scale systematic reviews, called the Integrated Science Assessments (ISAs)[1] that synthesize the latest research on each of six air pollutants to inform environmental policymaking. To guarantee the best possible coverage of relevant literature, EPA scientists spend months manually screening hundreds of thousands of references to identify a small proportion to be cited in an ISA. The challenge of extreme scale and the pursuit of maximum coverage of relevant literature calls for effective machine-assisted approaches to reducing the time and effort required by the screening process.

In this task, you are going to build a machine learning model that can help EPA scientists find articles that will be cited in an ISA. More specifically, given a pool of references, we want to order them such that relevant references that will eventually be cited are ranked high and therefore if EPA scientists screen the reference pool in that order, they can identify relevant references earlier in the process.

We focus on two successive ISAs reporting the state of research on ozone, one written in 2013, the other written in 2020.[2] As training data, you are given the candidate pool of references used in 2013, and labels that indicate whether each reference in the pool was cited or not in the 2013 ozone report. As test data, you are given the candidate pool of references used in 2020, but this time without any labels that indicate whether each reference was cited or not in the 2020 ozone report. Your goal is to compute a score for each reference in the test data, such that the higher the score, the more likely a reference would be cited in the 2020 ozone report.

Data Description

The training data contains 15,772 references. The following metadata fields are provided:

- REFERENCE_ID: a unique identifier of a reference record;
- TITLE: the title of this reference;
- AUTHORS: the author names of this reference. Multiple author names are separated by semicolons. An empty string if the author information was not recorded as metadata;
- YEAR: year of publication of this reference;
- ABSTRACT: abstract of this article. An empty string if an article does not have an abstract or its abstract was not recorded as metadata;
- CITED: whether this reference was cited in the 2013 ozone report. ”1: means cited; ”0” means not cited. In total, 2,063 references are cited.

[1] https://www.epa.gov/isa

[2] https://www.epa.gov/isa/integrated-science-assessment-isa-ozone-and-related-photochemical-oxidants

The test data contains 171,376 references. It contains the same metadata fields as the training data, except that it does not have the CITED column (a total of 1,153 references from this pool were cited in the 2020 ozone report). The much larger reference pool (compared to that of 2013) reflects the ever-increasing volume of scientific literature, which makes manual screening increasingly costly and machine-assisted screening a necessity.

Evaluation

The submission should be a .csv (comma separated text) file with a header line "REFERENCE_ID,Score" followed by exactly 171,376 lines. In each line, there should be exactly two values, separated by a comma. The first value is the REFERENCE_ID of a test example (an integer), and the second value is the score predicted by your model (a number that can take any real values, and a higher number should indicate a higher position in the ranked list/higher chance of being cited).

You can make 10 submissions per day. Once you submit your results, you will get an average precision computed based on 50% of the test data. This will position you somewhere on the leaderboard. Once the competition ends, you will see the final accuracy computed based on the other 50% of the test data.

The evaluation metric is the **average precision** of your ranking scores. This metric takes values in the range $[0, 1]$; the higher, the better the ranked list. It is calculated as follows. First, we sort all references in the test set in descending order of your ranking score (from high to low). Then, we record the rank positions for each cited article in the test set. If there are N cited articles, then the ideal rank positions for those articles are Rank $1, 2, \cdots, N$, which will achieve the highest possible average precision, 1. In general, however, the scores may put some or all cited articles out of the top N. If we denote the rank positions of all N cited articles as $\{r_1, r_2, \cdots, r_N\}$, then the average precision (AP) is

$$AP = \frac{1}{N} \sum_{i=1}^{N} \frac{i}{r_i}\,.$$

For example, suppose $N = 5$, and the rank positions of the cited articles are $\{3, 5, 6, 10, 35\}$. Then the average precision is $(1/3 + 2/5 + 3/6 + 4/10 + 5/35)/5 = 0.355$. The higher each of the cited articles are ranked in the list, the smaller their rank position values, and the higher the average precision. You can learn more about average precision here and here.

You can use any classifier or combination of classifiers, any combination or selection of features, and either supervised, semi-supervised, or even transfer learning approaches. You can be creative and make use of external data sources. However, **please do not attempt to link back the original dataset.**[3]

Try to implement a ranking method that can outperform the baseline method: a logistic regression model using unigram count features derived only from the title text, with regularization hyperparameter $C = 1$; its average precision on test data ≈ 0.07. Its implementation is released in Canvas → Files → Homework Assignment 3. The formula to compute your grade (rounded to the closest integer):

$$grade = 80 + 20 * 2/\log_2(2 + rank)$$

Yes! The winner will get 105 points (105% grade of this homework)! The 25^{th} position will get 88 points.

Have fun! And don't waste your quota of submissions!

What to hand in: a one-page memo describing the algorithms/features/tools you explored and the corresponding results. Please write down your name and the display name you used in the competition. In addition, please attach source code to implement the algorithm in your submissions.

[3] `https://catalog.data.gov/dataset/isa-literature-screening-dataset-v-1` or `https://ils.unc.edu/~wangyue/isa-dataset/`.