Stage 3

To submit your results, make a public repo on Github called “NLP Challenges” where you should upload your code, as an IPython Notebook, in different folders for different challenges. Share the link with us and we would review the code and your approach and get back to you.

If it is found that the code was plagiarised, the applicant would get disqualified immediately.

Please note: If you have already created a repo for an earlier challenge, just add solutions to these as different folders in the same repo.

Tagging System of Questions using Transfer Learning

Overview:

In this challenge, we provide the titles, text, and tags of Stack Exchange questions from six different sites. We then ask for tag predictions on unseen physics questions. Solving this problem via a standard machine learning approach might involve training an algorithm on a corpus of related text. Here, you are challenged to train on material from outside the field. Can an algorithm predict appropriate physics tags after learning from biology, chemistry or mathematics data? Let's find out!

Objective:

Main goal of this task is to train a model on questions belonging to domains like biology, chemistry, or mathematics but use that to predict tags of physics question. These tags describe the topic of questions.

Dataset:

In this dataset, you are provided with question titles, content, and tags for Stack Exchange sites on a variety of topics (biology, cooking, cryptography, diy, robotics, and travel). The content of each question is given as HTML. **The tags are words or phrases that describe the topic of the question.**

The test set is comprised of questions from the *physics.stackexchange.com*. For each question in the test set, you should use the title and question content in order to generate potential tags.

Available here

Example:

Question: When you see the atomic mass number for an element, does it take into account the atomic mass defect?

Tags: **mass** **nuclear-physics** **binding-energy** **metrology** **elements**

Evaluation:

The evaluation metric for this challenge is **Mean F1-score**. The F1 score measures accuracy using the statistics precision p and recall r. Precision is the ratio of true positives (tp) to all predicted positives (tp + fp). Recall is the ratio of true positives to all actual positives (tp + fn). The F1 score is given by

$$F1 = 2\frac{p \cdot r}{p + r} \quad \text{where} \quad p = \frac{tp}{tp + fp}, \quad r = \frac{tp}{tp + fn}$$

The F1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favoured over extremely good performance on one and extremely bad performance on other.

Submission:

For every question in the dataset, submission files should contain two columns: id and tags. The predicted tags should be a space-delimited list. The file must have a header and should look like the following:

```
id,tags

1,physics poetry

2,physics poetry chemistry

3,physics electrons
```

Explain the approach you took and your logic in a Readme file. It should make it clear to the evaluator why you did what you did.

Helpful Resource:

[Transfer Learning on Stack Exchange Tags]

Note:

You are free to use any approach you see fit. You need to be able to cite appropriate research papers that were used to make key decisions.