Training Data Requirements

The training data should be sourced from a robust text corpus (or corpora), preferably written by experts and peer reviewed. In the original model, the titles and abstracts of all heliophysics-related papers written in a single year were sourced in order to train the model on. This is to ensure that the model maintains a high level of accuracy when defining entities.

Training data for the model must take the form of individual sentences with at least one clear subject, and object, as well as a linking verb showing the relationship between the subject and object.

A suggested methodology for preparing training data for the model is as follows:

1. Select an appropriate source for the model. Some potential sources for training data are academic papers, abstracts, and book chapters relevant to the subject.
2. Examine the data for sentences that present examples of subjects, objects, and linking verbs that relate to the model topic. For example, the sentence “FBs form as a result of the interaction between solar wind discontinuities and backstreaming ion beams in the foreshock.” was selected for training the model.
3. Put the sentence in a list and annotate all entities/nouns. This is done by creating a dictionary with a key-value pair of “entities” and a list of entities – e.g.

```
TRAIN_DATA = {
      ("FBs form as a result of the interaction between solar wind
      discontinuities and backstreaming ion beams in the foreshock.",
      {"entities": [(48, 58, "ASTROPHYSICS"), (93, 97,
      "ASTROPHYSICS")]})
}
```

4. The annotation is done by marking the starting index of the first letter and the end index of the last letter and then assigning it a category. For example, the entity “solar wind” in the above example starts at index 48 and ends at index 58 and falls into the ASTROPHYSICS label category.
5. Repeat for as many entities as are deemed necessary. For a novel field, around 50-100 training samples at minimum are recommended.
6. An example of a filled-out list of examples is as follows:

```
TRAIN_DATA = [
("BepiColombo, a joint mission to Mercury by the European Space
Agency and Japan Aerospace Exploration Agency, will address
remaining open questions using two spacecraft, Mio and the
Mercury Planetary Orbiter.", {"entities": [(0, 11, "MISSION"),
(32, 39, "ASTROPHYSICS"), (47, 68, "ORG"), (74, 107, "ORG"),
(169, 172, "MISSION"), (181, 206, "MISSION")]}),
("Mio First Comprehensive Exploration of Mercury's Space
Environment: Mission Overview", {"entities": [(0, 85,
"PAPER")]}),
```

("Martian Thermospheric Warming Associated With the Planet Encircling Dust Event of 2018", {"entities": [(0, 86, "PAPER")]}),
("Ray-and-power tracing provided wave amplitudes as well as trajectories and wave normal angles throughout the plasmasphere.", {"entities": [(31, 46, "ASTROPHYSICS"), (109, 121, "ASTROPHYSICS")]}),
("Dynamical Evolution of Simulated Particles Ejected From Asteroid Bennu", {"entities": [(0, 70, "PAPER")]}),
("We use global and local hybrid kinetic ions and fluid electrons simulations to investigate the conditions under which foreshock bubbles FBs form and how their topology changes with solar wind conditions.", {"entities": [(48, 58, "ASTROPHYSICS")]}),
("FBs form as a result of the interaction between solar wind discontinuities and backstreaming ion beams in the foreshock.", {"entities": [(48, 58, "ASTROPHYSICS"), (93, 97, "ASTROPHYSICS")]}),
("The visible and near-infrared imaging spectrometer on board the Yutu-2 rover of ChangE-4 mission has conducted 2 sets of spectrophotometric measurements at two sites on its 10th lunar day.", {"entities": [(38, 51, "ASTROPHYSICS"), (64, 71, "PROJECT"), (80, 89, "MISSION")]}),
("The Mars Science Laboratory mission investigated Vera Rubin ridge, which bears spectral indications of elevated amounts of hematite and has been hypothesized as having a complex diagenetic history.", {"entities": [(4, 36, "MISSION"), (49, 60, "ASTROPHYSICS")]}),
("The InSight mission to Mars landed within Homestead hollow on an Early Amazonian lava plain.", {"entities": [(4, 20, "MISSION"), (23, 28, "ASTROPHYSICS"), (42, 52, "ASTROPHYSICS"), (81, 91, "ASTROPHYSICS")]}),
("The many completed studies show an Ice Giant mission with an in situ probe is feasible and would be welcomed by the international science community.", {"entities": [(35, 53, "MISSION")]}),
("NASA Parker Solar Probe mission is currently investigating the local plasma environment of the inner heliosphere <0.25 R_{\u2609} using both in situ and remote sensing instrumentation.", {"entities": [(0, 5, "ORG"), (5, 32, "MISSION"), (69, 76, "HELIOPHYSICS"), (101, 113, "HELIOPHYSICS")]}),
("We will relate the results of the Rosetta mission to those of the flybys.", {"entities": [(34, 50, "MISSION")]}),
("Cometary Nuclei: From Giotto to Rosetta", {"entities": [(0, 40, "PAPER")]}),
("A Maximum Rupture Model for the Southern San Andreas and San Jacinto Faults, California, Derived From Paleoseismic Earthquake Ages: Observations and Limitations", {"entities": [(0, 161, "PAPER")]}),
("The CESM2 is the version of the CESM contributed to the sixth phase of the Coupled Model Intercomparison Project CMIP6.", {"entities": [(4, 10, "PROJECT"), (32, 37, "PROJECT"), (75, 113, "PROJECT")]}),
("The datasets of two Ocean Model Intercomparison Project simulation experiments from the Climate Ocean Model Project, forced by two different sets of atmospheric surface data, are

described in this paper.", {"entities": [(20, 56, "PROJECT"), (89, 116, "PROJECT")]}),
("Model simulations in the Community Earth System Model Large Ensemble Project confirmed the physical connection between the warm CEP SST anomaly and the drought in EC.", {"entities": [(25, 77, "PROJECT")]}),
("The pickup process on the extended oxygen corona created by the strong EUV flux contributes to the total O+ loss.", {"entities": [(42, 49, "HELIOPHYSICS")]}),
("As systems become more complex over time, the impacts of space weather on space flights and humanity in general are likely to increase.", {"entities": [(57, 71, "ASTROPHYSICS"), (74, 88, "MISSION")]}),
("Humans will encounter extremely serious problems of space flight safety at the beginning of new phase of the Moon exploration.", {"entities": [(52, 65, "MISSION"), (109, 114, "ASTROPHYSICS")]}),
("Motivated by a successful prediction on the peak of solar cycle 24 81.7, comparable to the observed 81.9, Du in Astrophys.", {"entities": [(52, 64, "HELIOPHYSICS")]})

]