MODEL 5 (GEO):

pip install sense2vec

prodigy sense2vec.teach geo6_method /Users/sashaqanderson/Dropbox/USGS/NER_Work/s2v_old --seeds "water_flow, heat_transfer, subduction_zone,mass_transfer,soil_erosion "

prodigy terms.to-patterns geo6_method --label GEO

{"label":"GEO","pattern":[{"lower":"water_flow|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"heat_transfer|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"subduction_zone|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"mass_transfer|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"soil_erosion|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"evaporation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"surface_water|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"water_cycle|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"heat_exchange|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"ocean_currents|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"thermal_expansion|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"groundwater|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"water_table|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"ground_water|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"temperature_gradient|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"carbon_cycle|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"oceans|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"sedimentation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"boundary_layer|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"pollutants|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"compaction|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"solar_radiation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"sediments|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"natural_convection|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"liquefaction|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"thermal_energy|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"volcanism|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"currents|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"salinity|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"greenhouse_gasses|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"pressure_differentials|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"greenhouse_effect|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"evaporation_rate|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"hydrostatic_pressure|NOUN"}]}

{"label":"GEO","pattern":[{"lower":"soils|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"aquifer|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"greenhouse_gases|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"ocean_water|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"subduction_zones|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"particulate_matter|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"subsidence|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"surface_waters|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"reservoirs|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"volcanic_activity|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"wellbore|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"green_house_effect|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"water_pressure|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"air_currents|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"dissolved_oxygen|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"ocean_waters|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"surface_tension|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"advection|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"tectonic_activity|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"watersheds|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"temperature_differential|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"plate_tectonics|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"mechanical_stress|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"latent_heat|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"convection_currents|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"gas_exchange|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"water_pollution|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"meltwater|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"particulates|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"atmospheric_temperature|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"jet_stream|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"heat_energy|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"precipitation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"surface_layer|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"waste_water|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"ocean_surface|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"subduction|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"acidification|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"cavitation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"rainfall|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"CO2|ORG"}]}

{"label":"GEO","pattern":[{"lower":"methane|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"seawater|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"carbon_dioxide|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"atmospheric_oxygen|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"hydrocarbons|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"atmospheric_pressure|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"sea_water|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"upper_atmosphere|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"nitrogen|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"troposphere|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"atmospheric_carbon|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"ozone|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"plant_growth|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"dissolved_gasses|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"organic_carbon|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"organic_matter|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"methane_gas|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"algal_blooms|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"atmospheric_CO2|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"phytoplankton|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"sediment|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"thermal_radiation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"C02|ORG"}]}
{"label":"GEO","pattern":[{"lower":"CO^2|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"nitrogen_oxides|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"water_column|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"evaporated_water|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"hydrogen_gas|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"oxygen|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"atmospheric_composition|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"organic_material|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"dissolved_gases|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"sulfur_dioxide|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"dissolved_minerals|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"ambient_pressure|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"sea_ice|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"cloud_formation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"GHGs|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"surface_temperature|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"CO2_levels|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"hydrocarbon|NOUN"}]}

{"label":"GEO","pattern":[{"lower":"algae_blooms|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"seafloor|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"contaminants|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"radioactive_particles|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"hydrogen_sulfide|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"greenhouse_gas|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"photosynthesis|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"hydrogen|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"anaerobic_bacteria|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"radioactive_elements|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"particulate|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"radioactive_isotopes|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"solutes|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"green_house_gasses|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"biomass|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"green_house_gases|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"microorganisms|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"dissolved_CO2|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"radioactive_decay|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"heavy_metals|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"acid_rain|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"polar_caps|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"lower_atmosphere|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"sulfuric_acid|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"microbes|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"silicates|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"ozone_layer|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"crystal_formation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"radiation|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"CO2_concentration|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"carbonic_acid|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"volcanic_ash|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"plant_life|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"contaminates|NOUN"}]}
{"label":"GEO","pattern":[{"lower":"hydrothermal_vents|NOUN"}]}

prodigy db-out geo6_method > /Users/sashaqanderson/Dropbox/USGS/NER_Work/model6/geo6_patterns.jsonl

Prodigy ner.manual geo6_method_data blank:en
/Users/sashaqanderson/Dropbox/USGS/NER_Work/model6/train40.jsonl --label GEO

prodigy db-out geo6_method_data >
/Users/sashaqanderson/Dropbox/USGS/NER_Work/model6/geo6_data_model.jsonl

python -m spacy pretrain /Users/sashaqanderson/Dropbox/USGS/NER_Work/model6/train40.jsonl en_vectors_web_lg /Users/sashaqanderson/Dropbox/USGS/NER_Work/model6/pretrain_geo6/ --use-vectors

Prodigy ner.batch-train geo6_method_data en_vectors_web_lg --init-tok2vec /Users/sashaqanderson/Dropbox/USGS/NER_Work/model6/pretrain_geo6/model999.bin --output geo6_model --eval-split 0.2 --label GEO

46% Accuracy

Prodigy ner.make-gold geo6_method_data_correct ./geo6_model /Users/sashaqanderson/Dropbox/USGS/NER_Work/model6/train40.jsonl --label GEO

ValueError: [E103] Trying to set conflicting doc.ents: '(70, 83, 'GEO')' and '(70, 83, 'GEO')'. A token can only be part of one entity, so make sure the entities you're setting don't overlap.