Семинарское задание №2
Word2Vec

Мосиенко Константин Викторович

2019

Вам дана коллекция пар («название организации», «род занятий»), по одной паре на строку, поля разделены символами табуляции. Вам предстоит научиться предсказывать род деятельности организации по её имени. Обучающая выборка располагается в файле *names_and_rubrics_learn.tsv*, тестовая - *names_and_rubrics_test_no_rubric.tsv*. Для каждой строки тестовой выборки необходимо предсказать род занятий организации и в формате, аналогичном формату обучающей выборки, записать результат в файл, который следует прислать вместе с кодом и отчётом о том, как работает код.

Примеры из обучающей выборки:

Антикварный магазин Антик-маркет	Антикварный магазин
ЧОО АБ Снег	Охранное предприятие
ВекторЭлектро	Электромонтажные работы
Растяпино	Магазин алкогольных напитков

Обратите внимание, что род занятий - это не произвольная строка, точно описывающая виды деятельности компании. Это некоторая характеристика, выбираемая из некоторого заранее фиксированного множества. Приведём топ видов деятельности (рубрик):

Гостиница	9909
Ресторан	7864
Кафе	6192
Шиномонтаж	6108

Вы вольны решать задачу любым способом. Я лишь приведу несколько своих соображений. Можно научиться строить эмбеддинги имён и рубрик. Тогда, имея произвольное имя, можно сравнить его со всеми рубриками и выбрать самую близкую. Строка «Охранное предприятие» может являться как рубрикой, так и названием организации. Поэтому стоит добавить в обучение и сами рубрики. Также это отличный тест для реализации - организация с именем, совпадающим с именем какой-то рубрики, скорее всего должна иметь именно эту рубрику.

Авторы пяти наилучших работ получат дополнительный балл. Метрика - процент верных предсказаний на тестовой выборке. Бейзлайна нет - сравнивайтесь друг с другом, обсуждайте в чате. Будут зачтены все адекватные работы, использующие построение эмбеддингов. В отчёте обязательно надо написать что, зачем и как сделано. Программировать нейросети самому не обязательно, можно взять готовые библиотеки (например, fasttext), можно даже взять готовые нейросети(кроме тех, что обучены вашими коллегами), если такие найдутся. Также в отчёт необходимо добавить примеры удачных и неудачных классификаций.

Если вы пишете в ipynb, то отдельный отчёт можно не делать - всё можно описать и там.