Отчет

# **Поставленная задача.**

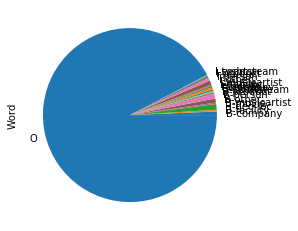
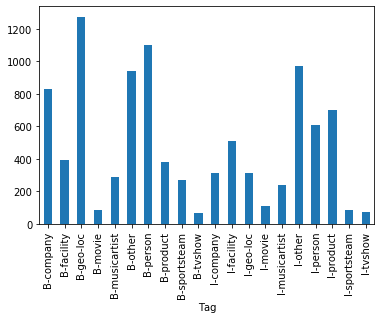
Извлечь именованные сущности из предоставленных текстовых данных.

# **Входные данные.**

Текстовый файл, где в каждой строке записаны токен и тэг, разделенные пробелом. Последовательности таких строк формируют предложения, разделенные пустой строкой.

# **Дисбаланс классов.**

Токены, не являющиеся именованными сущностями, составляют 92% всех токенов.



Так выглядит распределение остальных тэгов.

Наименее распространенным тэгом среди именованных сущностей является тэг 'I-tvshow', наиболее распространенным — 'B-geo-loc'.

# **Пунктуация.**

Токены, представляющие собой знаки препинания, составляют 13% всех токенов. Часть из них согласно разметке представляют собой именованные сущности. В действительности они ими не являются, но могут нести дополнительную информацию, которую алгоритм классификации сможет учитывать для определения настоящих именованных сущностей.

В работе алгоритм классификации будет обучен на двух датасетах — включающих токены пунктуации и не включающих.

# **Используемые признаки.**

В задаче распознавания именованных сущностей хорошим признаком является частеречная разметка. В работе используется библиотека NLTK, чтобы присвоить каждому слову частеречный тэг.

Кроме того, в качестве признаков для алгоритма классификации будут выбраны:

* длина слова
* шаблоны капитализации
* суффиксы и префиксы слов
* метки первого и последнего слов в предложении
* и другие

Кроме того, в списке признаков каждого слова будут присутствовать признаки его слов-соседей.

Хорошими признаками могли бы служить эмбеддинги слов, но за неимением достаточных вычислительных мощностей и оперативной памяти, чтобы использовать предобученные векторы, в данной работе они не присутствуют.

# **Обучение модели**

В качестве алгоритма классификации были опробованы Perceptron, SGDClassifier, CRF (метод условных случайных полей). Последний показал наилучшие результаты. Использование нейронных сетей и языковых моделей (ELMO, Bert) невозможно ввиду ограниченных вычислительных ресурсов (а также малой обучающей выборки).

# **Результаты**

Результаты классификации оказались приблизительно равными для двух датасетов.

Ниже представлен отчет по качеству работы классификатора для датасета, включающего знаки пунктуации. Так как тэг 'O' сильно искажал результаты, делая их завышенными, он исключен из оценки. Наиболее релевантной метрикой является f1-score. Она представляет собой среднее гармонической точности (precision) и полноты (recall). Красным выделена ячейка, значение которой лучшим образом отражает качество классификации для данной выборки с учетом её дисбаланса.

precision recall f1-score support

B-sportsteam 0.50 0.22 0.31 50

I-sportsteam 0.50 0.20 0.29 15

B-facility 0.78 0.53 0.63 87

I-facility 0.80 0.62 0.70 117

B-company 0.79 0.53 0.63 156

B-geo-loc 0.83 0.53 0.64 258

B-other 0.69 0.40 0.50 179

I-other 0.56 0.43 0.49 169

B-person 0.68 0.42 0.52 225

I-person 0.63 0.52 0.57 124

B-movie 0.80 0.18 0.30 22

I-movie 1.00 0.24 0.39 29

B-product 0.52 0.20 0.29 71

I-product 0.31 0.29 0.30 139

I-geo-loc 0.85 0.42 0.56 69

B-musicartist 0.40 0.15 0.22 54

I-musicartist 0.27 0.06 0.10 62

B-tvshow 1.00 0.06 0.11 17

I-company 0.60 0.48 0.53 73

I-tvshow 0.00 0.00 0.00 23

micro avg 0.65 0.41 0.50 1939

macro avg 0.63 0.32 0.40 1939

weighted avg 0.65 0.41 0.49 1939

Стоит обратить внимание на класс I-tvshow, оценки качества классификации которого равны нулю. Это самый малочисленный класс, поэтому либо классификатору не хватило данных, чтобы обучиться его классифицировать, либо объекты с данным тэгом вовсе не попали в обучающую выборку.