Задача поиска символов в текстах*

Северилов П. А.[1]***, Лемтюжникова Д. В.***[1]***, Апишев М. А.***[2]
severilov.pa@phystech.edu
[1]МФТИ; [2]МГУ

Поиск символов в текстах. Сравнение работы нескольких state-of-the-art алгоритмов. Предложена метрика качества классификатора для символов (символ/не символ). Определена применимость методов.

Ключевые слова: *sequence labeling, recurrent neural network, hidden markov model.*

1 Введение

В простейшем случае эта задача сводится к задаче Sequence Labeling на размеченной выборке. Сложность заключается в получении достаточного объёма обучающих данных, то есть требуется по имеющейся небольшой экспертной разметке получить выборку большего размера (автоматически путём поиска закономерностей или же путём составления несложной и качественной инструкции для разметки, например, в Толоке). Наличие разметки позволяет начать эксперименты с подбором оптимальной модели, здесь могут быть интересны разнообразные нейросетевые архитектуры (BiLSTM, Transformer и т.п.). Предлагаемый подход к анализу текста используется экспертами в ручном режиме и не был автоматизирован

2 Название раздела

Данный документ демонстрирует оформление статьи, подаваемой в электронную систему подачи статей `http://jmlda.org/papers` для публикации в журнале «Машинной обучение и анализ данных». Более подробные инструкции по стилевому файлу `jmlda.sty` и использованию издательской системы LaTeX 2$_\varepsilon$ находятся в документе `authors-guide.pdf`. Работу над статьёй удобно начинать с правки TeX-файла данного документа.

2.1 Название параграфа.

Нет ограничений на количество разделов и параграфов в статье. Разделы и параграфы не нумеруются.

2.2 Теоретическую часть работы

желательно структурировать с помощью окружений Def, Axiom, Hypothesis, Problem, Lemma, Theorem, Corollary, State, Example, Remark.

Определение 1. *Математический текст хорошо структурирован, если в нём выделены определения, теоремы, утверждения, примеры, и т. д., а неформальные рассуждения (мотивации, интерпретации) вынесены в отдельные параграфы.*

Утверждение 1. *Мотивации и интерпретации наиболее важны для понимания сути работы.*

Теорема 1. *Не менее* 90% *коллег, заинтересовавшихся Вашей статьёй, прочитают в ней не более* 10% *текста.*

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Задачу поставил: Лемтюжникова Д. В. Консультант: Апишев М. А.

Таблица 1 Подпись размещается над таблицей.

Задача	CCEL	boosting
`Cancer`	$\mathbf{3.46} \pm 0.37$ (3.16)	4.14 ± 1.48
`German`	$\mathbf{25.78} \pm 0.65$ (1.74)	29.48 ± 0.93
`Hepatitis`	18.38 ± 1.43 (2.87)	19.90 ± 1.80

Доказательство. Причём это будут именно те разделы, которые не содержат формул. ■

Замечание 1. Выше показано применение окружений Def, Theorem, State, Remark, Proof.

3 Некоторые формулы

Образец формулы: $f(x_i, \alpha^\gamma)$.

Образец выключной формулы без номера:

$$y(x,\alpha) = \begin{cases} -1, & \text{если } f(x,\alpha) < 0; \\ +1, & \text{если } f(x,\alpha) \geqslant 0. \end{cases}$$

Образец выключной формулы с номером:

$$y(x,\alpha) = \begin{cases} -1, & \text{если } f(x,\alpha) < 0; \\ +1, & \text{если } f(x,\alpha) \geqslant 0. \end{cases} \tag{1}$$

Образец выключной формулы, разбитой на две строки с помощью окружения align:

$$R'_N(F) = \frac{1}{N}\sum_{i=1}^{N}\Big(P(+1\,|\,x_i)C\big(+1, F(x_i)\big) + \\ + P(-1\,|\,x_i)C\big(-1, F(x_i)\big)\Big). \tag{2}$$

Образцы ссылок: формулы (1) и (2).

4 Пример илюстрации

Рисунки вставляются командой `\includegraphics`, желательно с выравниванием по ширине колонки: `[width=\linewidth]`.

Практически все популярные пакеты рисуют графики с подписями, которые трудно читать на бумаге и на слайдах из-за малого размера шрифта. Шрифт на графиках (подписи осей и цифры на осях) должны быть такого же размера, что и основной текст.

При значительном количестве рисунков рекомендуется группировать их в одном окружении `{figure}`, как это сделано на рис. **??**.

5 Пример таблицы

Подпись делается *над таблицей*, см. таблицу 1.

6 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.