Университет ИТМО, факультет программной инженерии и компьютерной техники
Двухнедельная отчётная работа по «Информатике»: аннотация к статье

Дата прошедшей лекции	Номер прошедшей лекции	Название статьи/главы книги/видеолекции	Дата публикации (не старше 2021 года)	Размер статьи (от 400 слов)	Дата сдачи
11.09.2024	1	Information Theory, Living Systems, Communication Engineering	18.05.2024	~5050	25.09.2024
25.09.2024	2	Research and Development of Data Compression Methods Based on Neural Networks	01.01.2023	~3122	09.10.2024
09.10.2024	3	Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application	03.11.2021	~9800	23.10.2024
23.10.2024	4	MarkupLM: Pre-training of Text and Markup Language for Visually-rich Document Understanding	11.03.2022	~2900	06.11.2024
	5				
	6				
	7				

Выполнил(а) Юксель Хамза, № группы *P3132*, оценка _____
Фамилия И.О. студента — не заполнять

Прямая полная ссылка на источник или сокращённая ссылка (bit.ly, tr.im и т.п.)
https://doi.org/10.48550/arXiv.2110.08518

Теги, ключевые слова или словосочетания (минимум три слова)
MarkupLM, Visually Rich Document Understanding (VRDU), Markup Language, XPath, Web-Based Tasks, Dynamic Rendering, Document Representation Learning, Evaluation Datasets

Перечень фактов, упомянутых в статье (минимум четыре пункта)
1. MarkupLM is a new way of teaching computers how to understand documents, especially documents that do not have a fixed layout, such as websites and instead of relying on visual layout like some other models, MarkupLM focuses on the actual structure of the document, using things like HTML tags to understand how different parts of the document are related.
2. MarkupLM uses something called a DOM tree and XPath to break down the document into its individual components and understand how they all connect.
3. One of the MarkupLM's feauture is an XPath embedding layer and it helps the model understand the hierarchy of the document, like which elements are nested inside others.
4. MarkupLM uses three different ways to learn from the data: **Masked Markup Language Modeling** (MMLM) for learning contextual information, **Node Relation Prediction** (NRP) for understanding relationships between nodes in the markup tree and **Title-Page Matching** (TPM) for capturing high-level semantics using the title and body elements.
5. Evaluation on WebSRC and SWDE datasets demonstrates that MarkupLM significantly outperforms existing latest models in web-based structural reading comprehension and information extraction tasks.

Позитивные следствия и/или достоинства описанной в статье технологии (минимум три пункта)
1. Efficient Handling of Dynamically Rendered Documents.
2. Using Markup Structure For Better Understanding.
3. Elimination Of Rendering Requirements.
4. Reduced Training Time and Sample.
5. Outstanding Performance in Web-Based Tasks.

Негативные следствия и/или недостатки описанной в статье технологии (минимум три пункта)
1. Reliance on Well-Structured Markup.
2. Challenges With Unstructured Data.
3. Need For Continued Development.

Ваши замечания, пожелания преподавателю *или* анекдот о программистах[1]

[1] Наличие этой графы не влияет на оценку