* **Мы считаем статистики топ, среднее, медиану, по годам, не по годам, для статей, и для цитируемости**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field Name | Field Type | Description | Example | F1ght TEAM |
| id | string | paper ID | 43e17f5b20f7dfbc07e8ac6e | Выкинуть |
| title | string | paper title | Data mining: concepts and techniques | 1. Top words (самое популярное слово Облако тегов) 2. Untop wordsx 3. Clusterization 4. Построить embedding (не нужно) |
| authors.name | string | author name | Jiawei Han | 1. Сделать нормальную таблицу… 2. Топ авторов по цитируемости + по количеству статей 3. Пары авторов (нграма). |
| author.org | string | author affiliation | Department of Computer Science, University of Illinois at Urbana-Champaign | 1. Топ огранизации 2. Если орг несколько ({}), то сделать нормальную таблицу 3. Bar chart по количеству статей за год, за 5 лет… 4. У автора может быть несколько организации 5. Сколько работает авторов в организации за промежуток вреня |
| author.id | string | author ID | 53f42f36dabfaedce54dcd0c | 1. PK |
| venue.id | string | paper venue ID | 53e17f5b20f7dfbc07e8ac6e | 1. PK |
| venue.raw | string | paper venue name | Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial | Название Журнала   1. Статистики 2. Топ, худшие лучшие 3. Самое популярные слово в журнале по цитируемости, и по количеству статей (может с разрезе времени) (для самых популярных слов) 4. … |
| year | int | published year | 2000 | 1. Очень важная фича 2. Mast have |
| keywords | list of strings | keywords | ["data mining", "structured data", "world wide web", "social network", "relational data"] | 1. One hot 2. Статистики (самый популярные, самый цитируемые, самые) |
| fos.name | string | paper fields of study | Web mining | 1. Облако тегов 2. Статистики |
| fos.w | float | fields of study weight | 0.659690857 | 1. Разобрать что это такое, связано на Fos.name или на имени автора |
| references | list of strings | paper references | ["4909282", "16018031", "16159250",  "19838944", ...] | 1. Потом когда будем делать модельки…. |
| n\_citation | int | citation number | 40829 | 1. Как цитируемость статьи зависит от количества человек и от организации 2. Как цитируемость зависит от языка |
| page\_start | string | page start | 11 |  |
| page\_end | string | page end | 18 | 1. Посчитать размер статьи отдельное переменной (article\_len) 2. Статистики для article\_len |
| doc\_type | string | paper type: journal, book title... | book | 1. Сделать статистики относительно doс\_type |
| lang | string | detected language | en | 1. Смотри n\_citation |
| publisher | string | publisher | Elsevier | Название издатель   1. Статистики |
| volume | string | volume | 10 | Разобраться с этим, что это…. |
| issue | string | issue | 29 | Тоже разобрать что это значит |
| issn | string | issn | 0020-7136 |  |
| isbn | string | isbn | 1-55860-489-8 |  |
| doi | string | doi | 10.4114/ia.v10i29.873 |  |
| pdf | string | pdf URL | //static.aminer.org/upload/pdf/1254/ 370/239/53e9ab9eb7602d970354a97e.pdf |  |
| url | list | external links | ["http://dx.doi.org/10.4114/ia.v10i29.873", "http://polar.lsi.uned.es/revista/index.php/ia/ article/view/479"] |  |
| abstract | string | abstract | Our ability to generate... | 1. Если есть время, повыдергивать слова и для них статистики... |
| indexed\_abstract | dict | indexed abstract | {"IndexLength": 164, "Inverte dIndex": {"Our": [0], "ability": [1], "to": [2, 7, ...]}} |  |