Languages

Level of linguistic self-assessment of Wikipedia users based on the latest revision of their user page. Each object is a json, one for each line of the file

Features

Raw

The features that are produced by the wikidump are shown as follows:

```
{
  "revisions": [
    {
      "num_languages_declared": 3,
      "user_id": 1426276,
      "user_name": "ChessBOT",
      "id": 58277661,
      "timestamp": "2012-07-26T19:44:53Z",
      "languages": [
        {
          "level": 6,
          "lang": "german"
        },
        {
          "level": 3,
          "lang": "english"
        },
        {
          "level": 1,
          "lang": "dutch"
        }
      ]
    }
  ],
  "namespace": 2,
  "id": 3715154,
  "title": "Soulman"
}
```

The `id` represents the user's id, the `title` indicates the user's username and the `namespace` is the namespace from which the information was obtained.

`languages` is an array which contains the self-assessed language knowledge levels of each languages known by the user.

You can consult here to see what the namespace corresponds to, in this case the `User` namespace.

Refactored

After the data refactor, the result is the following:

```
{
  "name": "Soulman",
  "id": 3715154,
  "languages": [
    {
      "level": 6,
      "lang": "german"
    },
    {
      "level": 3,
      "lang": "english"
    },
    {
      "level": 1,
      "lang": "dutch"
    }
  ],
  "num_languages_declared": 3,
  "edit_date": "2012-07-26T19:44:53Z"
}
```

Stats

The statistics remained as they were produced by the wikidump:

```
{
  "performance": {
    "pages_analyzed": 230305,
    "end_time": "2021-03-18 12:44:49.094082",
    "start_time": "2021-03-17 19:58:32.541808",
    "revisions_analyzed": 2498680
  },
  "users": {
    "languages": {
      "kirghiz": {
        "knowledge": [
          0,
          1,
          1,
          0,
          0,
          0,
          0
        ]
      },
      ...,
      "corsican": {
        "knowledge": [
          0,
          3,
          2,
          0,
          0,
          0,
          0
        ]
      }
    },
    "total": 13917,
    "num_unique_languages": 130
  }
}
```

`pages_analyzed` represents the number of users analyzed in the dump, hence the total number of users in the dump.

`Users` is the array representing the language and level of knowledge. The knowledge vector indicates, for each index, the number of users who have self-assessed with that level (value of the index) for that language.

The level ranges from 0 to 6 (representing N, i.e. native speaker level)

`num_unique_languages` is the number of element in the `Users` array.

`total` is the total number of users who have declared to know at least one language.

Wikibreaks

Wikibreaks and its category with the related parameters specified by the user on their user page or on the user talk page. Each object is a json, one for each line of the file.

Features

In this example there are two data which represent the user's wikibreak specified first in the user discussion page and then in the user page, where information such as name, category, subcategory and options are represented.

Raw

Revisions for user `Mago Merlino` associated to the user talk page:

```
{
  "revisions": [
    ...,
    {
      "user_name": "Lucas",
      "wikibreaks": [],
      "id": 27122869,
      "user_id": 34392,
      "timestamp": "2009-10-02T13:46:59Z"
    },
    ...,
    {
      "user_name": "Mago Merlino",
      "wikibreaks": [
        {
          "wikibreak_name": "occupato",
          "wikibreak_category": [
            "mental"
          ],
          "wikibreak_subcategory": "busy",
          "options": {
            "1": "[[Utente:Merlin89|Merlin89]]"
          },
          "at_least_one_parameter": true
        }
      ],
      "id": 27222996,
      "user_id": 386574,
      "timestamp": "2009-10-06T21:22:09Z"
    },
    ...,
    {
      "user_name": "Erik1991",
      "wikibreaks": [
        {
          "wikibreak_name": "occupato",
          "wikibreak_category": [
            "mental"
          ],
          "wikibreak_subcategory": "busy",
          "options": {
            "1": "[[Utente:Mago Merlino|Mago Merlino]]"
          },
          "at_least_one_parameter": true
        }
      ],
      "id": 29242443,
      "user_id": 318922,
      "timestamp": "2010-01-11T12:49:42Z"
    },
    {
      "user_name": "Mago Merlino",
      "wikibreaks": [],
      "id": 29458791,
      "user_id": 386574,
      "timestamp": "2010-01-20T13:07:52Z"
    }
    ...,
  ],
  "namespace": 3,
  "id": 2371347,
  "title": "Mago Merlino"
}
```

Revisions for user `Mago Merlino` associated to his/her user page:

```
{
  "revisions": [
    ...,
    {
      "user_name": "Mago Merlino",
      "wikibreaks": [],
      "id": 27137865,
      "user_id": 386574,
      "timestamp": "2009-10-03T09:32:45Z"
    },
    {
      "user_name": "Mago Merlino",
      "wikibreaks": [
        {
          "wikibreak_name": "occupato",
          "wikibreak_category": [
            "mental"
          ],
          "wikibreak_subcategory": "busy",
          "options": {
            "1": "[[Utente:Merlin89|Merlin89]]"
          },
          "at_least_one_parameter": true
        }
      ],
      "id": 27222993,
      "user_id": 386574,
      "timestamp": "2009-10-06T21:21:45Z"
    }
    ...,
    {
      "user_name": "Mago Merlino",
      "wikibreaks": [
        {
          "wikibreak_name": "occupato",
          "wikibreak_category": [
            "mental"
          ],
          "wikibreak_subcategory": "busy",
          "options": {
            "1": "Mago Merlino"
          },
          "at_least_one_parameter": true
        }
      ],
      "id": 27851636,
      "user_id": 386574,
      "timestamp": "2009-11-06T14:02:09Z"
    },
    {
      "user_name": "Mago Merlino",
      "wikibreaks": [],
      "id": 27851667,
      "user_id": 386574,
      "timestamp": "2009-11-06T14:03:27Z"
    }
    ...,
  ],
  "namespace": 2,
  "id": 2434621,
  "title": "Mago Merlino"
}
```

Refactored

Revisions for user `Mago Merlino` refactored and merged:

```
{
  "name": "Mago Merlino",
"id_talk_page": 2371347,
"ambiguous": false,
"id_user_page": 2434621,
"wikibreaks": [
  {
    "name": "occupato",
    "categories": [
      "mental"
    ],
    "subcategory": "busy",
    "to_date": "2009-11-06 14:03:27+00:00",
    "from_date": "2009-10-06 21:21:45+00:00",
    "parameters": [
      {
        "options": {
          "1": "[[Utente:Merlin89|Merlin89]]"
        },
        "timestamp": "2009-10-06 21:21:45+00:00"
      },
      {
        "options": {
          "1": "Mago Merlino"
        },
        "timestamp": "2009-10-06 21:21:45+00:00"
      }
    ]
  },
  {
    "name": "occupato",
    "categories": [
      "mental"
    ],
    "subcategory": "busy",
    "to_date": "2010-01-20 13:07:52+00:00",
    "from_date": "2009-10-06 21:22:09+00:00",
    "parameters": [
      {
        "options": {
          "1": "[[Utente:Merlin89|Merlin89]]"
        },
        "timestamp": "2009-10-06 21:22:09+00:00"
      },
      {
        "options": {
          "1": "[[Utente:Mago Merlino|Mago Merlino]]"
        },
          "timestamp": "2009-10-06 21:22:09+00:00"
        }
      ]
    }
  ]
}
```

The wikibreaks specified by the user are indicated with the date of specification, the date of removal and the respective list of parameters.

Parameters are represented as a list of options and timestamps which traks their changes.

If the `to_date` field is empty, it means that the pause is still present on the user page.

If one of the two fields, `id_talk_page` and `id_user_page`, are null, the information obtained does not come from that page.

The wikibreaks consist in the ones retrived from the user talk page and the user page.

As there may be inconsistent information, such as a template present on one side and not on the other, the `ambiguous` field has been introduced in order to specify whether there was a temporal overlap of the same template during the merge between the two pages.

Stats

```
{
  "performance": {
    "revisions_analyzed": 1385719,
    "pages_analyzed": 227581,
    "end_time": "2021-04-05 19:19:29.999694",
    "start_time": "2021-04-05 08:34:52.425883"
  },
  "wikibreaks": {
    "users_at_least_parameter": 102,
    "templates_at_least_one_parameter": 1175,
    "users": 252,
    "user_subcategories_occurences": {
      "user mental health": {
        "with_params": 0,
        "total": 1
      },
      "wikibreak": {
        "with_params": 79,
        "total": 154
      },
      "busy": {
        "with_params": 1,
        "total": 62
      },
      "exams": {
        "with_params": 20,
        "total": 69
      },
      "deceased wikipedian": {
        "with_params": 1,
        "total": 1
      },
      "retired": {
        "with_params": 1,
        "total": 6
      }
    },
    "templates": 5101,
    "user_categories_occurences": {
      "other": {
        "with_params": 2,
        "total": 7
      },
      "break": {
        "with_params": 99,
        "total": 204
      },
      "mental": {
        "with_params": 1,
        "total": 63
      },
      "health related": {
        "with_params": 0,
        "total": 1
      }
    }
  }
}
```

`users_at_least_parameter` number of users who have specified at least one parameter.

`templates_at_least_one_parameter` number of total wikibreaks templates with at least one parameter.

`users` total number of users who specified a wikibreaks.

`templates` total number of wikibreaks templates encountered.

`user_subcategories_occurences` dictionary which represents the total occurrences for each subcategory with at least one parameter.

`user_categories_occurences` dictionary which represents the total occurrences for each category with at least one parameter.

List of categories with subcategories associated: - Back/Not * can't retire * considering retirement * off and on wikibreak - break * wikibreak * in house * switch * at school * exams * vacation * out of town * personal issues - health-related * bonked * user grieving * user health inactive * user covid-19 * user mental health * user stress - mental * busy * discouraged * user contempt * user frustrated * user mental health * user stress - technical * computer death * no internet * no power * storm break - other * deceased wikipedian * not around * retired * semi retired * ex-wikipedia

Transcluded user warnings

Templates with parameters associated with the transcluded user warnings in a user talk page. Each object is a json, one for each line of the file.

Features

Raw

```
{
  "id": 13668,
  "title": "80.58.43.107",
  "revisions": [
    {
      "id": 118430,
      "user_name": "Llull",
      "user_warnings": [],
      "user_id": 7,
      "timestamp": "2004-07-18T13:05:09Z"
    },
    {
      "id": 275660,
      "user_name": "Joanjoc",
      "user_warnings": [
        {
          "lang": "ca",
          "at_least_one_parameter": false,
          "category": "not_serious",
          "options": {},
          "user_warning_name": "registre"
        }
      ],
      "user_id": 101,
      "timestamp": "2005-12-24T16:55:02Z"
    },
    {
      "id": 353758,
      "user_name": "SMP",
      "user_warnings": [],
      "user_id": 1734,
      "timestamp": "2006-03-09T13:42:02Z"
    }
  ],
  "namespace": 3
}
```

The `user_warnings` object in the `revisions` object contains all the user warnings associated with that particular revision with options, name and category.

Refactored

```
{
  "name": "Edd",
  "id_talk_page": 15535,
  "user_warnings_recieved": [
    {
      "category": "not_serious",
      "user_warning_name": "benvinguda",
      "transluded": true,
      "parameters": [
        {
          "options": {},
          "timestamp": "2004-12-03T11:11:59Z"
        }
      ]
    }
  ],
  "user_warnings_stats": {
    "2004": {
      ...
      "12": {
        "serious_transcluded": 0,
        "warnings_transcluded": 0,
        "warnings_substituted": 0,
        "not_serious_transcluded": 1,
        "serious_substituted": 0,
        "not_serious_substituted": 0
      }
    },
    ...,
    "2021": {
      "1": {
        "serious_transcluded": 0,
        "warnings_transcluded": 0,
        "warnings_substituted": 0,
        "not_serious_transcluded": 0,
        "serious_substituted": 0,
        "not_serious_substituted": 0
      }
     ...,
      "12": {
        "serious_transcluded": 0,
        "warnings_transcluded": 0,
        "warnings_substituted": 0,
        "not_serious_transcluded": 0,
        "serious_substituted": 0,
        "not_serious_substituted": 0
      }
    }
  }
}
```

The `user_warnings_stats` object contains statistics per month of the user warnings received by the user in question, while the `user_warnings_recieved` object contains the category of the warning received, the options, the name and any parameters.

Stats

```
{
  "user_warnings": {
    "total_user_talk_pages": 190338,
    "users": 58598,
    "users_at_least_parameter": 2058,
    "templates_at_least_one_parameter": 40267,
    "user_template_occurences": {
      "ca": {
        "av\u00eds d'edici\u00f3/p\u00e0gina/llista de b\u00e9ns culturals d'inter\u00e8s na
          "user_talk_occurences": 0,
          "user_talk_occurences_with_params": 0
        },
        ...,
        "benvinguda": {
          "user_talk_occurences": 56846,
          "user_talk_occurences_with_params": 4
        },
        ...,
        "ccbysanom\u00e9s": {
          "user_talk_occurences": 3,
          "user_talk_occurences_with_params": 3
        }
      }
    }
  }
  "categories": {
    ...,
  },
  "performance": {
    "pages_analyzed": 190338,
    "start_time": "2021-04-11 19:15:38.804847",
    "revisions_analyzed": 573753,
    "end_time": "2021-04-13 05:06:33.318531"
  }
}
```

`user_warnings` is the set of all warnings analyzed, including the total number of user talk pages that present at least one warning, users who have received at least one warning, with the classification if they have received one of them with at least one paremeter. In `user_template_occurences` there are the user warnings name with the occurrences associated. `categories` should have stored the occurrencies of each category, it seems to be buggy but it is not a problem since the information is saved in the features anyway.

Substituted user warnings

Each object is a json, one for each line of the file.

Features

Raw

```
{
  "namespace": 3,
  "revisions": [
    {
      "user_name": "VriuBot",
      "templates": [
        {
          "name": "benvinguda-taula",
          "category": "not_serious"
        }
      ],
      "id": 11154335,
      "timestamp": "2013-03-07T09:55:07Z",
      "user_id": 9468
    }
  ],
  "id": 9649,
  "title": "ArinArin"
}
```

The `templates` object contains the category and the name of the warning received. This association is probabilistic because it consists in searching for the most salient words present in a user warning template.

The words with the greatest value of the `tf-idf` metrics are chosen among all the other ones.

Refactored

```
{
  "name": "ArinArin",
  "user_warnings_stats": {
    "2013": {
      ...,
      "3": {
        "serious_transcluded": 0,
        "warnings_substituted": 0,
        "not_serious_substituted": 1,
        "warnings_transcluded": 0,
        "not_serious_transcluded": 0,
        "serious_substituted": 0
      },
      ...,
      "12": {
        "serious_transcluded": 0,
        "warnings_substituted": 0,
        "not_serious_substituted": 0,
        "warnings_transcluded": 0,
        "not_serious_transcluded": 0,
        "serious_substituted": 0
      }
    },
    "2021": {
      "1": {
        "serious_transcluded": 0,
        "warnings_substituted": 0,
        "not_serious_substituted": 0,
        "warnings_transcluded": 0,
        "not_serious_transcluded": 0,
        "serious_substituted": 0
      },
      ...,
    }
  },
  "user_warnings_recieved": [
    {
      "user_warning_name": "benvinguda-taula",
      "transluded": false,
      "parameters": [
        {
          "options": {},
          "timestamp": "2013-03-07T09:55:07Z"
        }
      ],
      "category": "not_serious"
    }
  ],
  "id_talk_page": 9649
}
```

Same as above but there couldn't be any options.

Stats

```
{
  "user_warnings_stats": {
    "total": 98034,
    "template_recognized": {
      "av\\u00eds d'edici\\u00f3/p\\u00e0gina/llista de b\\u00e9ns culturals d'inter\\u00e8s
        "category": "not_serious",
        "occurences": 55
      },
      "editant": {
        "category": "not_serious",
        "occurences": 1
      },
      "av\\u00eds d'edici\\u00f3/p\\u00e0gina/llista de b\\u00e9ns culturals d'inter\\u00e8s
        "category": "not_serious",
        "occurences": 123
      },
      ...,
      "welcome": {
        "category": "not_serious",
        "occurences": 784
      }
    }
  },
  "performance": {
    "pages_analyzed": 190338,
    "start_time": "2021-04-14 07:42:04.154495",
    "revisions_analyzed": 190338,
    "end_time": "2021-04-14 17:29:20.175345"
  }
}
```

The `user_warnings_stats` object contains the warning recognized, its category and the number of occurencies.

Substituted and transcluded user warnings

The resulting file is the union of the two refactored files by merging them. Each object is a json, one for each line of the file.

Features

Refactored

```
{
  "user_warnings_recieved": [
    {
      "category": "not_serious",
      "parameters": [
        {
          "timestamp": "2009-03-17T07:55:29Z",
          "options": {
            "1": "Civilitzaci\\u00f3 asteca"
          }
        }
      ],
      "transcluded": true,
      "user_warning_name": "av\\u00edsretiradaadq"
    }
  ],
  "user_warnings_stats": {
    "2004": {
      "1": {
        "not_serious_substituted": 0,
        "warning_transcluded": 0,
        "warning_substituted": 0,
        "serious_substituted": 0,
        "not_serious_transcluded": 0,
        "serious_transcluded": 0
      },
      ..
    },
    "2021": {
      ..,
      "12": {
        "not_serious_substituted": 0,
        "warning_transcluded": 0,
        "warning_substituted": 0,
        "serious_substituted": 0,
        "not_serious_transcluded": 0,
        "serious_transcluded": 0
      }
    }
  },
  "name": "Arnad\\u00ed",
  "id_talk_page": 19009
}
```

Stats

```
{
  "performance": {
    "start_time": "2021-04-11 19:15:38.804847",
    "end_time": "2021-04-14 17:29:20.175345",
    "pages_analyzed": 380676,
    "revisions_analyzed": 764091
  },
  "user_warnings": {
    "total_user_talk_pages_transcluded": 190338,
    "users_transcluded": 58598,
    "users_at_least_parameter_transcluded": 2058,
    "total_user_talk_pages_substituted": 98034,
    "user_template_occurences": {
      "ca": {
        "escolar": {
          "user_talk_occurences_transcluded": 0,
          "user_talk_occurences_with_params_transcluded": 0,
          "user_talk_occurences_substituted": 0
        },
        "av\\u00eds d'edici\\u00f3/p\\u00e0gina/llista de b\\u00e9ns culturals d'inter\\u00e
          "user_talk_occurences_transcluded": 0,
          "user_talk_occurences_with_params_transcluded": 0,
          "user_talk_occurences_substituted": 13
        },
        "benvinguda": {
          "user_talk_occurences_transcluded": 56846,
          "user_talk_occurences_with_params_transcluded": 4,
          "user_talk_occurences_substituted": 463
        },
        "welcome": {
          "user_talk_occurences_transcluded": 0,
          "user_talk_occurences_with_params_transcluded": 0,
          "user_talk_occurences_substituted": 784
        }
      }
    }
  },
  "categories": {
    "ca": {
      "warning": {
        "users_transcluded": 1,
        "users_substituted": 223,
        "total_transcluded": 0
      },
      "not_serious": {
        "users_transcluded": 1,
        "users_substituted": 100548,
        "total_transcluded": 0
      }
    },
    "it": {
      ..
    },
    "es": {
      ..
    },
    ..
  }
}
```

Merged version of transcluded and substituted user warnings