Anton Mishchuk, Oleh Kariuk, Volodymyr Byno, Volodymyr Maletskyi, Denis Porplenko

Mining Massive Datasets. Final project

Motivation

There are lots of articles in UA wiki that are not translated into EN (or other languages). When volunteers want to translate an article it is quite difficult for them to understand where to start. We are going to build a system that will recommend a UA article that will be likely quite popular in a target language.

Problem statement

Create a system that can recommend UA article which, after being translated, will be popular (in terms of page views) in EN wiki. (Actually, we can choose any target language, not only English)

Approach

We assume that the popularity of articles correlates with their categories. Knowing which categories more popular we can recommend the corresponding articles.
The following steps need to be implemented:

1. For each UA page extract categories.
2. For each translated UA page fetch pageviews of corresponding EN pageviews in 2018.
3. Build an ML model that uses “UA categories” as input and “EN pageviews” as output
4. Predict “EN pageviews” for UA pages without translation.
5. Rank pages based on the pageviews.

Data

UA wiki (https://dumps.wikimedia.org/ukwiki/20190420/) :

- ukwiki-20190420-page.sql - pages
- ukwiki-20190420-category.sql.gz - contains categories for articles
- ukwiki-20190420-categorylinks.sql - links from pages to categories
- ukwiki-20190420-langlinks.sql - links to other languages

EN wiki (https://dumps.wikimedia.org/enwiki/20190420/):

- enwiki-20190420-page.sql.gz - pages
- We suspect that the only way to get pageviews is HTTP API. ***Please assist here.***

Evaluation

Having “UA categories” as features and “EN pageviews” as output we train ML model. We can use k-fold cross-validation to measure model performance.