Matrix Generator

User manual

Table of Content

3.1.Download dumps and application....5

3.2.Setting paths....5

4.1User interface....6

4.2. Using application....8

5.1.Configuration....9

5.2.XML text processing....10

5.3.By words....11

5.4.By links....12

1. About this user manual

This document was created to instruct new users how to use the application called Matrix Generator. We tried our best to describe every function and feature of the application so that anyone can use it with ease and no trouble. However the application is still under development (as for January 2011) and it is possible that this manual may not cover the full functionality of Matrix Generator. Some of the described features may also change or become missing after this manual was finished.

We hope that this document will be helpful and that you will enjoy reading it and using Matrix Generator. When possible we tried to smuggle some sense of humor, you will judge if it was worth it ;-)

2. About Matrix Generator

Matrix generator is an application for:

- generating special data – matrix with value of similarity of articles
- fast processing Wikipedia dumps
- analyse connections between categories and/or articles
- working on Wikipedia data without Internet access
- advanced display and browse of category structure

Originally Matrix Generator was planned to be only a little program helping our group of four students in developing a different application. Plans however change and you never know when one month little support project changes into a big application which evolves for two semesters and takes a year to finish it… well, at least we hope to believe it is finished ;-)

During development there were many different ideas on how the project should evolve, some of them were later implemented, some failed and some got later excluded. Unfortunately usually the coolest ideas fail or get excluded, that is why Matrix Generator is maybe not very cool, but it is functional instead.

Application has many features and options, it is optimized for performance, good thing that you have not seen how slow was the 1.03 version – since that time we have boosted the performance a lot. We tried to make it resistant to errors but there may be still some situations that we have not foreseen.

The application is run together with a Console window, closing it would close the whole application, we decided to make it like this because 90% of information during run-time is written to the Console. It was the easiest way to quickly present a lot of information, especially because the code was written by more than one person at a time (everyone coded everything in his own way :-P) and it would be difficult to accomplish that in another way.

3. Getting started

3.1. Download dumps and application

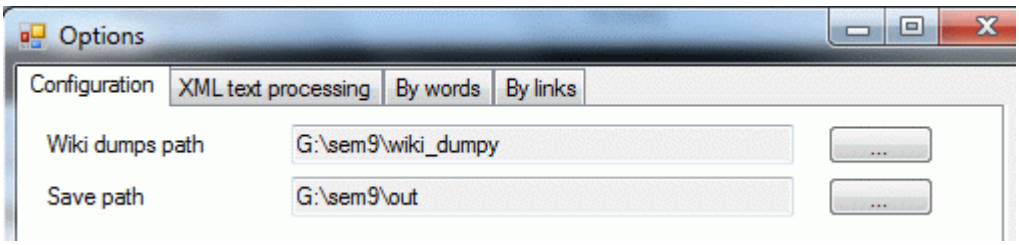
First you need to download Matrix Generator application and save it to disk, it does not require installation, you can just run it.

Next you need to download wikipedia dump files from the url: http://dumps.wikimedia.org, there you must choose enwiki or plwiki or simplewiki and download the files:

- *-page.sql
- *-pagelinks.sql
- *-categorylinks.sql
- *-redirect.sql
- *-pages-articles.xml

Save the files and unpack it to a directory like D:\wiki_dumps and remember the path because you will need it later.

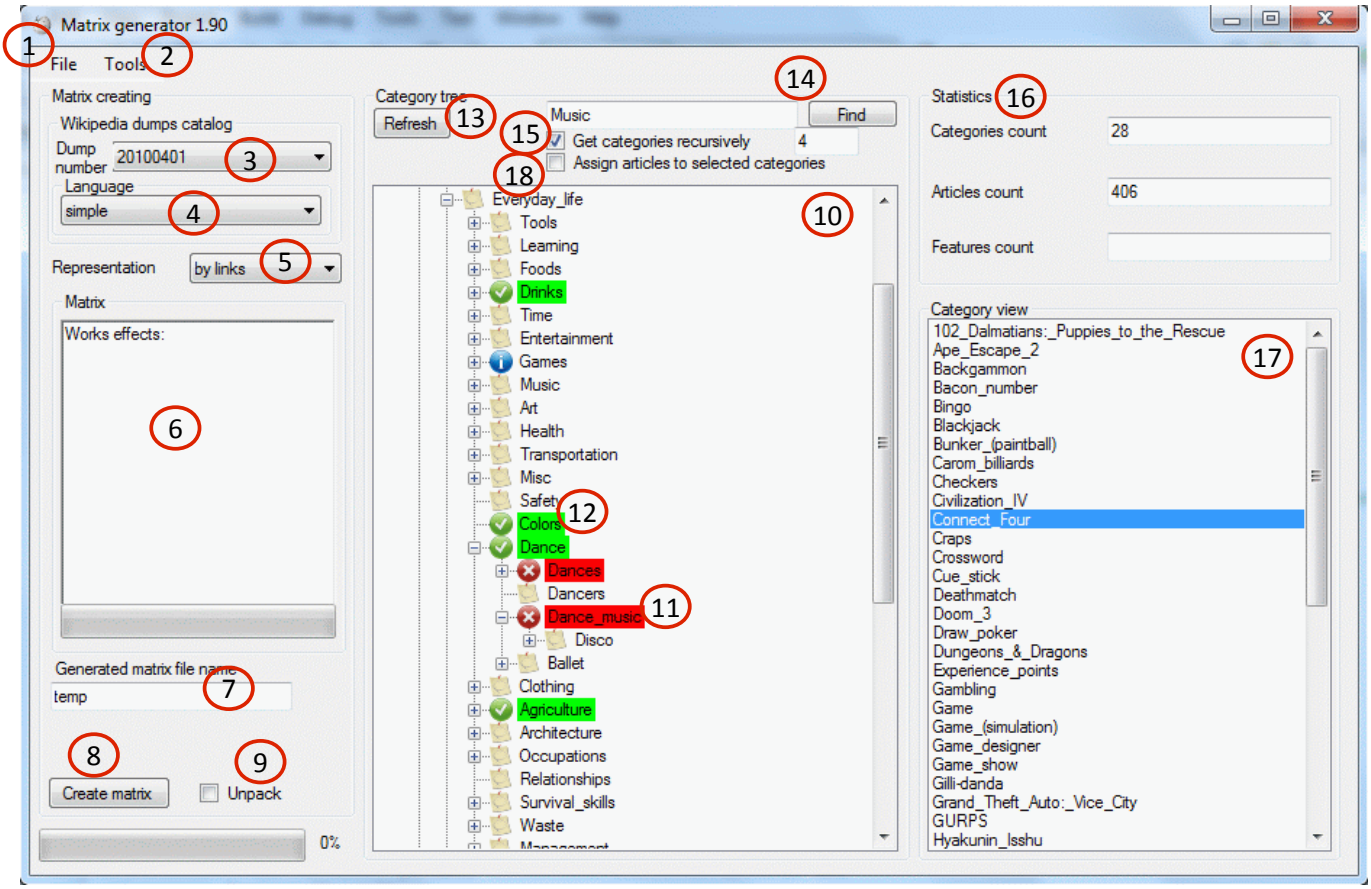
3.2. Setting paths



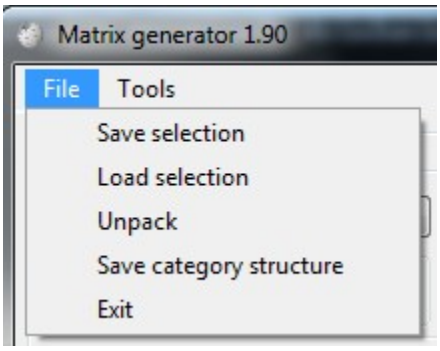
We need to launch the application and chose Options from Tools menu. Then we need to set path to wiki dumps catalog and to catalog for output files. This steps is all that you need to work with application.

4. Main application window

4.1 User interface



1) File Menu



Save selection – save selection of categories

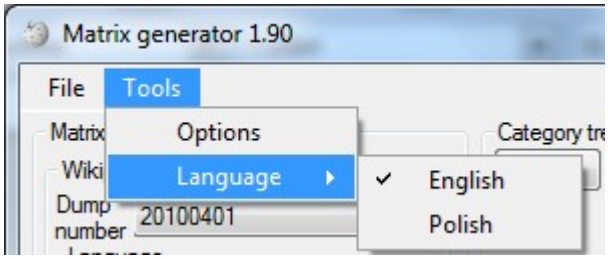
Load selection - load previously saved selection of categories

Unpack – unpack packed matrix

Save category structure – save structure od selected category

Exit – Exit and go for a beer

2) Tools



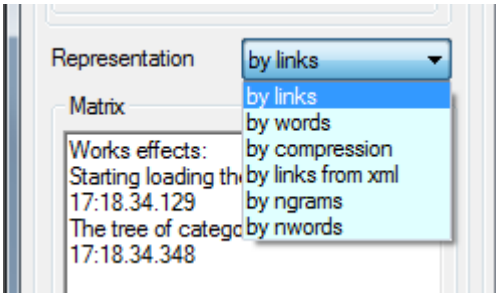
Options – options of application

Language – application languages

3) Dump number – number of dump in wikipedia catalog – date released

4) Wiki language – for application only simple, pl and en

5) Representation



1. by links – representation using links from articles
2. by words – representation using words from xml articles
3. by compression - representation using compare of compressed pairs of articles
4. by links from xml – the same like in by links, but we get links from xml dump, not from sql
5. by ngrams - representation using ngrams from xml articles
6. by nwords - representation using nwords from xml articles

6) Work effects – although Console gives more and better information

7) Matrix name – prefix of output files names

8) Create matrix – go go go!!!

9) Unpack – matrix will be unpacked when we click Create matrix

10) Category viewer – in this area we see categories tree

11) Deselected categories – category that will not be processed (with her childs)

12) Selected categories – category that will be processed

13) Refresh – refresh category view – also click in this button will generate application files

14) Find – we can find category with this textbox and button, we can use regex (* for unidentified number of characters, ? for one character)

15) Recursively – we can assign depth level

16) Statistics – when we select category in this labels we see number of categories and articles in selected space, labels feature count display only when we click create matrix

17) Category view – this listBox displays all articles from Category

18) Assign articles to selected category – when checked, this will assign articles only to selected categories, when unchecked – articles will be assigned to their original category

4.2. Using application

When we have downloaded some wiki dumps and set the paths, we can start using application.

1. First we need to select the wiki language that we want to use (4) and then select the dump number (3) which is the date of release. Then we can click Refresh (13) to see the category viewer (10) – during the first time for each dump version there will be application files generated.

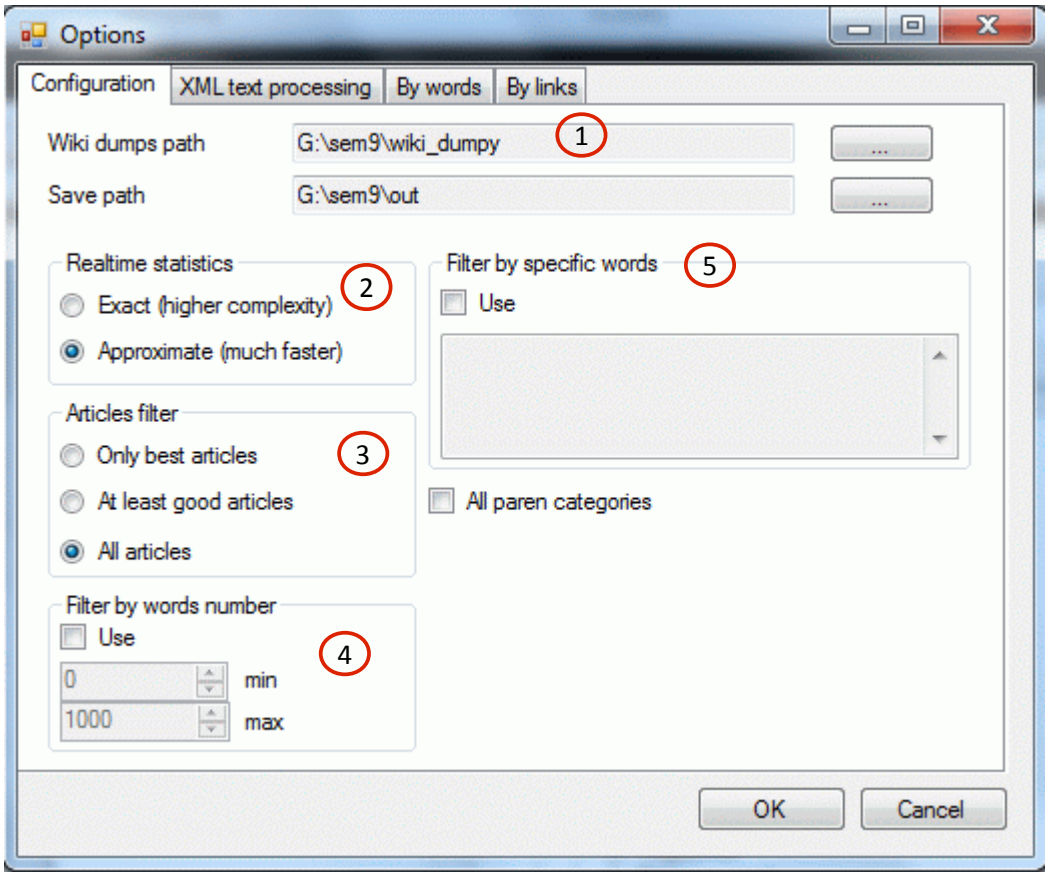
2. Now we can view the category tree, select categories by right mouse click, mark as deselected by second right click and remove any selection by third right click. We can search categories (14), search supports regex (look 4.1.14). We can choose options (15, 18).

3. After completing points 1 and 2 we can finally choose representation (5) and a prefix for output files (7). We can choose if we want the matrix to be unpacked right away (9) and click Create matrix (8) to get the job done.

4. GL&HF (good luck and have fun)!

5. Options

5.1. Configuration



1) Application paths – in text boxes we have catalog with wikipedia dumps and with output files

2) Realtime statistics – we can choose method for realtime statistics:

- exact – slow and get many memory, but it's exact

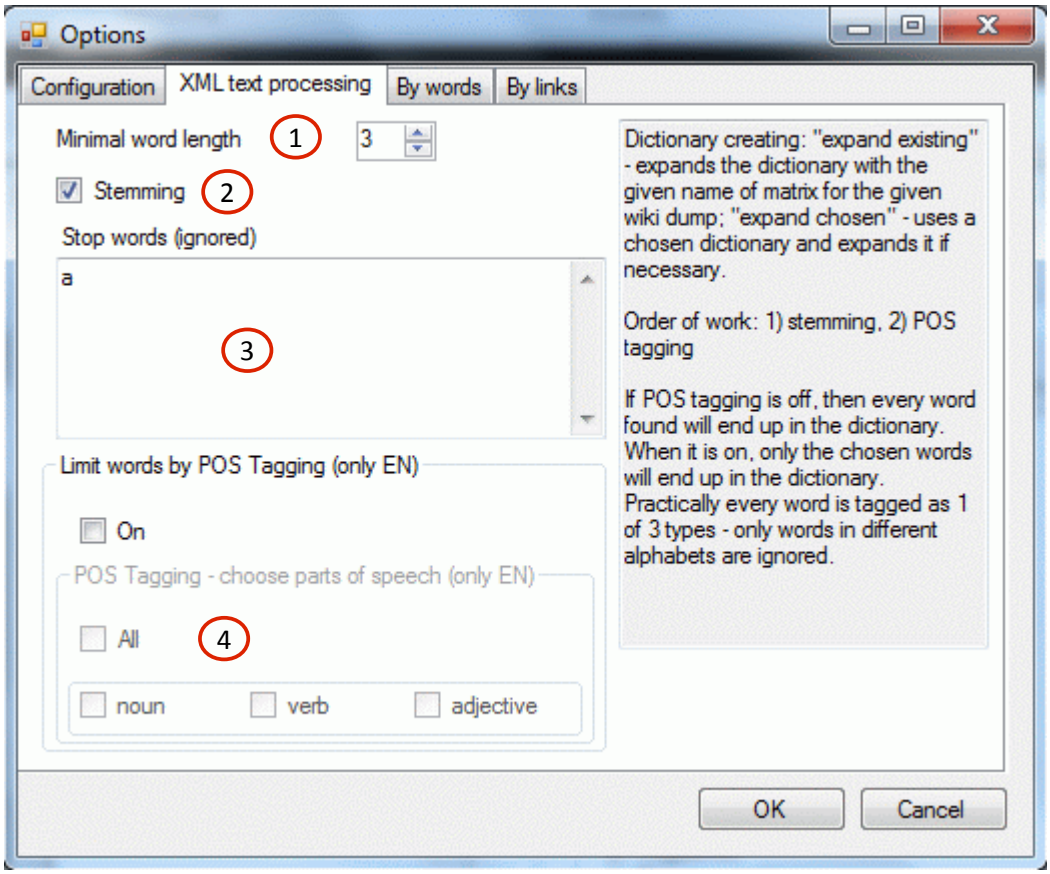
- approximate – fast and use little memory, but it's not exacts

3) Article filter – filters articles from categories “best articles” and “good articles”

4) Filter by words number – filters articles by number of words in this articles

5) Filter by specific words – filters only articles with all of specific words

5.2. XML text processing



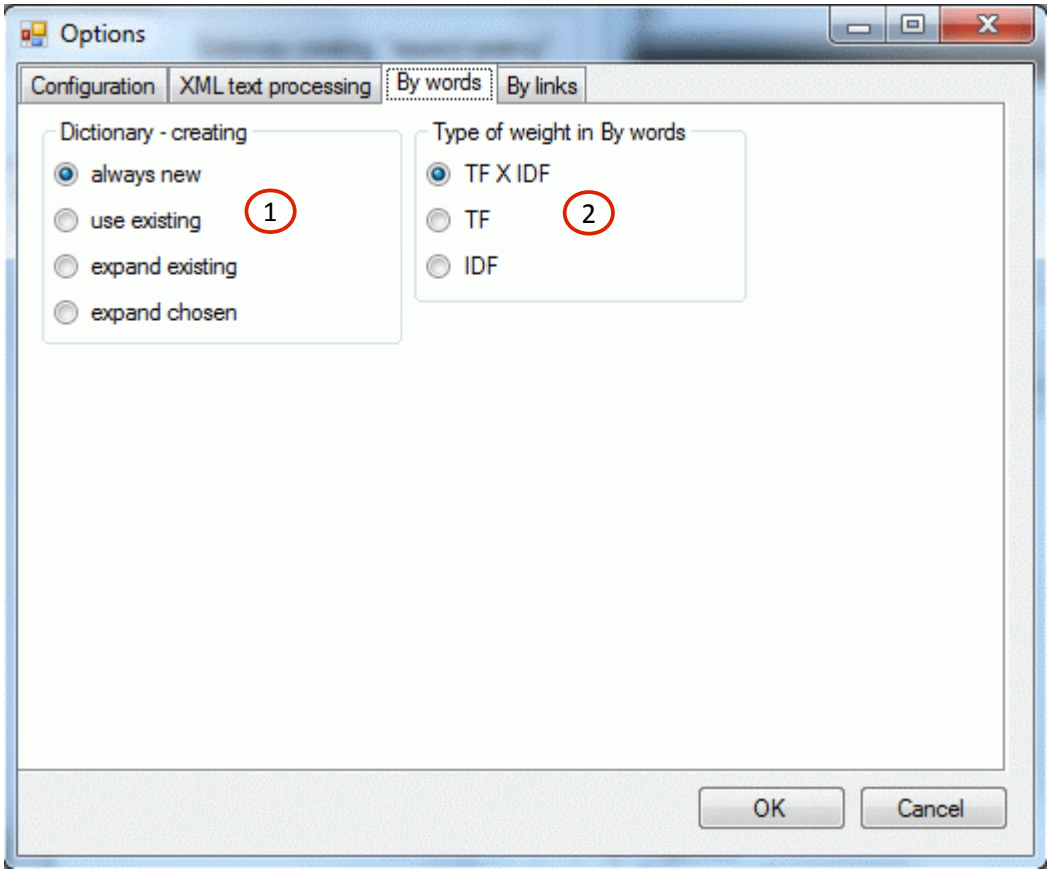
1) Minimal word length – shorter words will be annihilated!

2) Stemming – words will be brought to its base form

3) Stop words – words from this list will be ignored, separate words using spacebar or semicolon or comma

4) POS tagging – will include (in output) selected parts of speech (all words will we tagged as one of the possible types)

5.3. By words



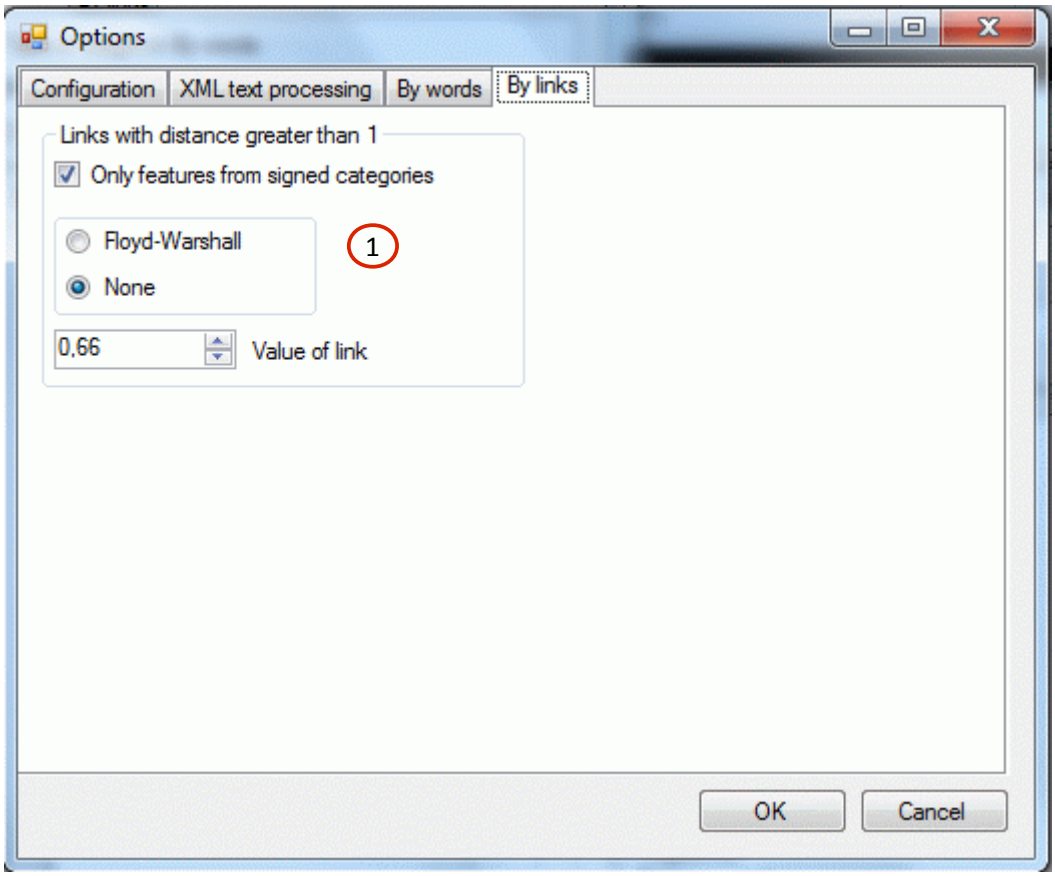
1) Dictionary is used in representation “by words”

- always create a new dictionary or
- use existing dictionary – works when you don't change prefix of the output files – uses the dictionary generated in earlier execution of matrix creation
- expand existing – as above but will add new words if necessary
- expand chosen – this will let you choose an existing dictionary to use

2) Type of weight – this is type of weight in matrix when we chose representation by words, we can use

- TF X IDF – product of Term Frequency and Inverse Document Frequency
- TF – Term Frequency
- IDF – Inverse Document Frequency

5.4. By links



1) Links with distance greater than – in this group box we can choose method to calculate distance greater than 1, for now we only can use algorithm Floyd-Warshall

6. Output files

In output catalog we have six files (or five if don't checked unpacked):

- articles_dict – file with articles dictionary, we have article name and his id
- feature_dict – file with features dictionary, we have feature name and his id, what does mean “feature”? In different representation features are different things:
  - by links/links from xml – features are links from articles
  - by words/nwords/ngrams – features are words/nwords/ngrams from articles
  - by compression – features are the same thing like articles dictionary, because we compress articles and compare each other
- cats_dict – file with categories dictionary, we have all categories from selected category space and their names
- categories – file with structure of categories from selected category space
- lista – packed matrix
- matrix – like name says it's the output matrix, this is how it looks

```
fc   f0    f1   …  fn
a0 v00 v01 … v0n
a1 v10 v11 … v1n
…    …    …   …   …
an vn0 vn1 …  vnn
```

fc – features count
fx – feature, where x is id of feature
ax – article, where x is id of articles
vxy – value of feature x for article y