Project 2018 – Multi-domain social recommender system

Use the WikiMID dataset downloadable from http://wikimid.tweets.di.uniroma1.it/wikimid/ (use WikiMID dataset in the form of tab separated values (TSV)). For about 500000 users (English and Italian users), it associates a set of interests extracted from their messages or from their friendship list (by selecting those friends in the list that indicate an **interest** more than a peer friendship relation). A Wikipedia page is associated to every interest. Read more in the related paper. Use only the ENGLISH dataset.

1. For every user u_i , given his/her Wikipedia-mapped interests, finds a set of CATEGORIES representing a synthesis of the shown preferences. In general, you should have less categories than interests (categories should "synthesize" a user's main interests).

$$u_i{:}\ \left(W_1^i, W_2^i, \dots W_k^i\right) \rightarrow\ c_1^i, c_2^i, .. c_{k_i}^i$$

Example:

user_176236916

Interests:

WIKI:EN:American_Tabloid
WIKI:EN:Chester_Himes
WIKI:EN:The_Outsider_(Wright_novel)
WIKI:EN:Sidetracked_(novel)
WIKI:EN:The_Human_Factor_(Graham_Greene_book)
WIKI:EN:Colin_Cotterill
WIKI:EN:Whispers_Under_Ground
WIKI:EN:Adrian_McKinty
WIKI:EN:Monkey_Man_(The_Rolling_Stones_song)
WIKI:EN:Sgt._Pepper's_Lonely_Hearts_Club_Band_(song)
WIKI:EN:Band_of_the_Castle_Guards_and_the_Police_of_the_Czech_Republic
Etc.

Preferred interest categories[1]**:**

Series
Writers
Journalists
Magazines
Politicians
Seasons
Drama
High_schools
Networks
Actors
Screenwriters
Television
Band
Companies
Etc.

[1] Nothe that there is no pairwise correspondence between wikipages and semantic interests. The latter are collectively extracted from the full set of wikipages for each user.

You may use any semantic resource (Wikipedia categories, DBPedia, Babelnet..) and any method you can invent or find in literature (we do not expect anything particularly innovative, so don't worry)

2. Generate clusters G_j^S of similar users (i.e. with similar interests) – using a whatever community detection method among those presented in class.

3. Use a whatever simple method to evaluate clusters (e.g. average distance between cluster members and non-cluster members: you want that any two elements in a cluster are more similar to each other than any two elements belonging each to a different cluster).

4. You are further given a set of 1500 Twitter IDs, file **S21.tsv**. Using Twitter API, download profile and friendship information, and try to associate each user u_j with the cluster of other users most similar to u_j. Explain the adopted similarity method.

5. You are given 500 additional users (file **S22-preferences.tsv**), for which you have the user ID and the list of preferred items in terms of Wikipedia pages. You are further given for each user a list of 6 Wikipedia pages (file **S23.tsv**). Recommend to each user 3 out of the 6 proposed items. In selecting 3 items and discarding the other 3, you should define and implement some algorithm that ranks the 6 items according to the "induced" user's interests. Explain the method you use to decide which recommendations are more likely to fit each user's interests.