Text Analysis of Social Networks: Working with FB.com and VK.com Data

Seminar of CENTAL, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Alexander Panchenko

panchenko@lt.informatik.tu-darmstadt.de

Language Technology Group, TU Darmstadt, Germany

November 27, 2015

Outline

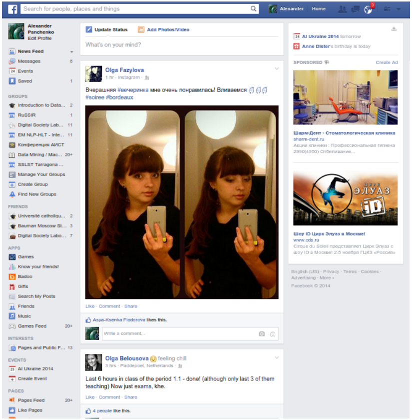
Outline

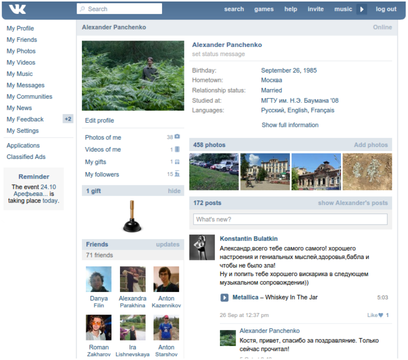
Acknowledgment: Digital Society Laboratory LLC

http://socialkey.ru/

Social networks from the users's standpoint

Facebook (FB) and VKontakte (VK)





Social networks from the data miner's standpoint

Facebook (FB) and **VKontakte (VK)**

- **Profiles**: a set of user attributes
  - categorical variables (region, city, profession, etc.)
  - integer variables (age, graduation year, etc.)
  - text variables (name, surname, etc.)
- **Network**: a graph that relates users
  - friendship graph
  - followers graph
  - commenting graph, etc.
- **Texts**:
  - posts
  - comments
  - group titles and descriptions

Gathering of VK and FB data

- **Big Data**: VK worth tens or even hundreds of TB
- **Decide** what do you need (posts, profiles, etc.).
- **Download**:
  - API
  - Scraping
- **Download limits** and **API limitations** are specific for each network.
- **Parallelization** is very practical, especially horizontal one:
  - Amazon EC2, Distributed Message Queues

Storing VK and FB data

- Again, **Big Data**
- **NoSQL** solutions are helpful
- **Raw data**: Amazon S3
- **For analysis**: HDFS
- **Efficient retrieval**: Elastic Search

Outline

Social Network Analysis

- **Structure analysis**: friendship graph, comments graph, etc.
- **Content analysis:** profile attributes, posts, comments, etc.
- **Combined approaches**.

What scientific communities analyze social networks?

- 60s – the first structural methods
- 00s – online social network analysis boom
- **Social Network Analysis** community (Sociologists, Statisticians, Physicists)
- **Data and Graph Mining** community
- **Natural Language Processing** community

Technologies for analysis of social networks

- Machine Learning: **hidden vs observable** user attributes
- **Training** of the model often can be scaled vertically

python™ scikit learn

- **Applying** the model should be scaled horizontally

hadoop Spark

Outline

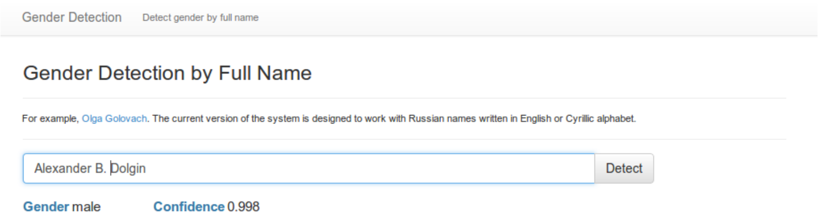
Problem

Joint work with Andrey Teterin.

- **Detect gender of a user**
  - to profile a user;
  - user segmentation is helpful in search, advertisement, etc.
- **By text written by a user**:
  - Ciot et al. [2013], Koppel et al. [2002], Goswami et al. [2009], Mukherjee and Liu [2010], Peersman et al. [2011], Rao et al. [2010], Rangel and Rosso Rangel and Rosso [2013], Al Zamal et al.Al Zamal et al. [2012] and Lui et al. Liu et al. [2012].
- **By full name**: Burger et al. [2011], Panchenko and Teterin [2014]

Online demo

http://research.digsolab.com/gender



Training Data

- 100,000 full names of Facebook users with known gender
- full name – first and last name of a user
- gender: male or female
- names written in both Cyrillic and Latin alphabets
- "Alexander Ivanov", "Masha Sidorova", "Pavel Nikolenko", etc.

Training Data

		Ivanova	Ivanov	Kuznetsova	Kuznetsov	Vasilyeva	Smirnov	Smirnova	Petrov	Shevchenko	Popova	Petrova	Popov	Bondarenko	Morozova	Volkova	Novikova	Sokolova	Mihailova	Vasilyev	Kovalenko	Romanova	Pavlova	Andreeva	Kravchenko	Alekseeva	Kim
		264	251	167	131	130	128	126	117	116	115	115	106	105	96	94	92	89	89	88	83	81	81	76	74	71	70
3193	Aleksandr	0	25	0	13	0	16	0	11	7	0	0	16	6	0	0	0	0	0	12	4	0	0	0	4	0	4
2650	Elena	19	0	11	0	11	0	13	0	3	9	7	0	7	5	11	11	4	5	0	3	5	7	3	3	4	2
2620	Sergey	0	20	0	6	0	13	0	5	1	0	0	5	11	0	0	0	0	0	9	6	0	0	0	2	0	0
2222	Tatyana	12	0	10	0	10	0	9	0	7	8	11	0	0	13	4	4	9	5	0	1	0	6	4	3	5	2
2174	Olga	19	0	14	0	12	0	7	0	2	7	6	0	2	7	7	4	5	0	0	4	6	2	3	1	0	3
1976	Andrey	0	16	0	10	0	11	0	8	3	0	0	7	1	0	0	0	0	0	3	2	0	0	0	1	0	1
1914	Irina	16	0	6	0	5	0	8	0	0	5	7	0	1	3	4	4	10	2	0	2	8	3	6	2	3	1
1895	Natalya	14	0	13	0	6	0	4	0	1	5	5	0	4	9	3	6	2	7	0	1	3	3	5	2	2	1
1793	Aleksey	0	13	0	7	0	6	0	10	1	0	0	7	4	0	0	0	0	0	1	1	0	0	0	1	0	1
1721	Dmitry	0	14	0	8	0	8	0	3	5	0	0	8	4	0	0	0	0	0	4	1	0	0	0	0	0	0
1576	Svetlana	12	0	6	0	6	0	4	0	1	5	5	0	0	1	6	10	4	3	0	1	1	4	2	2	5	1
1449	Vladimir	0	13	0	5	0	4	0	7	1	0	0	2	5	0	0	0	0	0	2	0	0	0	0	3	0	4
1399	Yulia	4	0	9	0	3	0	7	0	4	1	0	0	1	0	1	2	2	3	0	3	1	1	1	0	3	2
1348	Anna	10	0	7	0	6	0	7	0	0	3	6	0	2	3	1	0	7	5	0	0	4	3	4	0	1	2
1216	Ekaterina	8	0	5	0	5	0	5	0	5	1	3	0	2	4	5	4	5	5	0	3	3	3	2	0	2	0
1199	Marina	8	0	5	0	5	0	4	0	0	6	5	0	1	4	5	2	3	4	0	1	1	4	3	2	4	3
1154	Evgeny	0	8	0	3	0	4	0	3	3	0	0	7	4	0	0	0	0	0	4	1	0	0	0	2	0	2
945	Igor	0	6	0	4	0	3	0	4	2	0	0	1	2	0	0	0	0	0	3	0	0	0	0	1	0	1
920	Anastasiya	5	0	7	0	5	0	3	0	1	0	1	0	0	2	3	3	2	1	0	1	6	0	0	3	2	0
857	Mariya	7	0	0	0	1	0	2	0	0	3	4	0	1	3	1	3	1	1	0	0	2	6	1	0	2	0
846	Oleg	0	5	0	3	0	5	0	2	2	0	0	3	0	0	0	0	0	0	1	1	0	0	0	1	0	2
822	Mihail	0	8	0	2	0	5	0	3	2	0	0	1	0	0	0	0	0	0	2	1	0	0	0	1	0	0
783	Ludmila	5	0	5	0	4	0	3	0	3	0	0	0	1	3	4	2	1	3	0	0	3	3	3	2	2	0
745	Oksana	5	0	1	0	1	0	0	0	3	2	3	0	1	2	1	1	1	2	0	4	0	3	0	0	1	0

Character endings of Russian names

- 72% of first names have typical male/female ending
- 68% of surnames have typical male/female ending
- a typical male/female ending splits males from females with an error less than 5%
- gender of $\geq$ 50% first names recognized with 8 endings
- gender of $\geq$ 50% second names recognized with 5 endings

Conclusion

Simple symbolic ending-based method cannot robustly classify about 30% of names. This motivates the need for a more sophisticated statistical approach.

Character endings of Russian names

Type	Ending		Gender	Error, %	Example
first name	na	(на)	female	0.27	Ekateri**na**
first name	iya	(ия)	female	0.32	Anastas**iya**
first name	ei	(ей)	male	0.16	Serg**ei**
first name	dr	(др)	male	0.00	Alexan**dr**
first name	ga	(га)	male	4.94	Sere**ga**
first name	an	(ан)	male	4.99	Iv**an**
first name	la	(ла)	female	4.23	Luidmi**la**
first name	ii	(ий)	male	0.34	Yur**ii**
second name	va	(ва)	female	0.28	Morozo**va**
second name	ov	(ов)	male	0.21	Objedk**ov**
second name	na	(на)	female	2.22	Matyushi**na**
second name	ev	(ев)	male	0.44	Serge**ev**
second name	in	(ин)	male	1.94	Teter**in**

Table: Most discriminative and frequent two character endings of Russian names.

Gender Detection Method

- input: a string representing a name of a person
- output: gender (male or female)
- binary classification task

Features

- endings
- character *n*-grams
- dictionary of male/female names and surnames

Model

- L2-regularized Logistic Regression

Features

Character *n*-grams

- males: Alexander Yaroskavski, Oleg Arbuzov
- females: Alexandr**a** Yaroskavsk**aya**, Nayali**ya** Arbuzov**a**
- BUT: "Sidorenko", "Moroz" or "Bondar"!
- two most common one-character endings: "a" and "ya" ("я")

Dictionaries of first and last names

- probability that it belongs to the male gender: $P(c = male|firstname)$, $P(c = male|lastname)$.
- 3,427 first names, 11,411 last names

Results

Model	Accuracy	Precision	Recall	F-measure
rule-based baseline	0,638	**0,995**	0,633	0,774
endings	0,850 ± 0,002	0,921 ± 0,003	0,784 ± 0,004	0,847 ± 0,002
3-grams	0,944 ± 0,003	0,948 ± 0,003	0,946 ± 0,003	0,947 ± 0,003
dicts	0,956 ± 0,002	**0,992 ± 0,001**	0,925 ± 0,003	0,957 ± 0,002
endings+3-grams	0,946 ± 0,003	0,950 ± 0,002	0,947 ± 0,004	0,949 ± 0,003
3-grams+dicts	0,956 ± 0,003	0,960 ± 0,003	0,957 ± 0,004	0,959 ± 0,003
endings+3-grams+dicts	**0,957 ± 0,003**	0,961 ± 0,003	**0,959 ± 0,004**	**0,960 ± 0,002**

Table: Results of the experiments on the training set of 10,000 names. Here *endings* – 4 Russian female endings, *trigrams* – 1000 most frequent 3-grams, *dictionary* – name/surname dict. This table presents precision, recall and F-measure of the female class.

Results

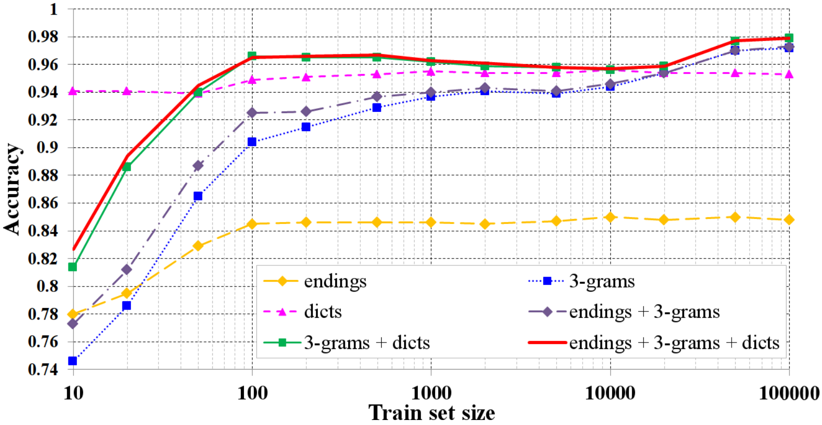


Figure: Learning curves of single and combined models. Accuracy was estimated on separate sample of 10,000 names.

Results

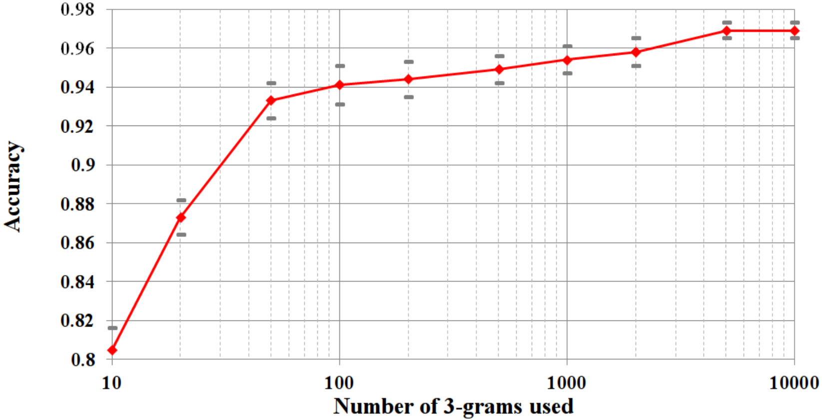


Figure: Accuracy of the model *3-grams* function of the number of features used k.

Error Analysis

There are several types of errors:

- Inconsistent annotation, such as "Anna Kryukova (male)" or "Boris Krolchansky (female)".
- Name string is neither male nor female, but rather a name of a group, e.g. "Wikom Tools", "Kazakh University of Humanities" or "Privat Bank".
- Name string represents a foreign name, e.g. "Abdulloh Ibn Abdulloh", "Brooke Alisson", "Ulpetay Niyetbay" or "Yola Dolson". Our model was not trained to deal with such names.
- Meaningless or partially anonymized names names, e.g. "Crazzy Ma", "Un Petit Diable", "Vv Tt", "Vio La Tor" or "Muu Muu". Additional information is required to derive gender of such users.
- People with rare names or surnames, e.g. "Guldjan Reyzova", "Yagun Zumpelich" or "Akob Saakan". These are people with

Error Analysis

	Train Set Errors		Test Set Errors	
	name	true class	name	true class
1	Lea Shraiber	female	Ilya Nadorshin	male
2	Profanum Vulgus	female	Rustem Saledinov	male
3	Anna Kryukova	male	Erkin Bahlamet	male
4	Gin Amaya	male	Gocha Lapachi	male
5	Gertrud Gallet	female	Muttaqiyyah Abdulvahhab	female
6	Dolores Laughter	female	Yola Dolson	female
7	Di Nolik	male	Heiran Gasanova	female
8	Jlija Hotieca	female	Hadji Murad	male
9	Gic Globmedic	female	Jenya Chekulenko	female
10	Ulpetay Niyetbay	female	Tury.Ru Domodedovskaya Metro Office	male
11	Olga Shoff	male	Elmira Nabizade	female
12	Phil Golosoun	male	Niko Liparteliani	male
13	Tsitsino Shurgaya	female	Oleg Grin'	male
14	Anna Grobov	female	Santi Zarovneva	female
15	Linguini Incident	female	Misha Badali	male
16	Toma Oganesyan	female	Che Serega	male
17	Swon Swetik	female	Petr Kiyashko	male
18	Adel Simon	female	Sandugash Botabaeva	female
19	Ant Kam-	male	Jenya Sergienko	female
20	Xristi Xitrozver	female	Abdulloh Ibn Abdulloh	female
21	Anii Reznookova	female	Naikaita Laitvainenko	male
22	Aurelia Grishko	male	Fil Kalnitskiy	male
23	Alex Bu	female	Helen Hovel'	female
24	Karen Karine	female	Valery Kotelnikov	male
25	Russian Spain	female	Max Od	male
26	Lucy Walter	male	Jean Kvartshelia	male
27	Aysah Ahmed	female	Adjedo Trupachuli	female

Outline

Problem

Motivation

- Goal: to detect Russian-speaking users
- Cyrillic alphabet is used also by Ukrainian, Belorussian, Bulgarian, Serbian, Macedonian, Kazakh, etc

Research Questions

- Which method is the best for Russian language?
- How to adopt it to the FB profile?

Contributions

- Comparison of Russian-enabled language detection modules.
- A technique for identification of Russian-speaking users.

Method

- input: a FB user profile
- output: is Russian-speaker? (or a set of languages user speaks)

Common Russian character trigrams

"на ", " пр", " то", " не", " ли", " по", "но ", " в ", " на", " ть", " не", " и ", " ко", " ом", "про", "то ", " их", " ка", "ать", "ото", " за", " ие", "ова", "тел", "тор", " де", "ой ", "сти", " от", "ах ", " ми", "стр", " бе", " во", " ра", "ая ", "ват", "ей ", "ет ", " же", "иче", "ия ", "ов ", "сто", " об", "вер", "го ", "и в", "и п", "и с", "ии ", "ист", "о в", "ост", "тра", " те", "ели", "ере", "кот", "льн", "ник", "нти", "о с"

Existing modules for language identification

- **langid.py**
  - `https://github.com/saffsd/langid.py`
  - Advanced n-gram selection
- **chromium compact language detector (cld)**
  - https://code.google.com/p/chromium-compact-language-detector
- **guess-language**
  - `https://code.google.com/p/guess-language`
- **Google Translate API**
  - `https://developers.google.com/translate/v2/using_rest#detect-language`
  - 20$/1M characters
- **Yandex Translate API**
  - `http://api.yandex.ru/translate`
  - Free of charge, 1M of characters / day (by September 2013)
- **Many more**, e.g. `language-detection` for Java

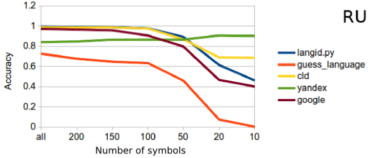
DBpedia Dataset

Language	Dataset	Number of texts	Size
RU	Dbpedia short abstracts	435058	Big
RU	Dbpedia labels	361148	Big
BG	Dbpedia short abstracts	85448	Big
BG	Dbpedia labels	77778	Big
RU	Dbpedia short abstracts	750	Small
BG	Dbpedia short abstracts	750	Small
EN	Dbpedia short abstracts	750	Small

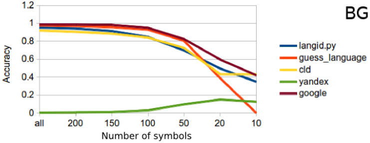
Accuracy of Different Language Detection Modules

Dbpedia short abstracts, Small

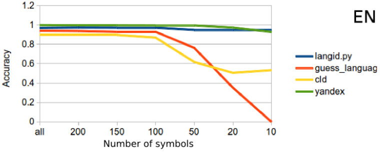
RU



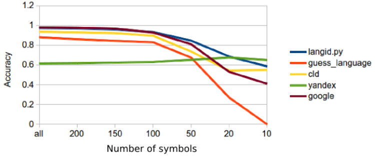
BG



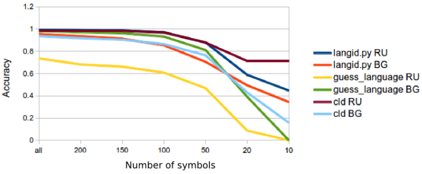
EN



Dbpedia short abstracts, Small (avg.)



Dbpedia short abstracts, Big (avg.)



Facebook Dataset: Method

- **Profile text**: posts + comments + user names – Latin symbols.
- **Profile text length**: 3,367 +- 17,540
- **Russian-speakers**: P(ru) > 0.95
- **Core Russian-speakers**:
    - P(ru) > 0.95
    - # Cyrillic symbols >= 20%
    - locale is `ru_RU`

Facebook Dataset: Results

- **9,906,524** public FB profiles (>= 50 cyr. characters)
- 8,687,915 **(88%)** Russian-speaking users
- 3,190,813 **(32%)** core Russian-speaking public Facebook users
- 5,365,691 **(54%)** of profiles with no profile text (<= 200 characters)

Outline

Problem

Joint work with Dmitry Babaev and Sergei Objedkov.

- **input**: *some* SN data representing a user
- **output**: list of user interests

Motivation

- Advertisement: targeting, user segmentation, etc.
- Recommendations of content and friends
- Customization of user experience
- ...

Data: FB and VK groups

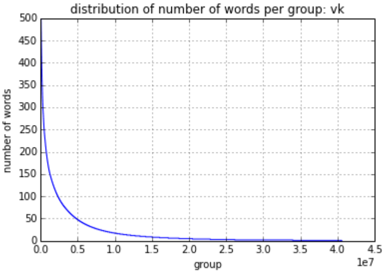
Text corpus

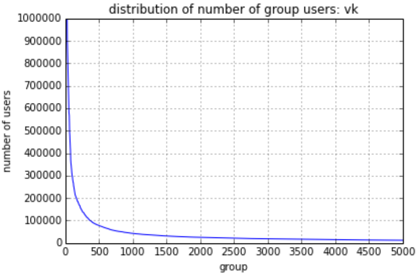
- 41 million of VK groups
- 11 million of FB publics
- 1.5 million of FB groups

Data format

- Title and/or description
- List of members
- Number of comments, likes, posts by a member

Data: VK groups





253 interests detected by our system

academy, advertising_offline, advertising_online, agrarian_univ, air_sports, alcohol_drinks, american_auto, animals, aquabike, aquatics, architecture, armed_forces, art_school, art_univ, art_vocational, asian_auto, auction_house, auto, auto_chemicals, auto_class_a, auto_class_b, auto_class_c, auto_class_d, auto_class_e, auto_class_f, auto_class_m, auto_class_s, auto_credits, auto_repair, auto_sound, auto_tuning, ballet, bank_cards, bank_deposit, beach_sports, beauty, boarding_schools, books, british_auto, bsnss_support, building_cars, burse, bus, business_train, cadet_corps, camera, car_insur, cats, celebration, cell_phone, cheap_auto, child_creativity_center, child_food, child_med, child_psy, child_sport_school, child_wares, child_wear, cinema, classical_concerts, classic_univ, clothes, clubs, combat_sports, commerc_serv, comm_realty_buy, comm_realty_rent, computer, concerts, consumer_credits, cookery, cosmetology, credits, culture_univ, dance, dance_sports, dating_sites, decorative_art, design, diet, diet_products, diving, dogs, doping, e_business, ecology_terrorism, economic_law_univ, ecoproducts, educational_center, elections, erotomania, ethnic, european_auto, everyday_wares, expensive_auto, extreme, extremism, fake_docs, family_kindergarten, fanatism, fastfood, federal_univ, fitness, food_delivery, foreign_college, foreign_realty, foreign_school, foreign_univ, forest_school, forex, furniture, gambling, games, garage, garden, german_auto, gifts, government, heavy_truck, hiking, hipsters, hobbies, homosexual, household_appliances, household_chemicals, house_rent, houses, housing, humane_univ, humorous_show, hunting, hypothec, insurance, intelligent_sports, isp, japanese_auto, job_law, job_search, job_support_orgs, kindergarten, korean_auto, land, lang_univ, laws, learn_gov, learn_lang, learn_non_gov, life_safety, light, light_duty_truck, local_authorities, low_alcohol_drinks, massage, media_period, media_themed, medical_univ, micro_credits, middle_cost_auto, military_univ, military_vocational, minibus, mlm, moto, movie_theater, museum, mushing, music, music_school, music_univ, music_vocational, nationalism, night_school, non_trad_med, non_trad_psy, npo, office_appliances, office_furniture, opposition, painting, parks, pedagogical_univ, photo_art, pif, plastic_surgery, playing_sports, plumbing_supplies, poetry, political_parties, politics, postgraduate, pregnancy, private_kindergarten, pro_government, pubs, quadricycle, real_buy, real_rent, realty, realty_development, refresher_course, religion, repair_wares, restaurant, retraining_course, road_motorcycle, rock_opera, russian_auto, sauna, school, scooter, sculpture, sea_rest, skiing, snowmobile, social_org, spares, special_vehicle, sport, sport_equipment, sport_motorcycle, sport_nutrition, sport_school, sport_univ, stationery, stomatology, summer_sports, swimming_pools, tabacco, technical_univ, textile, theatre, theologic_univ, ticket_fun, ticket_transp, tires_wheels, tourism, tourism_russia, trad_med, trad_psy, training_complex, travel, tutoring, very_expensive_auto, vocational, weapon, web_masters, wedding, wedding_agency, winter_sports, world_politics, yoga

Method

1. Create a text index of groups
2. Create a keyword list for each of 253 interests
3. **KW classifier**:
   - Retrieve top *k* groups retrieved by a set interest keywords
   - Rank by TF-IDF
   - Associate group's interests with its users
   - A group may have multiple interests
4. **ML classifier:**
   - Use top *k* groups as a training data
   - BOW features
   - Keyword features
   - Linear models: L2 LR, Liner SVM, NB
   - Classify all groups
   - A group may have up to three top interests
   - Associate group's interests with its users

Association of group's interests with its users

Engagement of a person into an interest category is proportional to the activity of the person in groups of this category:

$$e \approx w_{like} \cdot l + w_{s.comm} \cdot cs + w_{l.comm} \cdot cl + w_{repost} \cdot r$$

- l – the number of post likes
- cs – the number of short comments
- cl – the number of long comments
- r – the number of reposts

Association score of a user and an interest depends on engagement in a group and on the number of groups:

$$all \approx \alpha \cdot e_{fb} \cdot g_{fb} + \beta \cdot e_{vk} \cdot g_{vk}.$$

- e_{vk}, e_{fb} – engagement into VK/FB interest
- g_{vk}, g_{fb} – number of groups a user has in FB/VK

Results

Model	ML-groups1000-lr-30000	ML-groups3000-lr-30000	KW
Number of groups	2,913,212 (40,589,797)	3,952,806 (40,589,797)	6000 per category
Number of labels	3,008,354 (40,589,797)	4,090,816 (40,589,797)	1,022,813 (40,589,797)
Accuracy	0.91 +- 0.02	0.91 +- 0.03	--

Results per category: the best and the worst

	precision	recall	f1-score	support
agrarian_univ	1	0.9	0.95	117
cats	1	0.98	0.99	640
foreign_college	1	0.86	0.92	7
foreign_school	1	0.5	0.67	6
forest_school	1	0.42	0.59	26
job_law	1	0.18	0.3	17
lang_univ	1	0.17	0.29	6
sport_univ	1	0.71	0.83	17
training_complex	1	0.67	0.8	6
private_kindergarten	0.99	0.84	0.9	91
tabacco	0.99	0.99	0.99	924
air_sports	0.98	0.98	0.98	896
animals	0.98	0.98	0.98	900
beauty	0.98	0.99	0.98	926
dogs	0.98	0.99	0.99	904
erotomania	0.98	0.96	0.97	899
hipsters	0.98	0.94	0.96	751
boarding_schools	0.8	0.78	0.79	58
concerts	0.8	0.86	0.83	884
tourism	0.8	0.79	0.8	511
fastfood	0.79	0.86	0.82	892
media_period	0.79	0.74	0.76	792
ticket_fun	0.79	0.73	0.76	592
economic_law_univ	0.78	0.86	0.82	464
realty	0.77	0.67	0.72	399
technical_univ	0.77	0.63	0.69	265
educational_center	0.76	0.83	0.8	384
politics	0.76	0.66	0.7	453
npo	0.75	0.7	0.72	667
humane_univ	0.73	0.67	0.7	315
middle_cost_auto	0.73	0.64	0.68	25
music_univ	0.69	0.78	0.73	40
comm_realty_buy	0.68	0.59	0.63	313
cadet_corps	0.5	0.44	0.47	9

Top 30 interests on FB and VK

vk groups		fb publics		fb groups	
pregnancy	167100	books	15268	learn_lang	1611
games	153659	school	10654	media_themed	1229
school	109070	cinema	10076	photo_art	1170
music	94606	music	10018	dating_sites	1122
clothes	88252	media_themed	9567	clothes	1005
photo_art	72007	learn_lang	9162	tourism_russia	941
media_themed	70783	vocational	8321	design	937
poetry	63678	bsnss_support	6918	books	927
cats	62965	concerts	6067	hobbies	911
beauty	59363	religion	5340	wedding_agency	879
cinema	57298	advertising_online	4883	child_creativity_center	856
dogs	53818	poetry	4881	religion	836
summer_sports	52734	movie_theater	4827	gifts	753
clubs	48454	educational_center	4387	cookery	723
movie_theater	45892	british_auto	4330	celebration	718
painting	42096	games	4205	web_masters	706
wedding_agency	39808	summer_sports	4175	beauty	649
extreme	38567	fastfood	4013	games	648
gifts	37370	cookery	3853	music	643
cell_phone	35861	opposition	3844	cinema	598
books	35609	sport	3817	poetry	598
hiking	34223	child_creativity_center	3800	isp	568
parks	33730	dating_sites	3655	painting	566

Intersection of the top 30 interests on FB and VK

FB groups & FB publics & VK groups	VK groups & FB groups
1 games	1 wedding_agency
2 music	2 beauty
3 media_themed	3 cinema
4 cinema	4 gifts
	5 music
	6 games
	7 photo_art
	8 media_themed
	9 clothes

Interests co-occurrences

ML-groups3000-lr-30000		
cinema	movie_theater	3869
dance	music	2001
theatre	ticket_fun	1939
cell_phone	computer	1670
celebration	wedding	1597
concerts	music	1579
cosmetology	mlm	1367
clubs	concerts	1334
child_wear	clothes	1234
realty_development	repair_wares	1224
cinema	games	1224
car_insur	insurance	1121
parks	winter_sports	1108
extremism	nationalism	1050
computer	office_appliances	1015
camera	photo_art	986
ticket_fun	ticket_transp	979
low_alcohol_drinks	pubs	919
moto	road_motorcycle	814

Outline

Problem

Joint work with Dmitry Babaev and Segei Objedkov.

Motivation

- **input**: a user profile of one social network
- **output**: profile of the same person in another social network
- immediate applications in marketing, search, security, etc.

Contribution

- user identity resolution approach
- precision of 0.98 and recall of 0.54
- the method is computationally effective and easily parallelizable

Dataset

	VK	Facebook
Number of users in our dataset	89,561,085	2,903,144
Number of users in Russia [1]	100,000,000	13,000,000
User overlap	29%	88%

- **training set**: 92,488 matched FB-VK profiles

[1] According to comScore and `http://vk.com/about`

Profile matching algorithm

1. **Candidate generation**. For each VK profile we retrieve a set of FB profiles with similar first and second names.
2. **Candidate ranking**. The candidates are ranked according to similarity of their friends.
3. **Selection of the best candidate**. The goal of the final step is to select the best match from the list of candidates.

Candidate generation

- Retrieve FB users with names similar to the input VK profile.
- Two names are similar if the first letters are the same and the edit distance between names ≤ 2.
- Levenshtein Automata for fuzzy match between a VK user name and all FB user names
- Automatically extracted dictionary of name synonyms:
  - "Alexander", "Sasha", "Sanya", "Sanek", etc.

Candidate ranking

- The higher the number of friends with similar names in VK and FB profiles, the greater the similarity of these profiles.
- Two friends are considered to be similar if:
  - First two letters of their last names match
  - **Similarity between first/last names** sim_s are greater than thresholds α, β:

$$sim_s(s_i, s_j) = 1 - \frac{lev(s_i, s_j)}{\max(|s_i|, |s_j|)},$$

- Contribution of each friend to **similarity** sim_p **of two profiles** p_{vk} and p_{fb} is inverse of name expectation frequency:

$$sim_p(p_{vk}, p_{fb}) = \sum_{j: sim_s(s_i^f, s_j^f) > \alpha \wedge sim_s(s_i^s, s_j^s) > \beta} \min(1, \frac{N}{|s_j^f| \cdot |s_j^s|}).$$

Here s_i^f and s_i^s are first and second names of a VK profile, correspondingly, while s_j^f and s_j^s refer to a FB profile.

Best candidate selection

- FB candidates are ranked according to similarity sim_p to an input profile p_{vk}
- The best candidate p_{fb} should pass two thresholds to match:
  - its score should be higher than the *score threshold* γ:

$$sim_p(p_{vk}, p_{fb}) > \gamma.$$

  - either the only candidate or score ratio between it and the next best candidate p'_{fb} should be higher than the *ratio threshold* δ:

$$\frac{sim_p(p_{vk}, p_{fb})}{sim_p(p_{vk}, p'_{fb})} > \delta.$$

Results

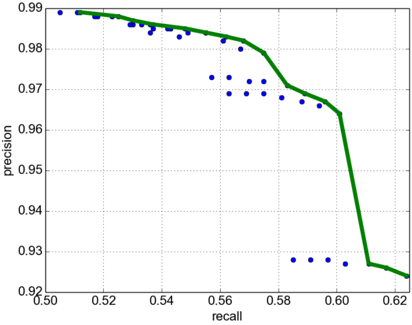


Figure: Precision-recall plot of the matching method. The bold line denotes the best precision at given recall.

Results: matching VK and FB profiles

First name threshold, α	0.8
Second name threshold, β	0.6
Profile score threshold, γ	3
Profile ratio threshold, δ	5
Number of matched profiles	644,334 (22%)
Expected precision	0.98
Expected recall	0.54

Outline

Statistics of the Facebook corpus 2013

Number of anonymized users	**3,190,813**
Language	Russian
Number of posts	426,089,762
Number of comments	147,140,265
Number of texts (posts + comments)	**573,230,027**
Number of tokens in posts	20,775,837,467
Number of tokens in comments	2,759,777,659
Number of tokens (posts + comments)	**23,535,615,126**
Average post length, tokens	49
Average comment length, tokens	19

Figure: Statistics of the Facebook corpus.

Most frequent positive and negative adjectives in the Facebook corpus

Positive adjectives		Negative adjectives	
хороший	good	плохой	bad
новый	new	старый	old
первый	first	долгий	long
нужный	helpful	неблагоприятный	unfavorable
бесплатный	free of charge	скучный	boring
любимый	beloved	сложный	complicated
интересный	interesting	голодный	hungry
спокойный	quiet	страшный	scary
социальный	social	скучно	bored
добрый	kind	немой	mute

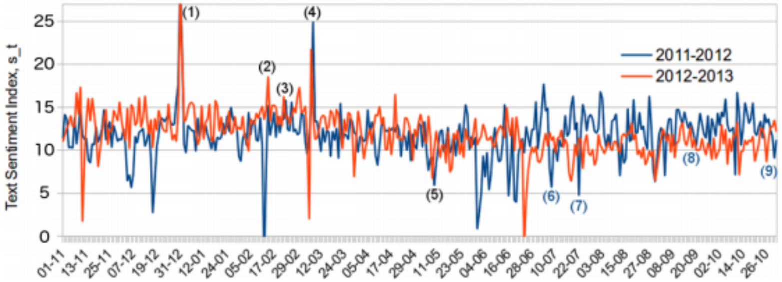
Performance of the dictionary-based sentiment classification approach, as compared to other methods (ROMIP-2012)

RunID	Object	Macro_P	Macro_R	Macro_F1	Accuracy	P_1	P_0	P_-1	R_1	R_0	R_-1
xxx	books	0.379	0.443	0.350	0.536	0.873	0.157	0.106	0.604	0.157	0.569
sentistrength	books	0.368	0.403	0.327	0.448	0.848	0.154	0.103	0.468	0.375	0.367
yyy	books	0.399	0.494	0.377	0.560	0.908	0.154	0.136	0.620	0.183	0.678
nb-blinov	books	0.408	**0.528**	**0.390**	**0.675**	0.909	0.157	0.157	0.785	0.042	**0.757**
dict ($\alpha = 0.02$)	books	0.42	0.431	0.348	0.445	0.893	**0.175**	0.191	0.42	0.677	0.197
dict ($\alpha = 0.05$)	books	0.437	0.404	0.274	0.327	0.919	0.163	0.229	0.246	0.844	0.122
dict ($\alpha = 0.07$)	books	**0.446**	0.381	0.217	0.261	**0.934**	0.157	0.248	0.155	**0.911**	0.077
Xxx	movies	0.395	0.454	0.361	0.493	0.819	0.235	0.131	0.586	0.148	0.628
Sentistrength	movies	0.371	0.401	0.343	0.436	0.774	0.219	0.119	0.485	0.274	0.445
Yyy	movies	0.411	**0.497**	0.390	0.522	0.849	0.221	0.165	0.610	0.173	**0.708**
nb-blinov	movies	0.347	0.489	**0.400**	**0.705**	0.788	0.000	0.253	**0.943**	0.000	0.525
dict ($\alpha = 0.02$)	movies	0.453	0.438	0.383	0.465	0.852	0.264	0.241	0.416	0.723	0.177
dict ($\alpha = 0.05$)	movies	0.473	0.412	0.315	0.382	0.892	**0.249**	0.278	0.259	0.869	0.109
dict ($\alpha = 0.07$)	movies	**0.48**	0.388	0.26	0.329	**0.908**	0.239	**0.293**	0.172	**0.923**	0.069
Xxx	cameras	0.388	0.390	0.370	0.561	0.864	0.111	0.190	0.632	0.138	0.400
Sentistrength	cameras	0.373	0.359	0.319	0.429	0.855	0.106	0.157	0.461	0.370	0.247
Yyy	cameras	0.445	**0.488**	**0.443**	0.628	0.903	**0.127**	0.303	0.687	0.312	**0.464**

Results of the social sentiment index calculation

	posts, %	comments, %	posts + comments, %
positive words	1.38	2.06	1.72
negative words	0.37	0.47	0.42
Word Emotion Index, e_w	1.75 (0.017)	2.53 (0.025)	2.14 (0.021)
Word Sentiment Index, s_w	3.72 (0.037)	4.38 (0.044)	3.81 (0.038)
positive texts	13.43	18.42	14.71
negative texts	1.83	2.28	1.94
Text Emotion Index, e_t	15.26 (0.153)	20.70	16.65 (0.166)
Text Sentiment Index, s_t	7.34 (0.073)	8.08	7.58 (0.076)

Results of the social sentiment index calculation



(1) 31-12-*—New Year (+);
(2) 14-02-*—St.Valentine's day (+);
(3) 23-02-*—Man's day (+);
(4) 08-03-*—Woman's day (+);
(5) 09-05-*—Victory Day, World War 2 commemorative day (–);
(6) 07-07-2012—Krasnodar Krai floods in Russia[6] (–);
(7) 22-07-2012—A new unpopular law regulating non-profit organizations in Russia[7] (–);
(8) 16-09-2012—A mass protest against government in Russia[8] (–);
(9) 25-10-2012—Hurricane Sandy in US (–).

Outline

Much more fun stuff can be done with the FB/VK data

- **Neologism Detection**
  - Build frequency dictionary of posts and comments.
  - Filter out dictionary words and grammatical errors.
  - Interpret most frequent non-dictionary words linguistically.
- **User Age & Region Detection**
  - Tell me who are your friends, and I will say who you are.
  - Most frequent age/region of friends.
  - Reject users with high variation of age/region among friends.
  - Up to 85-90% of accuracy.
- **User Income Detection**
  - Transfer learning: target variable is not present in SNs.
  - Training a model on a set of users with known income.
  - Applying the model on the social network profiles.

Thank you! Questions?

Al Zamal, F., Liu, W., and Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Goswami, S., Sarkar, S., and Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In *Third International AAAI Conference on Weblogs and Social Media*.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002).

Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Liu, W., Zamal, F. A., and Ruths, D. (2012). Using social media to infer gender composition of commuter populations. *Proceedings of the When the City Meets the Citizen Worksop*.

Mukherjee, A. and Liu, B. (2010). Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217. Association for Computational Linguistics.

Peersman, C., Daelemans, W., and Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Rangel, F. and Rosso, P. (2013). Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, page 177.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.