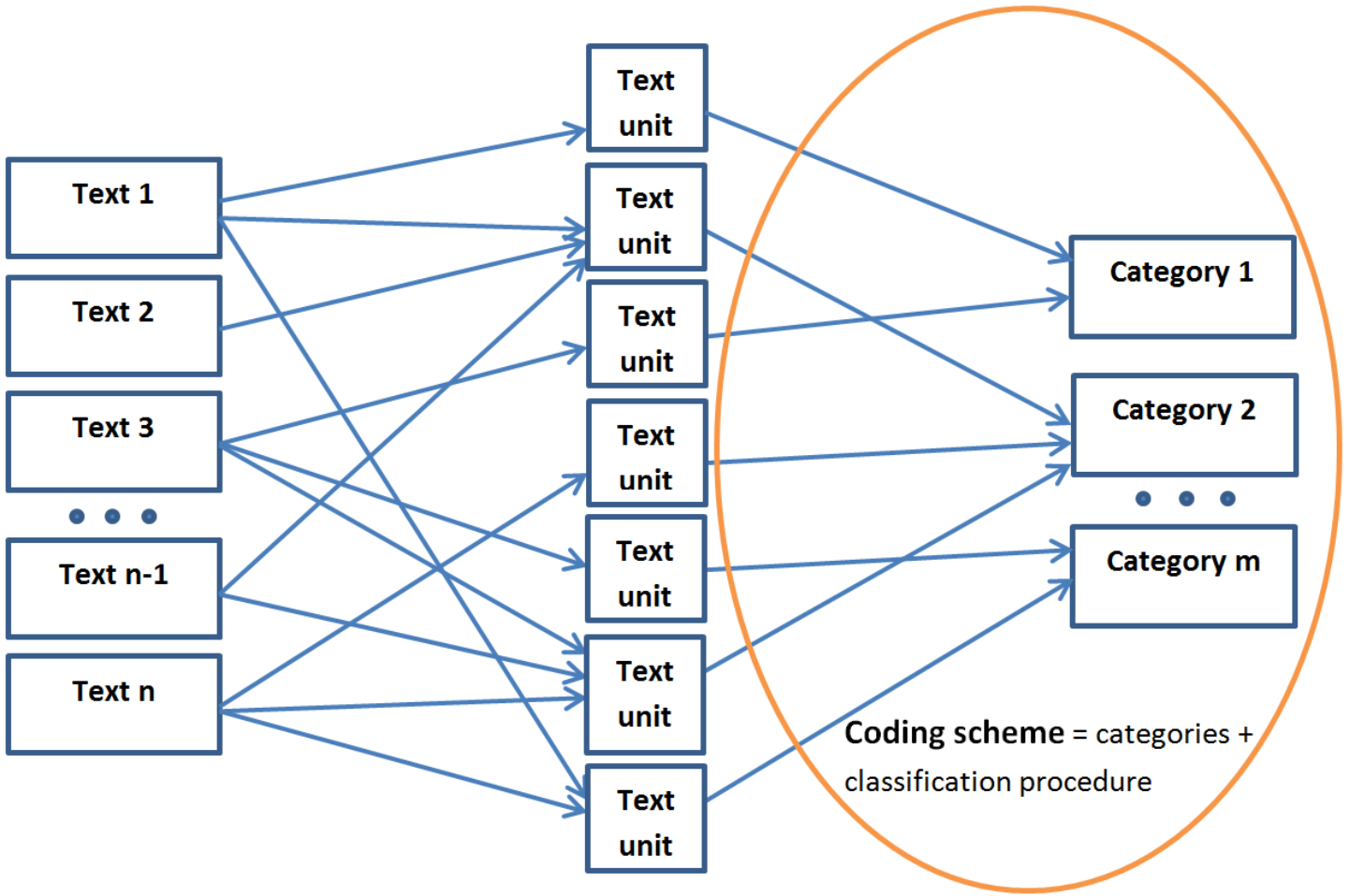
A Low Effort Approach to Quantitative Content Analysis

Maria Saburova Archil Maysuradze

Lomonosov Moscow State University

01.10.2015

Traditional content analysis



Traditional content analysis

Traditional workflow of quantitative content analysis:

- Define categories ("deductive" approach)
- Define the basic text units to be classified (e.g. individual words, phrases, or paragraphs)
- Develop a code guide (classification procedure)
- Apply coding scheme to a text corpus (classification)
- Quantification

Coding scheme abstractions

1. Question list

Coding scheme properties:

- Requires the understanding of implicit topics
- Cannot be answered automatically
- Requires to teach people to assess text units coherently
- Many assessors should be included

2. Value dictionary

Properties:

- Coding using dictionary can be automated
- Standard large dictionary or demand the dictionary from users
- The main complexity is to obtain the domain-oriented dictionary.

Statements

Research goal

Provide individual researchers with a low effort content analysis workflow.

Our research aims

To automate the domain-specific dictionary creation by means of machine learning.

To perform it

We had to recognize and solve nonstandard type of machine learning problem — feature distribution among content categories.

Interviewing workflow

A typical workflow of an individual researcher includes **interviewing** of a small number of respondents.

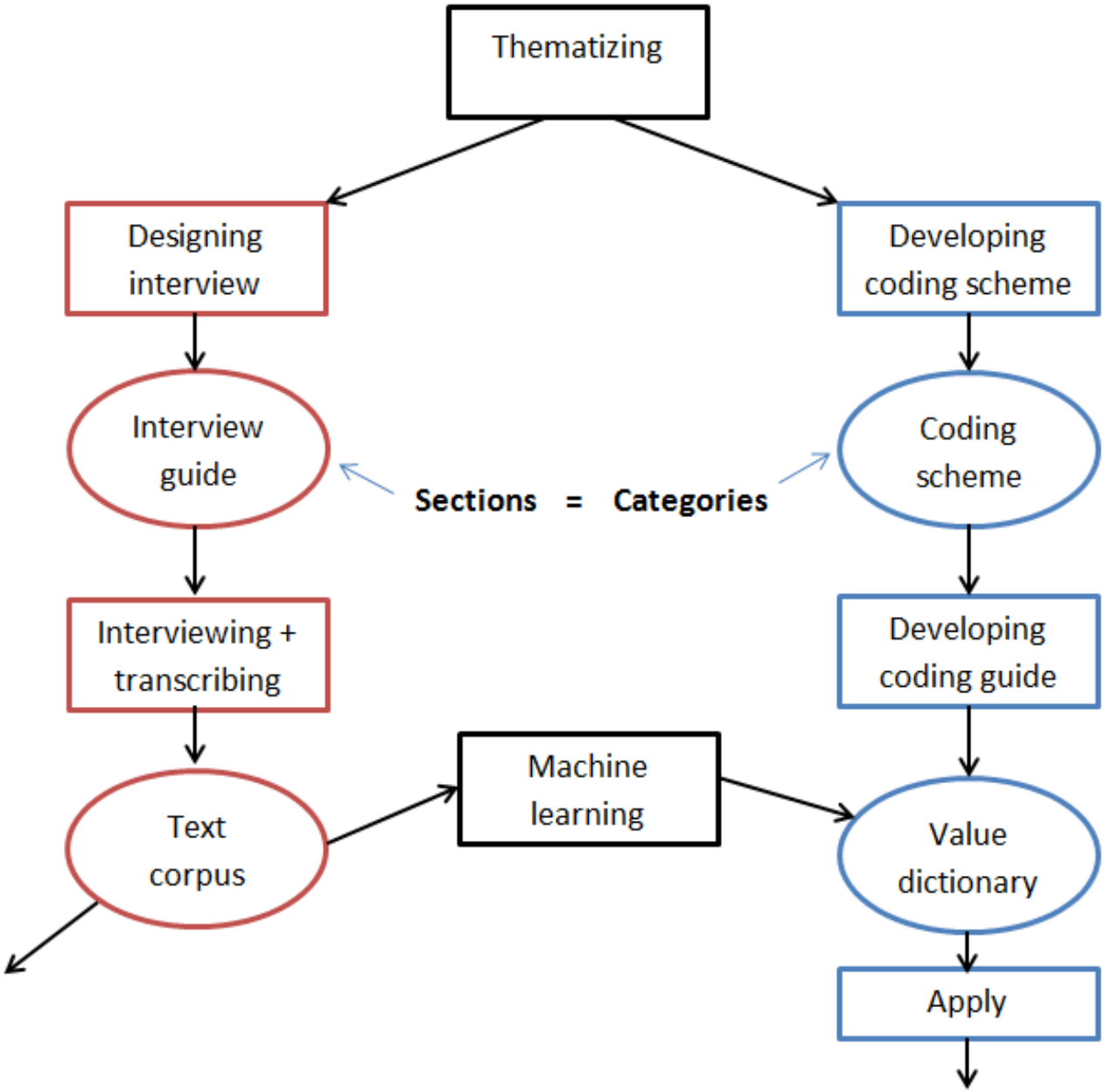
A complete interviewing process includes the following steps:

- Thematizing
- Designing
- Interviewing
- Transcribing
- Analyzing and Reporting

We claim

The data collected during the interview design and the interviewing may be used to develop a domain-specific dictionary.

Compare two workflows



Summary

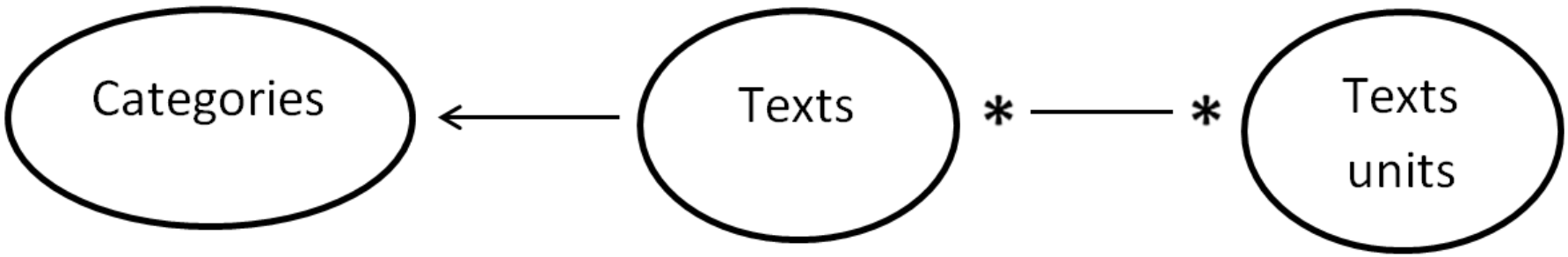
We propose:

- A workflow when an individual researcher proceeds to automatic quantitative content analysis after interviewing and transcribing steps
- A method of automatic construction of a domain-specific dictionary that only uses data collected during the interviewing

Need to formalize:

- Data model
- Formal problem
- Solution method

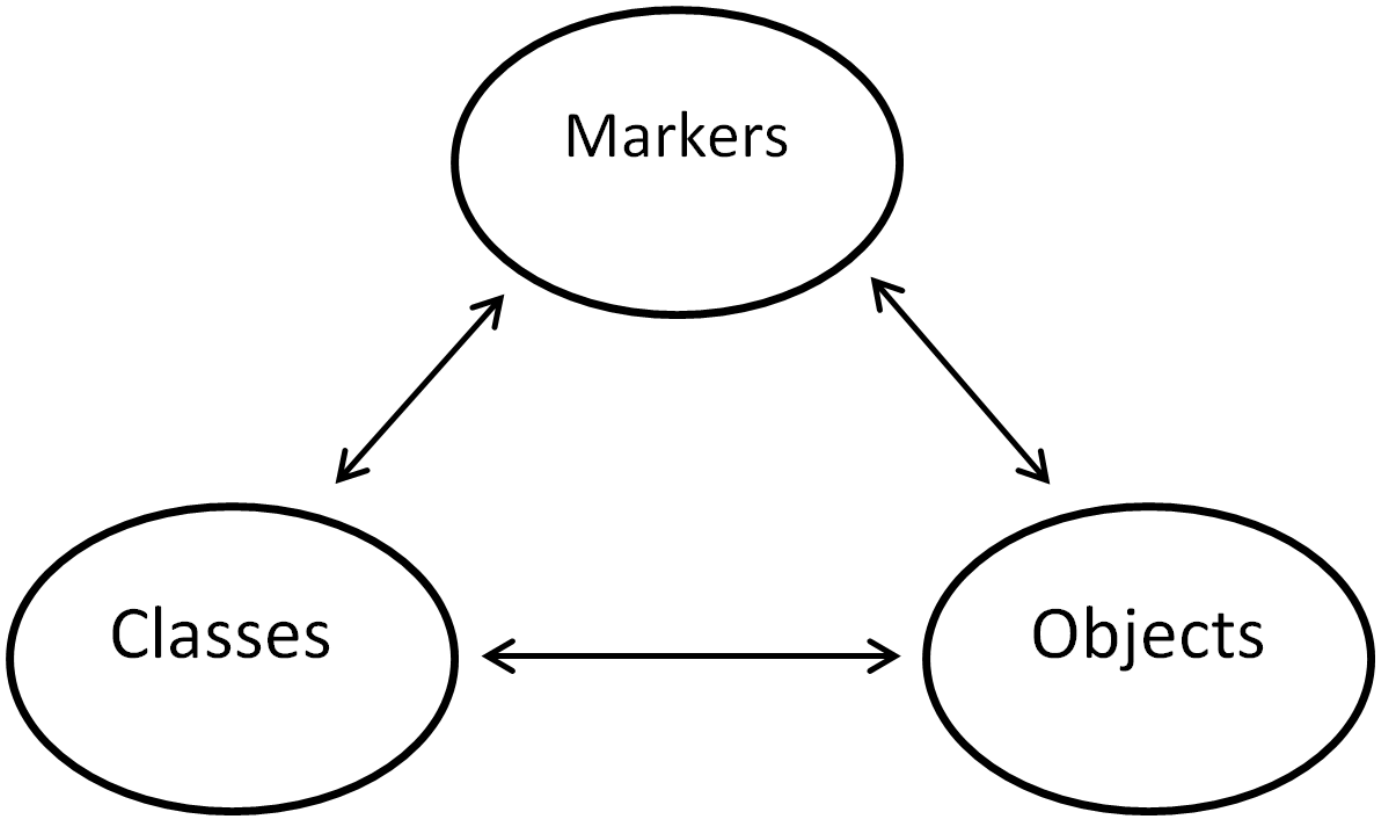
Content analysis workflow



Data model

Tripartite data model:

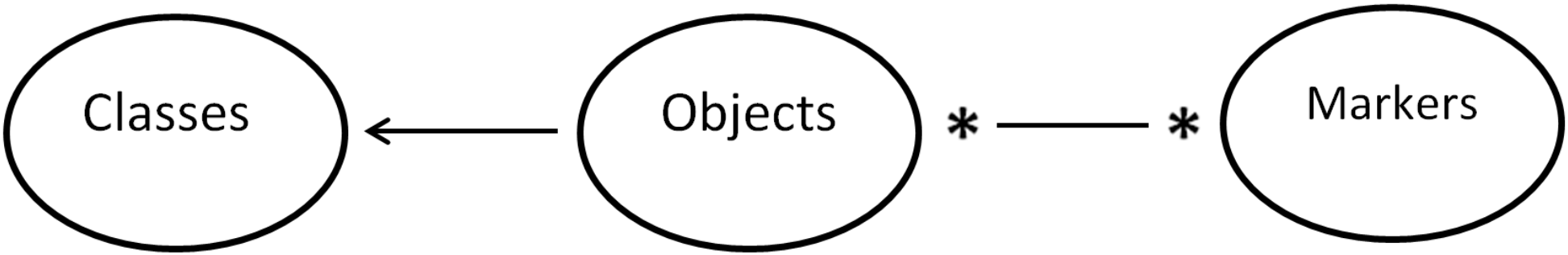
- three components: objects, markers and classes
- three binary heterogeneous relations between the units of analysis



Data model

Tripartite data model:

- three components: objects, markers and classes
- three binary heterogeneous relations between the units of analysis



Problem statement

The problem of feature distribution by classes consists of the following:

- Given predefined classes
- Given a training set
- Each object has a feature description and a class label
- Requires every feature assigned uniquely to either one of the predefined classes

Note: there is no initial marking of features relating to classes.

We consider

The problem of creating a dictionary as the problem of distributing the features.

Problem formalization

We introduce the following notation:

1. T — the number of features, t — feature number from 1 till T,
2. I — the number of labeled objects, i — object number from 1 till I,
3. J — the number of classes, j — class number from 1 till J,
4. f_{it} — value of feature t for object i, in particular, 0 or 1 for binary relation object-feature,
5. c_i — real object label i.

6. Required function: a_t — class number, which is mapped with feature t.

Feature distribution properties

- The results obtain their own interpretation and value.
- We are required to make a decision on each feature.
- Each feature should be labelled uniquely.
- The ratio of the number of features to the number of precedents is much greater than in traditional classification problems.
- Marker is interpreted as the presence of some object properties.
- In our problem statement features have a Boolean type or can be naturally reduced to this type.

Methods of collecting dictionary

Function as a classifier parameter

The markup is given for objects rather than features. Therefore, we may start with reducing the problem of feature classification to the task of object classification.

Required partial function from attributes to the classes

Will be obtained automatically after fitting classifier of this model to the data.

Function as a classifier parameter

We use linear information model as one of the simplest.

The estimate of object belonging to the class k:

$$\Gamma_k = \sum_{t=1}^{T} w_t f_t[a_t = k], \quad \text{where } w_t \text{ is non-negative feature's weight}$$

Decision rule:

$$A = \text{argmax}_k \Gamma_k.$$

Multiclass SVM analogue method

Parameters $\{a_t\}$ and $\{w_t\}$ are configured by solving margin maximization problem:

$$\frac{1}{2}||w||^2 + C\sum_{i,j}\xi_{ij} \to \min_{w,\{\xi_{ij}\}_{i,j}} \tag{1}$$

$$\sum_t w_t f_{it}([a_t = c_i] - [a_t = j]) \geq 1 - \xi_{ij},\ \forall i,\ \forall j \neq c_i \tag{2}$$

$$w_t \geq 0,\ \forall t \tag{3}$$

$$\xi_{ij} \geq 0,\ \forall i,\ \forall j \neq c_i \tag{4}$$

Multiclass SVM analogue method

Particularly, when sample is linearly separable, problem can be simplified:

$$\begin{aligned} &\tfrac{1}{2}||w||^2 \to \min \\ &\textstyle\sum_t w_t f_{it}([a_t = c_i] - [a_t = j]) \geq 1,\ \forall i,\ \forall j \neq c_i \\ &w_t \geq 0,\ \forall t \end{aligned} \tag{5}$$

To solve the problem, dual problem is defined:

$$\begin{aligned} &\textstyle\sum_{i,j} \alpha_{ij} - \tfrac{1}{2}||\beta_t + X_{ijt}^T \alpha_{ij}||^2 \to \max \\ &0 \leq \alpha_{i,j} \leq C, \\ &\beta_t \geq 0,\ \forall t \end{aligned} \tag{6}$$

where α_{ij} and β_t are dual variables, X_{ijt} is defined as

$$X_{ijt} = f_{it}([a_t = c_i] - [a_t = j]). \tag{7}$$

Multiclass SVM analogue method

After dual problem solving, we return to initial features:

$$\begin{aligned} \beta_t &= 0, \quad \text{if } X_{ijt}^T \alpha_{ij} \geq 0 \\ \beta_t &= -X_{ijt}^T \alpha_{ij}, \quad \text{otherwise.} \end{aligned} \tag{8}$$

Weights can be found with formula:

$$w_t = \beta_t + X_{ijt}^T \alpha_{ij}. \tag{9}$$

The dual problem is a linear programming problem, if a_i are fixed. Interior-point method can be used to solve this problem.

Multiclass SVM analogue method

Coordinate descent method is used to train a_i:

1. The algorithm starts from initial point: every feature is assigned to the class in which it often occurs.
2. On each iteration random feature s is selected. For this feature look over all classes for which this feature vote.
3. Weights w_t are optimized for each of these classes, when a_t are held.
4. Class a_s with a maximum value of the dual problem functional is assigned to the feature.
5. The procedure is repeated until convergence or until the specified number of iterations will be reached.

One-vs.-one SVM in relation to multiclass problem

We consider SVM algorithm for binary classification problem:

$$\begin{aligned} &\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i \to \min_{w,\xi};\\ &y_i\langle w, x_i\rangle \geq 1 - \xi_i, \quad \forall i = 1,\ldots,\ell;\\ &\xi_i \geq 0, \quad \forall i = 1,\ldots,\ell. \end{aligned} \tag{10}$$

The problem of multiclass classification can be reduced to the set of binary classification problems. We use one-vs.-one scheme:

1. We train binary classifiers a_{sk} for all classes pairs $s \neq k$;
2. Each of them distinguishes documents of class s from documents of class k;
3. Weights w_t^{sk} is considering to each classifier;
4. If $w_t^{sk} > 0$, then feature t vote for class s, else k;
5. Feature t is assigned to class s, if $w_t^{sk} > 0$ for more than a half pairs $k \neq s$.

Experiments

Data description:

- 20 interviews with the leaders of volunteering organizations.
- 6 categories: 'Supervisor portrait', 'Objectives and content of the organization's activities', 'The Concept of volunteerism', 'Working with volunteers', 'Volunteers portrait', 'Incentives and barriers to volunteering activities'.
- 7241 normalized words
- 120 documents

Multiclass SVM analogue results

Supervisor portrait			Objectives and content of the organization's activities			The Concept of volunteerism		
special'nost'	specialty	100%	finansirovanie	financing	100%	nazyvaju	call	100%
uchus'	learn	100%	itog	summary	100%	nazvat'	call	100%
skol'ko	how much	67%	reshenie	solution	100%	razovyj	one	100%
nemnogo	a little	50%	budushhee	future	100%	ponjatie	the concept	100%
sozdanie	creation	60%	istochnik	source	100%	obshhestvennik	public man	100%
gde	where	67%	voznikaju	arise	83%	bezvozmezdnyj	free	100%
ozhidanie	waiting	71%	zasluga	merit	86%	aktivist	activist	100%
universitet	University	75%	poslednij	last	75%	inogda	sometimes	100%
okonchanie	the end	78%	naibolee	the most	67%	znachimyj	significant	100%
god	year	70%	reshit'	to solve	70%	sistematicheskij	systematic	100%
reshil	decided	73%	cel'	goal	73%	dobrovol'chestvo	volunteering	100%
davno	long	75%	trudnost'	the difficulty	75%	social'no	social	100%

Working with volunteers			Volunteers portrait			Incentives and barriers to volunteering activities		
shtatnyj	staffing	100%	chashhe	more often	0%	meshaju	disturb	100%
dovolen	happy	100%	muzhchina	man	50%	otnoshus'	am	50%
navyk	skill	100%	zhenshhina	woman	67%	municipal'nyj	municipal	67%
special'nyj	special	100%	stanovljus'	become	75%	gosudarstvennyj	state	75%
proishozhu	happen	80%	molodoj	young	80%	prestizhen	prestigious	80%
internet	Internet	83%	dumaju	think	67%	naselenie	population	83%
pishu	write	86%	dobryj	good	71%	bol'shinstvo	most	86%
vazhno	important	88%	starshe	older	75%	modno	fashionable	88%
obojtis'	do	89%	sluchaj	case	67%	struktura	structure	89%
obraz	the way	90%	edinyj	single	60%	strana	country	90%
lichnyj	personal	91%	portret	portrait	64%	kazhetsja	it seems	82%
privlekat'	to attract	92%	procent	percentage	67%	doverie	trust	83%

‘One-v.s.-one SVM’ method

Supervisor portrait			Objectives and content of the organization’s activities			The Concept of volunteerism		
special’nost’	specialty	100%	itog	summary	100%	aktivist	activist	100%
uchit’sja	to learn	100%	reshenie	solution	100%	nazyvat’	call	100%
skol’ko	how much	100%	finansirovanie	financing	100%	razovyj	one	100%
gde	where	100%	istochnik	source	100%	obshhestvennik	public man	100%
rasskazat’	to tell	100%	budushhee	future	100%	inogda	sometimes	100%
nemnogo	a little	83%	naibolee	the most	83%	social’no	social	100%
posle	after	71%	cel’	goal	86%	znachimyj	significant	100%
sozdanie	creation	75%	trudnost’	the difficulty	88%	ponjatie	the concept	100%
ozhidanie	waiting	78%	zasluga	merit	89%	vozmezdnyj	reimbursable	100%
zanjat’sja	to do	80%	postavit’	to put	90%	jepizodicheskij	episodic	100%
okonchanie	the end	82%	poslednij	last	82%	schitat’	take	100%
universitet	University	83%	sposob	method	83%	sistematicheskij	systematic	100%
Working with volunteers			**Volunteers portrait**			**Incentives and barriers to volunteering activities**		
shtatnyj	staffing	100%	chastyj	frequent	0%	meshat’	disturb	100%
navyk	skill	100%	muzhchina	man	50%	otnosit’sja	apply	50%
dovol’nyj	happy	100%	zhenshhina	woman	67%	dobrozhelatel’nyj	friendly	67%
special’nyj	special	100%	portret	portrait	75%	bol’shinstvo	most	75%
internet	Internet	100%	molodoj	young	80%	municipal’nyj	municipal	80%
obojtis’	do	100%	duh	the spirit	67%	gosudarstvennyj	state	83%
jetap	stage	100%	edinyj	single	57%	prestizhnyj	prestigious	86%
kontrolirovat’	control	100%	stanovit’sja	to become	63%	naselenie	population	88%
zatrata	cost	100%	blagopoluchnyj	safe	67%	doverie	trust	89%
stimulirovat’	to stimulate	100%	aktivnyj	active	70%	modno	fashionable	90%
privlekat’	to attract	100%	cennost’	value	73%	razvitie	development	91%
dorogoj	dear	92%	starshij	senior	75%	strana	country	92%

Conclusion

- We proposed and implemented a low effort sociological workflow
- Our technique makes it possible to generate the dictionary from interview data.
- The problem of dictionary development was formalized as the problem of feature distribution.
- We proposed two solution methods, both were implemented and tested on real data.
- Future work will shift the focus to multiword phrases.

Thank you!

Your questions?

Contacts:
Maria Saburova, MSU
saburova.mi@yandex.ru

Archil Maysuradze, MSU
maysuradze@cs.msu.su