Andrey Kutuzov
National Research University Higher School of Economics
Olga Miryasova
Institute of Sociology, Russian Academy of Science

Social unrest through the prism of language: computational linguistics at sociology service

Институт социологии
Российской академии наук
INSTITUTE OF SOCIOLOGY OF RUSSIAN ACADEMY SCIENCE

The Why Linguistics Conference
May 7-9 2015, Tartu, Estonia

What's all this about

- Recently, sociologists started turning from **singular interviews** to analysis of "**big data**" collected from natural people interaction.
- Often this data comes as **raw text**: forum posts, e-mail messages, chat communication, etc.
- **Computational linguistics** and **natural language processing** provide methods to extract meaning and structure from raw texts.

Outline

1) What we study: sociological material

2) Why linguistics: massive text processing

3) Questions answered

4) Questions unresolved

2012 kindergarten catering case

Grass-roots social movement:

- Issue: **bad catering in Moscow kindergartens**;
- Participants: Parents from Moscow;
- Mostly mothers, age 22 to 45;
- Communicated via Internet forum http://forum.materinstvo.ru;
- 807 active users (at least one forum post) at the peak of activity in 2012;
- Activists organized rallies, met with the officials and addressed protest letters to the authorities.

Rally in February 2012

Transformation from philistines to activists

(1) Problems transform from **private** to **common** ones.

(2) Individual problem-solving transforms into **collective action**.

(3) Moving from "I am an object of political action" attitude towards "I am a subject of political action" attitude.

(4) Growing feeling of **solidarity** and **empowerment**.

Sociologists face problems

(1) **Traditional interviews** are limited with regards to number of correspondents and are somehow “artificial” (people act and speak differently in comparison with their natural habits).

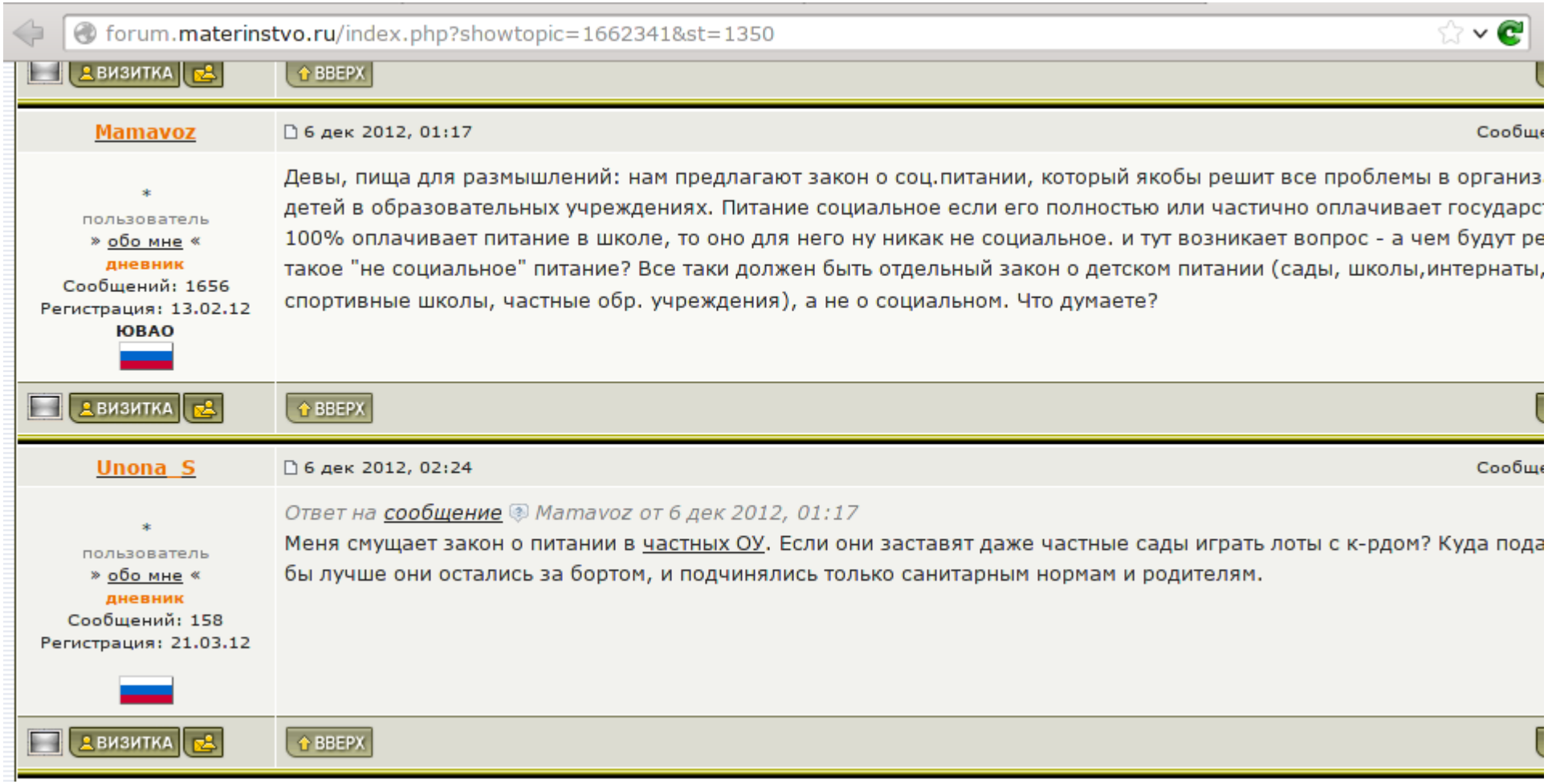
(2) **Participant observation** provides good personal understanding of the movement, but is difficult to formalize and generalize.

(3) All these approaches are highly **subjective** and have issues with **reproducibility**.

(4) Hence the need to automatically process and extract meaningful data from forum posts: **unstructured natural language**.

Source of linguistic data

http://forum.materinstvo.ru/ - large chain of threads about the kindergarten catering campaign



Sociologists' aims

(1) To compare **relations to power and authorities** within different groups inside the movement: how these sub-groups perceive their problems

(2) To support expert estimation of sub-groups division with **statistical evidence based on language use** by sub-groups members.

(3) To reveal **qualitative and quantitative differences** between the subgroups.

(4) To observe **opinion and behaviour evolution** over time.

Outline

1) What we study: sociological material
2) **Why linguistics: massive text processing**
3) Questions answered
4) Questions unresolved

What linguistics has to offer: data-driven analysis

(1) Using raw **word frequencies** to find **keywords** in one sub-corpus against another sub-corpus (with log-likelihood or other similar measures).
Software: AntConc

(2) Using word frequencies to **estimate and visualise changes** in activists' attitude over time.
Software: Python, Matplotlib

(3) Using **bag-of-words vector representations of texts** to automatically cluster posts or users into groups. Machine learning allows to do it in a completely unsupervised way.
Software: Weka, Orange, scikit-learn, etc.

Structuring the data

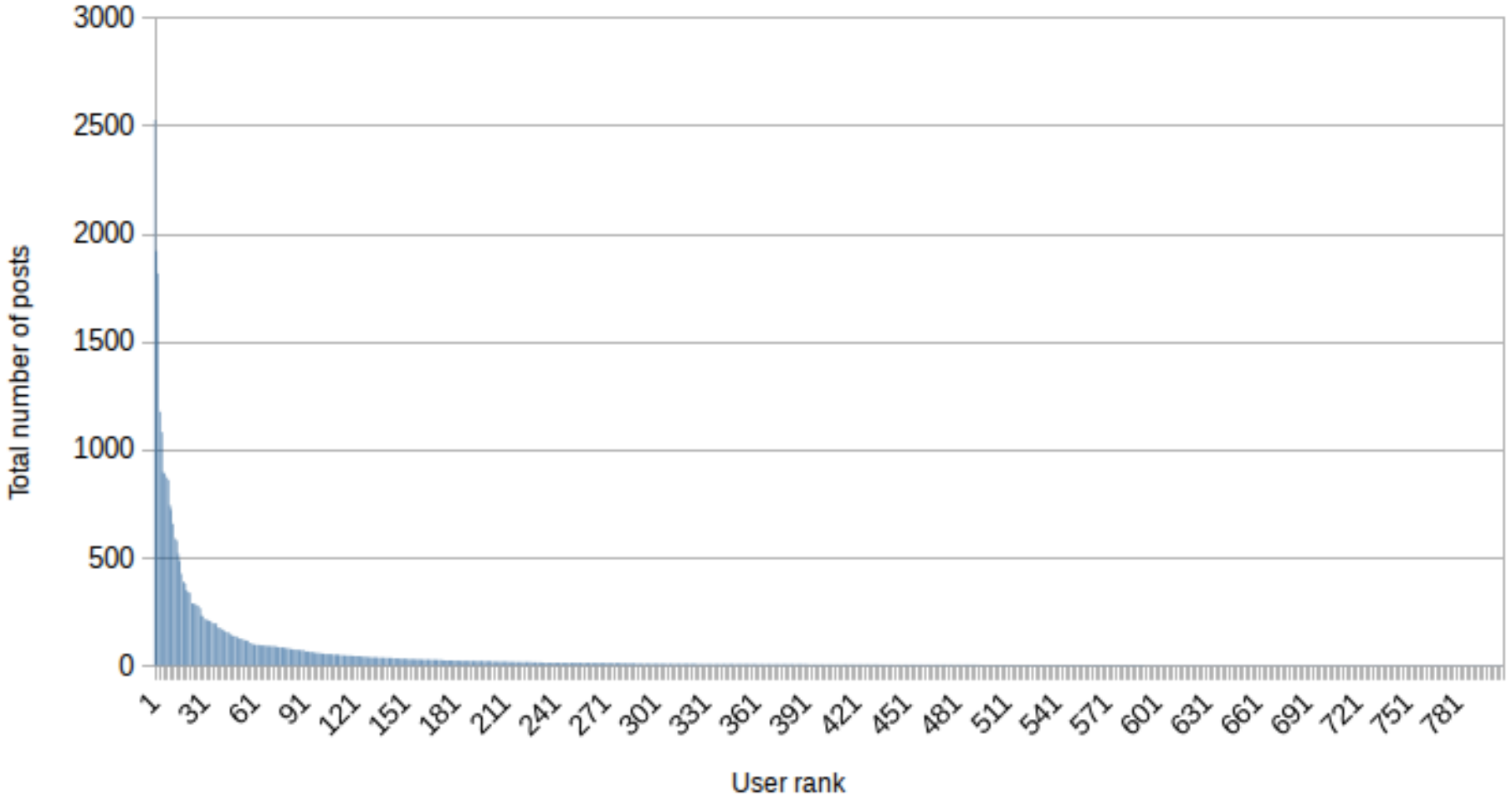
- Forum pages are **crawled and downloaded**.
- Data fields we are interested in are extracted using **regular expressions**; texts are **lemmatized**, functional words removed.
- Now we have **structured text corpus**: 34 134 posts from 807 users, 954 thousand words.
- **Reference corpus** used to control experiments (random threads from the same forum, not related to kindergarten catering): 805 thousand words.
- Data example:

Thread ID	Post ID	Date	Post text	User name	Total posts by user	Sign up date	Location
1657534	1105	2012-10-23	Пишут, что замораживают	LuMix	14286	23.11.2007	Moscow

Sub-corpora

- Structured data allows to analyse arbitrary “slices” of the whole corpus, for example:
  - Only posts by “activists” (“activists” selected by a sociologist);
  - Only posts by those who signed up in February 2012;
  - Only posts in August 2012;
  - ...etc.
- Huge amount of natural utterances, without any interviewer's influence.

Active members of the community can be traced by sheer number of posts: the distribution closely follows power law



“Political” keywords manifesting difference between “kindergarten” thread and other threads (as produced by log-likelihood measure)

Frequency	Log-Likelihood	Word
130	122	elections
188	190	member of parliament
206	167	president
161	118	government
191	122	authorities
89	84	protest
71	80	political
126	105	to unite
68	85	bureaucrat
304	130	to struggle
176	95	citizen
148	134	major
111	121	political party

Keywords manifesting difference between sub-groups (as produced by log-likelihood measure)

Active participants:

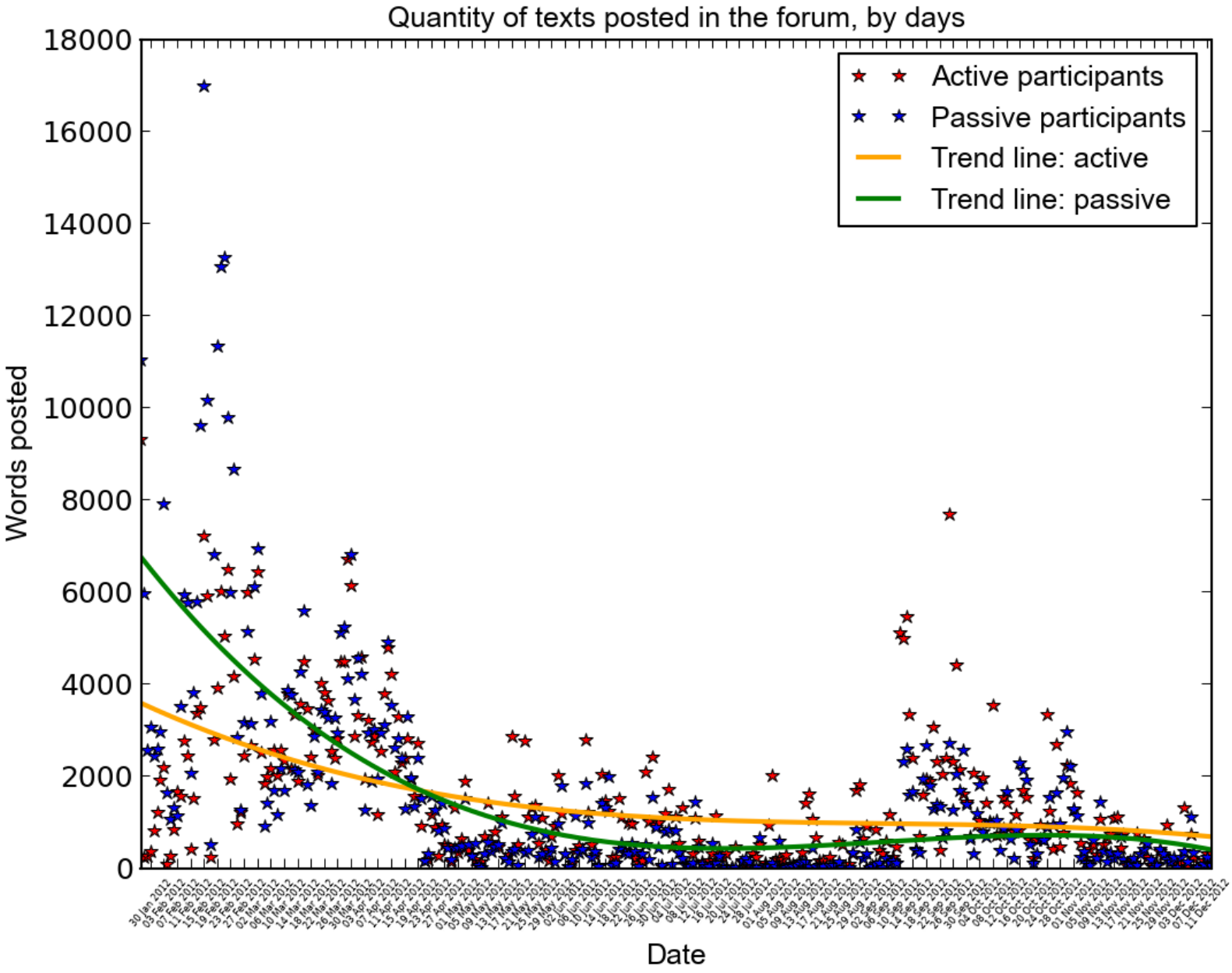
Department, organization, district, service, administrator, supplier, etc.

Passive participants:

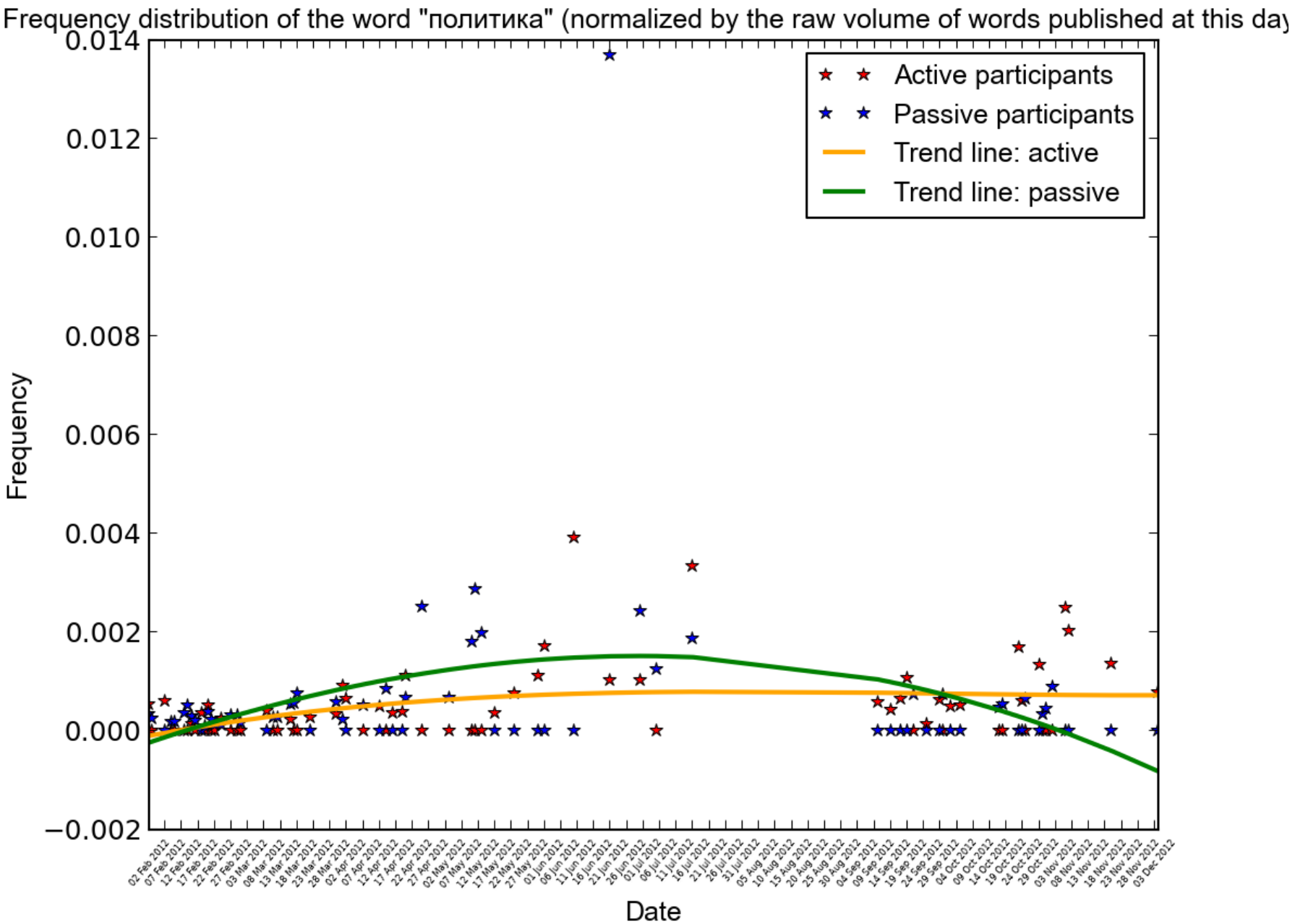
Kindergarten, allergen, nurse, child, drink, feed, menu, etc.

Often words are used in diminutive forms.

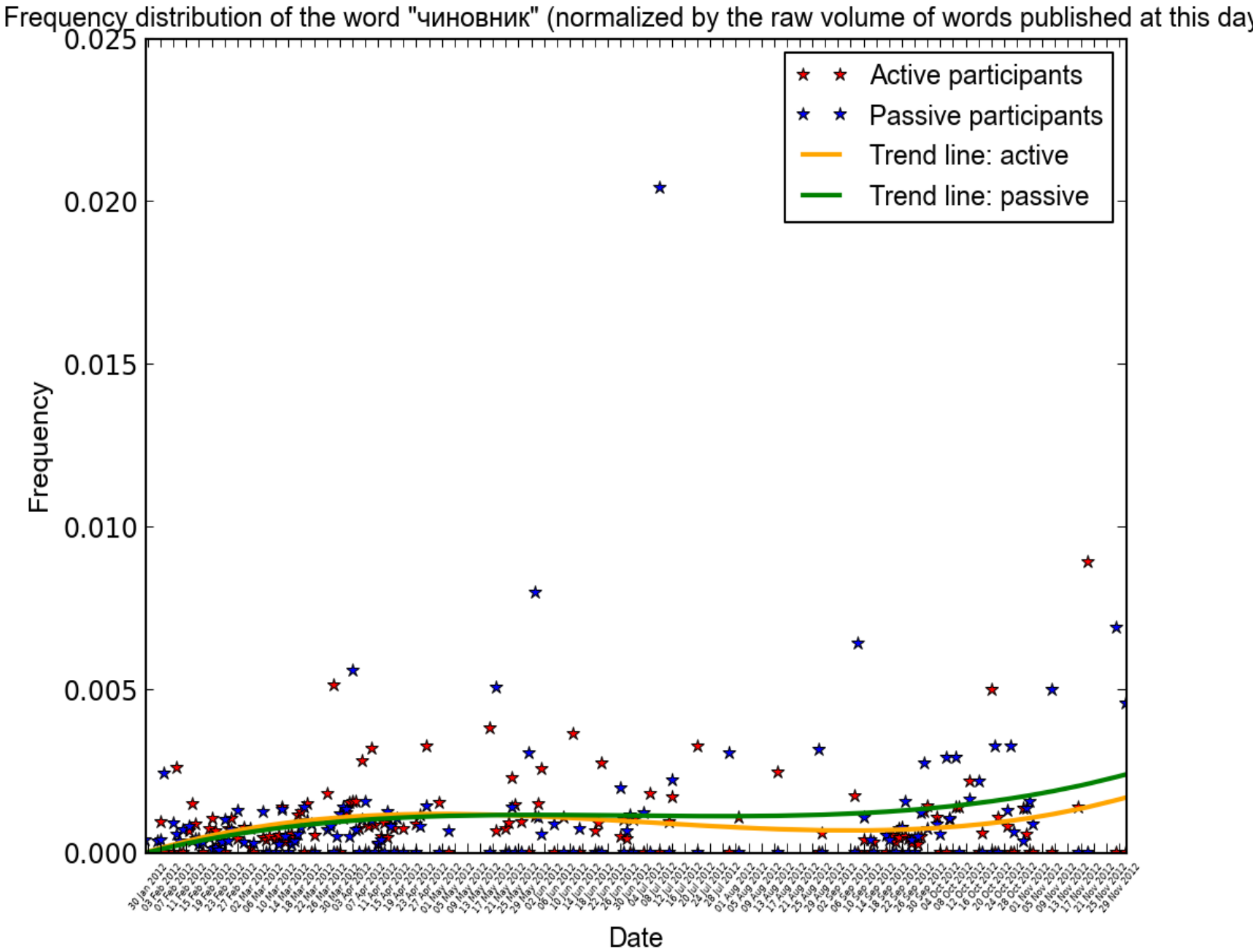
Structured data allows building “time lines” of language usage



How the word “**politics**” is used through 2012



How the word “**clerk/bureaucrat**” is used through 2012



Frequency distribution interpretation

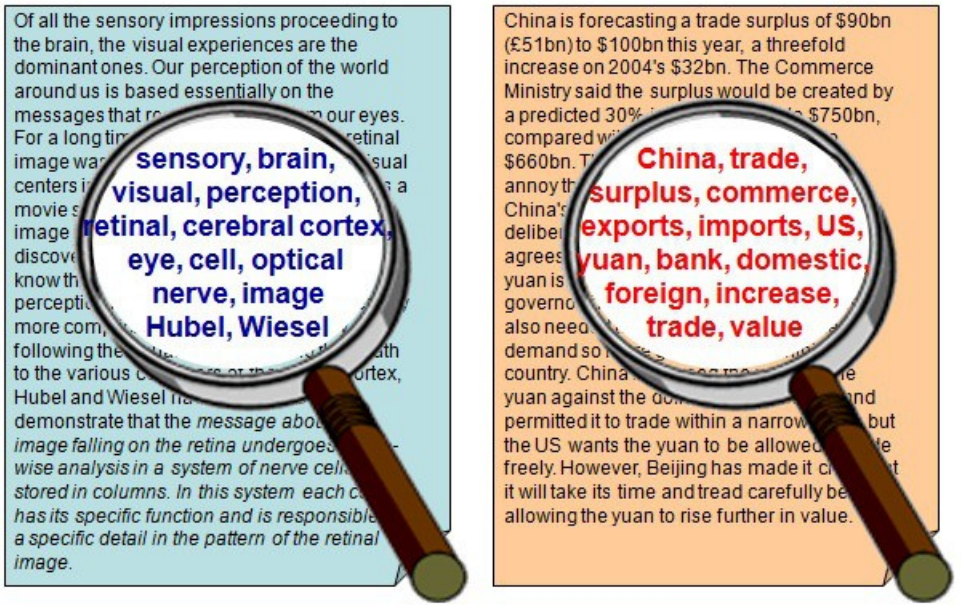
- **25 “active participants”** produced more posts than other 700 participants together. During 2012, post frequency dropped for all users, but it is less so for active ones.
- **“Active participants”** more frequently use words related to **interaction with authorities**; other participantsprimarily use lexicon related to **child care, health and kindergartens.**
- In comparison with reference corpus, our thread is more “politics-oriented”: words like “**rally**” or “**picketing**” are significantly more frequent.

Text as a vector of word counts

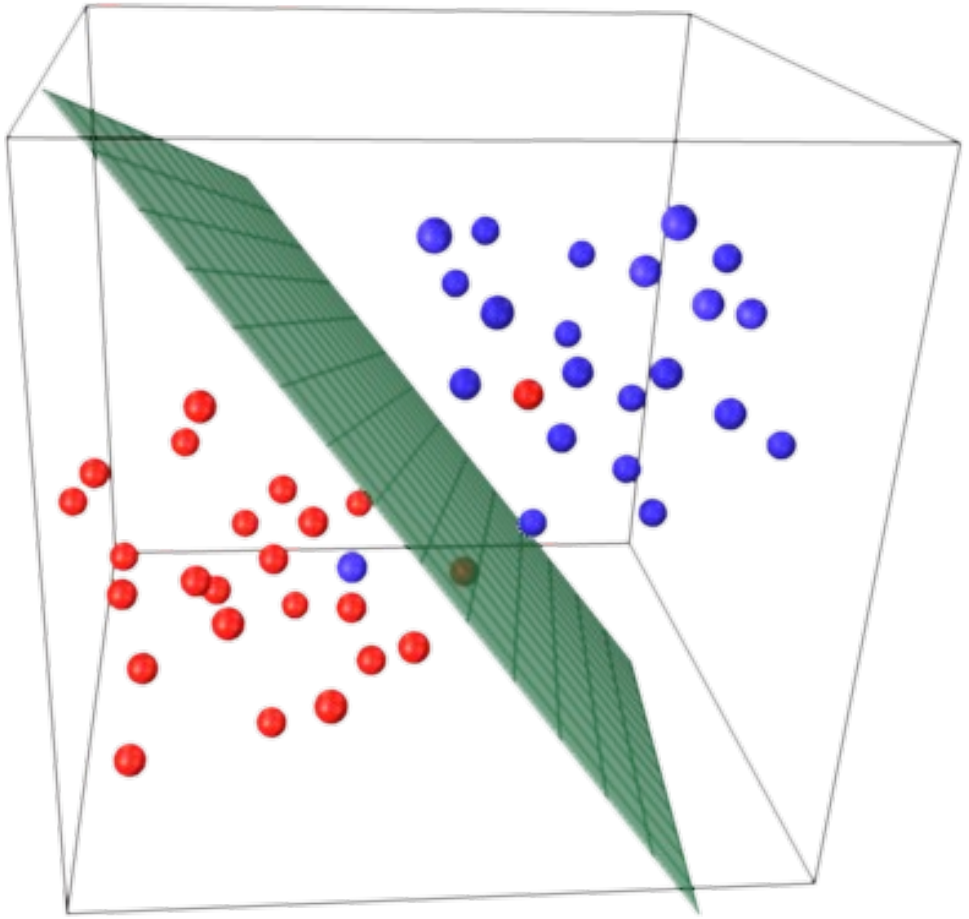
To compare texts in a formal and reproducible way, one has to convert them into numbers and project them in a **vector space**.

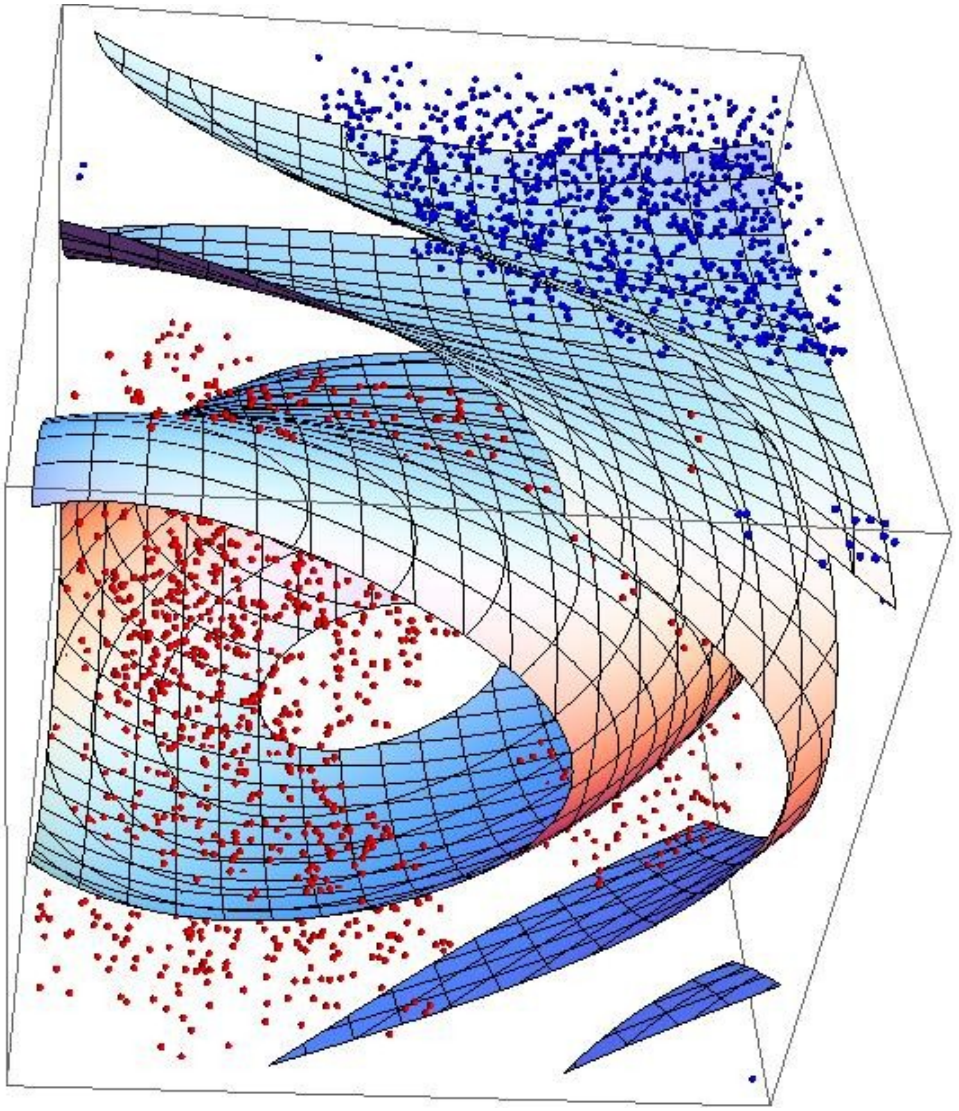
The simplest model is **bag-of-words**, where a text is a point in the space of dimensionality *n*, where *n* is the size of our lexicon, and each coordinate of this point is the n^{th} word frequency in the document.

Once documents are vectors, we can train **classifiers** on them or **cluster** them with any suitable algorithms.



The aim of **automatic classification** is actually to find a hyperplane which can correctly separate documents in the lexical vector space





Classification results

- **Posts from “protest” threads** about kindergarten catering are almost linearly separated from **random other posts** by **SVM** algorithm.

  Precision: 0.98
  Recall: 0.81
  kappa: 0.8535

- It is enough to initiate vector space with **Top-500** frequent words

- Inside these threads, **posts by “active participants”** are less different from **other participants' posts. Rotation Forest** algorithm gave the best results:

- **Precision: 0.8**
  Recall: 0.73
  kappa: 0.7556

- We had to use not less than **Top-1000** frequent words

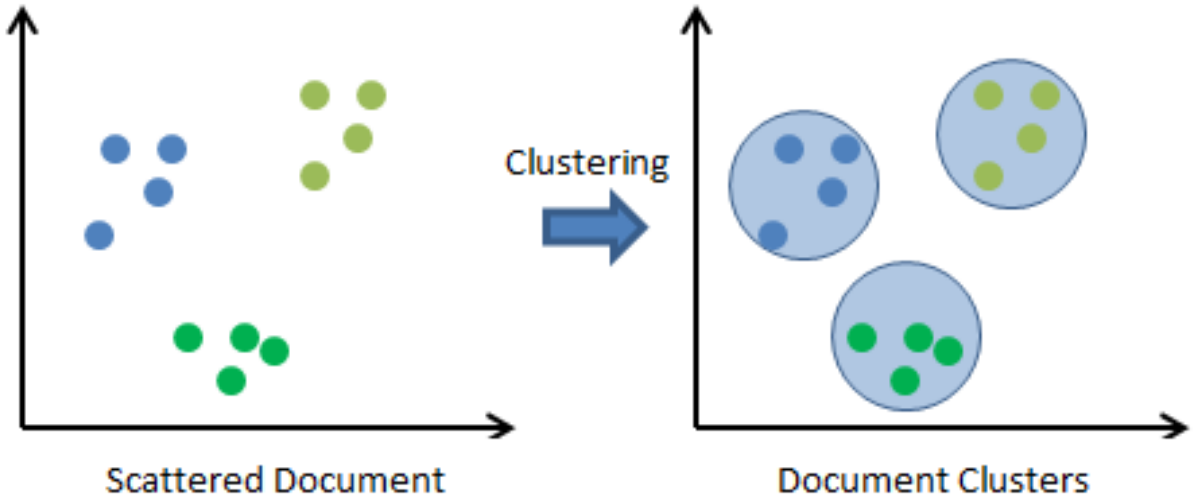
Automatic classification refining expert decisions

6 of 25 “active participants” selected by an expert sociologist, were consistently marked by Rotation Forest classifier as **“passive”** (based on their typical vocabulary).

After reconsidering, the expert **agreed** that these users should be excluded from “active participants”.

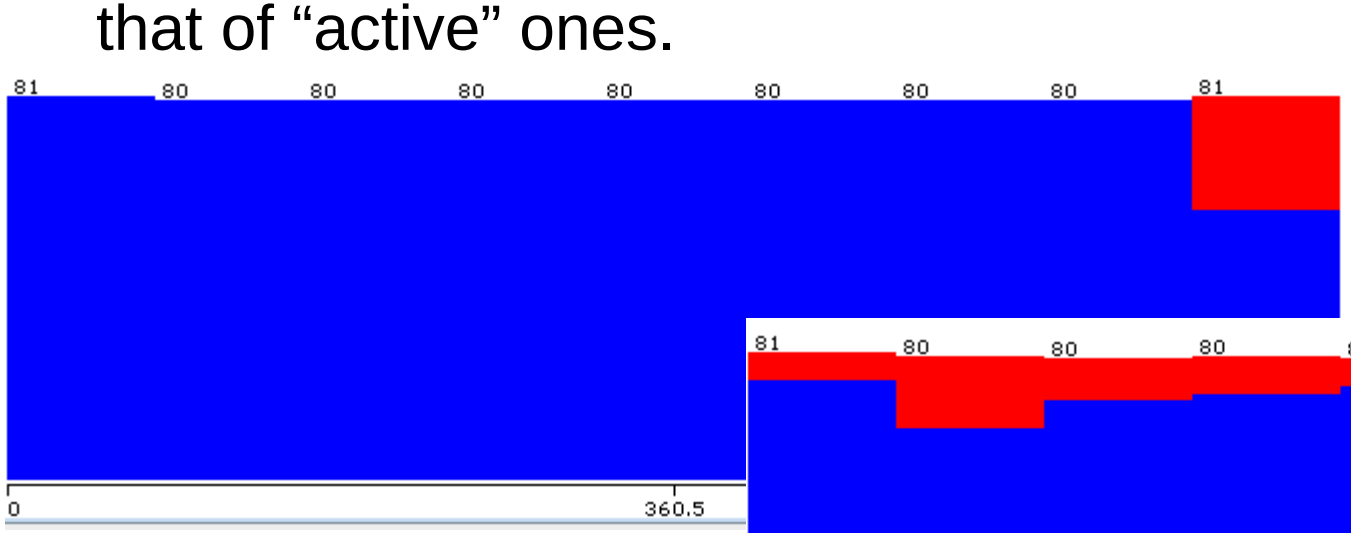
Text clustering

- Clustering is basically classification without pre-determined classes (unsupervised).
- Community participants can be clustered into groups based on what words they typically use.



Clustering results

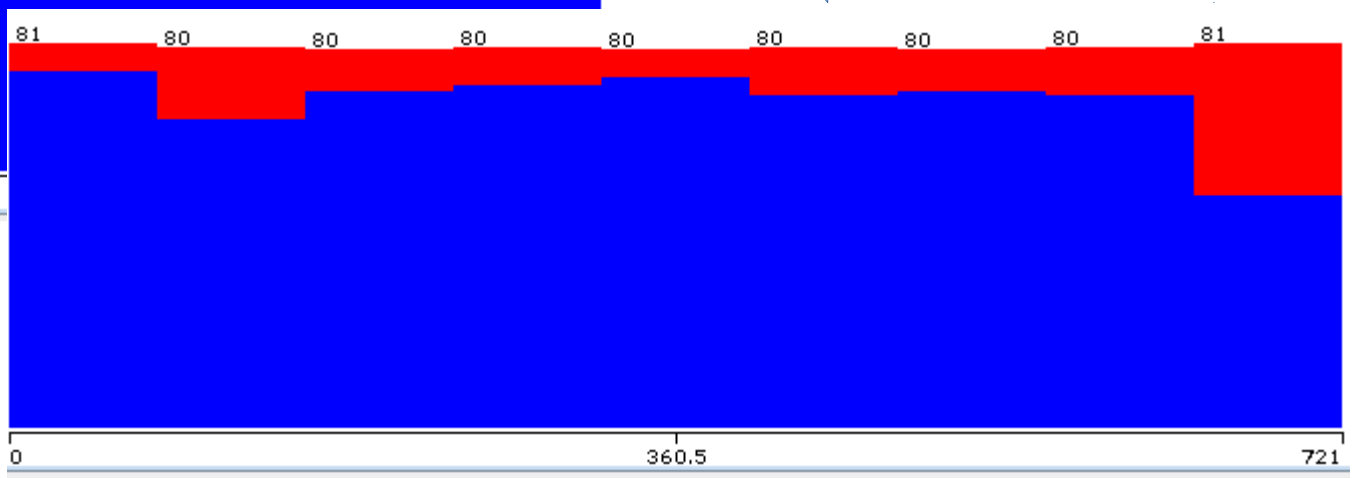
- There are no clusters among “active participants” selected by an expert.
- However, if clustering all “protest thread” participants, there appear **cluster 1** (218 users) including all “active participants” and **cluster 2** (588 users) including only “passive” users.
- It means there is a large group of users, whom experts had not selected as “active”, but in fact their typical vocabulary is similar to that of “active” ones.



Expert estimation

- active

Unsupervised clustering



Outline

1) What we study: sociological material
2) Why linguistics: massive text processing
3) **Questions answered**
4) Questions unresolved

Questions answered

- **Expert estimation** of sub-groups in the activists' community was **supported with lexical statistical data**.
- **Distribution of lexical frequencies** do show differences between sub-groups (more and less active users).
- The **vocabulary** of the social movement core witnesses that these people **conceive problems as common and question the authorities' right to make decisions without consulting the public.**

Questions answered

- “Active participants” more frequently use **names of officials or titles of governmental bodies.** However, we observe higher usage of words like “**supplier**” or “**administrator**” among “passive participants” since October 2012.
- It means they also switched to **perceiving problems as common**, but on a more **local level** of particular kindergartens.
- General trend throughout the year: less mentioning for the authorities (Moscow Education Department), more mentioning for administrators of local kindergartens.
- Low amount of specifically “political” words. Forum users are **still on their way from philistines to activists**.

Outline

1) What we study: sociological material
2) Why linguistics: massive text processing
3) Questions answered
4) **Questions unresolved**

Questions unresolved

- How to process **ellipsis, irony, euphemisms and co-reference**? All these phenomena abound in our material ("*they*" to refer to the authorities, etc.).
- Natural language processing itself may be unsupervised; but still we need an **informed expert** to qualitatively **interpret** the results and to evaluate the data.

Questions unresolved

- It would be useful to extract **discussion chains** (initial posts, replies and replies to replies) for further expert analysis.
- **Read-only** forum users are “invisible” to this approach (unlike interviews).
- To be confident about “**philistines to activists”** transformation, one should take into account not only linguistic features, but also facts of peoples' biographies.

Andrey Kutuzov
National Research University Higher School of Economics
Olga Miryasova
Institute of Sociology, Russian Academy of Science

Thanks for attention!

Questions are welcome.