MOSCOW ANALYTICAL CENTER

Machine learning methods and artificial intelligence technologies in the semantic analysis of questionnaire observation results in official statistics

Elena Zarova,
Doctor of Economics, Professor

Analytical Center by Moscow City Government

MOSCOW ANALYTICAL CENTER

The concept of semantic analysis

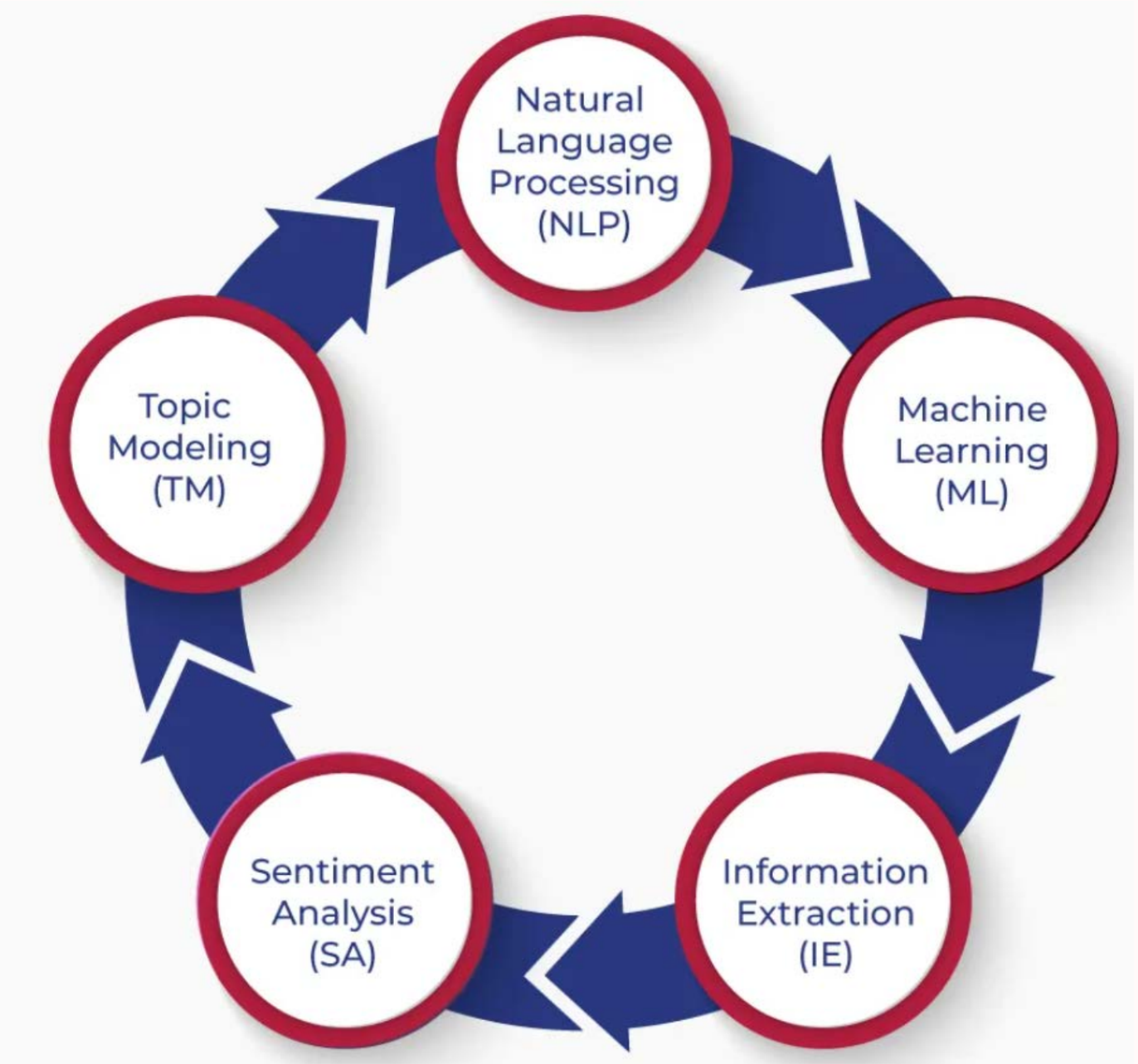
Origin of the word

- ***Semantics – from the Greek word σημαντικός, meaning "signifying" or "indicating***

The essence of the method

- ***Semantic analysis*** *is the process of extracting* ***explicit and implicit meaning from text****.*
- *Unlike traditional methods of text analysis, which focus on reading and* ***systematizing texts according to certain features****, semantic analysis seeks to understand the deeper meanings of texts and their collections (corpora) by analyzing syntactic structures, psycho-lexical connections, context, and other factors influencing meaning*

Statistical and Machine Learning Methods in Intelligent Analysis of Text Data (Text Mining) – Functions



ФЕДЕРАЛЬНОЕ СТАТИСТИЧЕСКОЕ НАБЛЮДЕНИЕ

КОНФИДЕНЦИАЛЬНОСТЬ ГАРАНТИРУЕТСЯ ПОЛУЧАТЕЛЕМ ИНФОРМАЦИИ

ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ ИСПОЛЬЗОВАНИЯ СУТОЧНОГО ФОНДА ВРЕМЕНИ НАСЕЛЕНИЕМ

ДНЕВНИК ИСПОЛЬЗОВАНИЯ ВРЕМЕНИ

ДЛЯ ЛИЦ В ВОЗРАСТЕ 15 ЛЕТ И БОЛЕЕ

ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ

Представляют:	Сроки представления
интервьюеры выборочного наблюдения использования суточного фонда времени населением - территориальному органу Росстата по установленному им адресу	до 3 октября 2024 г.

Форма № 3 – бюджет времени
Приказ Росстата об утверждении формы от 26.06.2024 № 254 О внесении изменений (при наличии) от № от №
1 раз в 5 лет

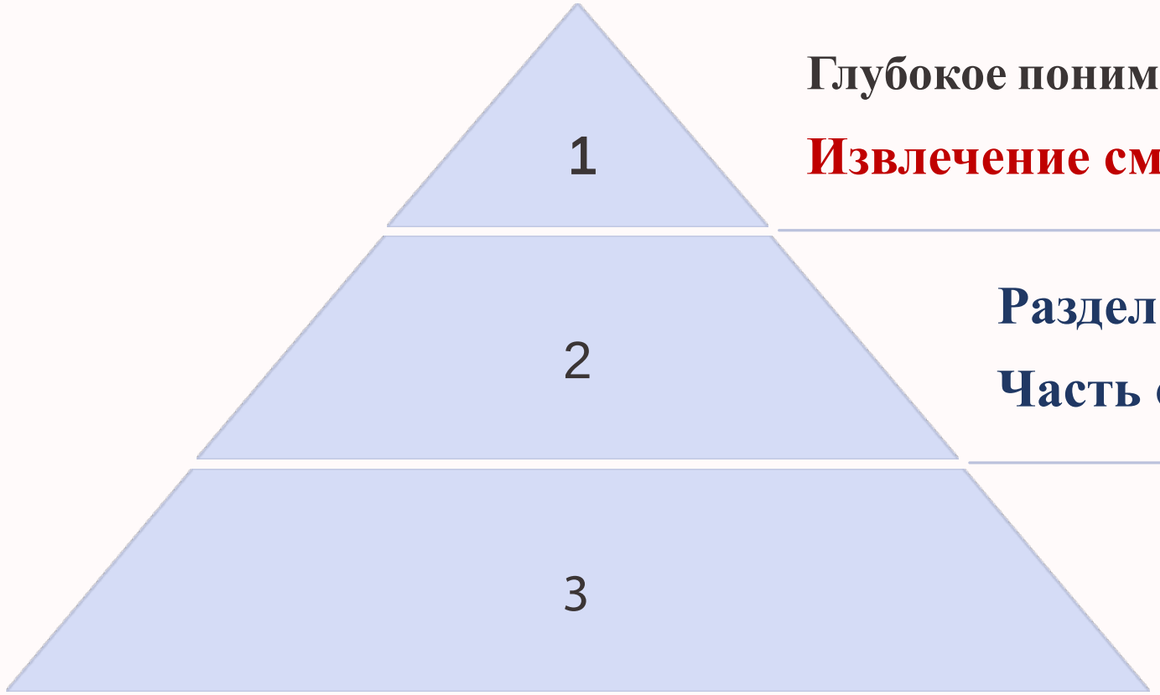
Территория _____

Время	***Что Вы делали?*** *Запишите свое основное занятие в каждый 10-минутный интервал с 16.00 до 19.00*		***Что еще Вы делали?*** *Запишите самое важное параллельное занятие*		***Использование сети Интернет***	***Где вы были?*** *Запишите расположение или вид транспорта например дома, дома у друзей, в школе, на рабочем месте, в магазине, пешком, в автомобиле, автобусе*		***Вы были один (одна) или с кем-то из знакомых Вам людей?*** *Отметьте «Да» крестиком*			
	Вписывайте только одно основное занятие в строке. Разделяйте собственно передвижение от деятельности, являющейся его причиной. Не забывайте указать вид транспорта. Отделяйте основную работу от дополнительной.	КОД вида основной деятельности		КОД вида параллельной деятельности	Отметьте «Да» крестиком		КОД места нахождения	Одна/(один)	С детьми 0-9 лет, живущими в Вашем домохозяйстве	С другими членами Вашего домохозяйства	С другими знакомыми людьми
16.00-16.10	*Ждала автобус на остановке*					*На улице*		x			
16.10-16.20	*Ехала на автобусе за детьми в детский сад*		*Читала новости на смартфоне*		**X**	*В автобусе*		x			
16.20-16.30	*Разговаривала с воспитательницей*		*Помогала детям одеваться*			*В детском саду*			x		x
16.30-16.40	*Пошла пешком в магазин*		*Разговаривала с детьми*			*пешком*			x		
16.40-16.50	*Покупала продукты*					*В магазине*			x		
16.50-**17.00**	*Шла пешком домой*					*пешком*			x		
17.00-17.10	*Пришла домой, переодевалась*		*Помогала детям переодеться*			*дома*			x		x
17.10-17.20	*Убирала свои продукты в холодильник*					- " -		x			
17.20-17.30	*Готовила ужин*		*Смотрела видео на планшете*		**X**	- " -		x			
17.30-17.40	- " -		- " -		**X**	- " -		x			
17.40-17.50	*Ужинала*		*Разговаривала с семьей*			- " -			x	x	

Example – diary-based time registration form. This is a structured form of self-registration where some variation in description is allowed, but the main information must be entered according to established rules.

SEMI-STRUCTURED FORM

Семантический анализ для обработки результатов анкетного опроса



Глубокое понимание

Извлечение смысла и контекста из текстов

Раздел NLP

Часть обработки естественного языка

Основа интерпретации

Помогает машинам понимать человеческий язык

Семантический анализ позволяет компьютерам извлекать значение из слов. Это критически важно для работы с естественным языком, который может использоваться при заполнении респондентами опросников в свободной или отчасти свободной форме.

Языковая неоднозначность при обследовании рабочего времени составляет от 8 до 15 %

Языковая неоднозначность возникает, когда респонденты формулируют свои ответы в свободной форме, и эти ответы могут быть интерпретированы по-разному в зависимости от контекста, личного опыта, культурных особенностей и других факторов.

В результате часть ответов оказывается "размытыми", двусмысленными или слишком общими для автоматической категоризации.

Примеры:

- **Ответ: "Работаю дома" — может означать как удалённую работу, так и работу по дому (домашние обязанности).**

- **Ответ: "Занимался делами" — не раскрывает конкретный вид деятельности.**

Examples of frequency analysis of time-use diary entries – WORD CLOUDS

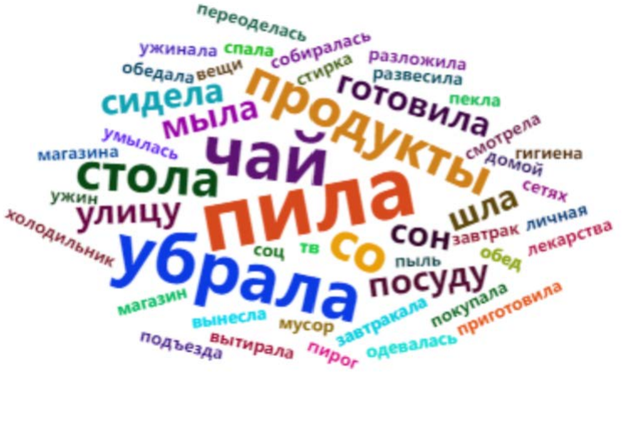
Respondent №1

Respondent № 2

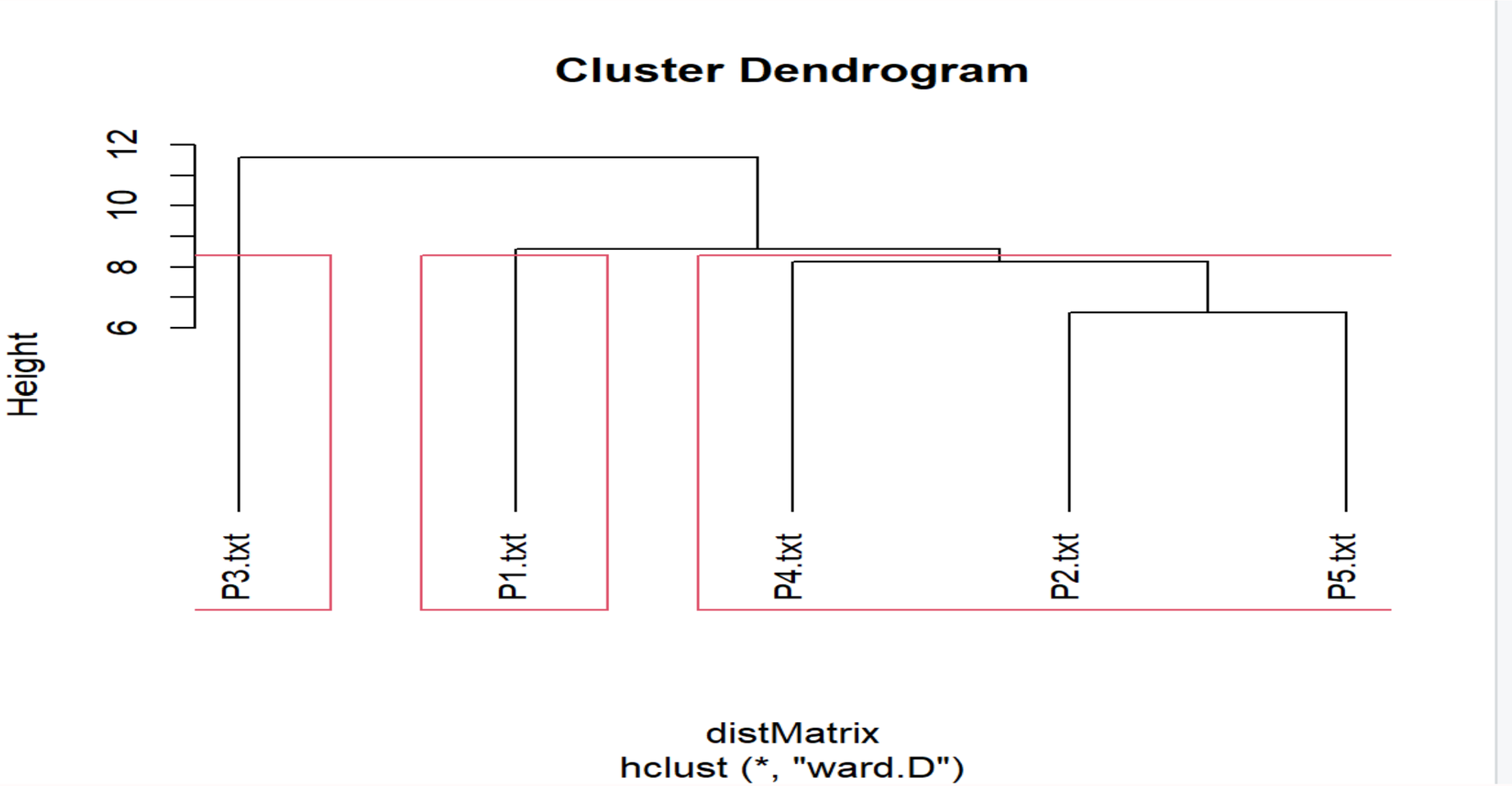
Respondent №3

Respondent №4

Respondent №5



Clustering of time-use diaries by content



Respondents 2 and 5 gave very similar answers (the interviewer should be checked).

Assessment of associations reveals redundancy or insufficiency of words in "typical responses."

Respondent №1

Ассоциации к слову «работал»

```
$работал
    сарае хозяйству      поле
     0.55      0.55      0.33
```

Respondent № 2

Ассоциации к слову «готовила»

```
$готовила
завтрак    обед    ужин
    0.56    0.56    0.56
```

- The results obtained are a "trial run," but they have confirmed the necessity of applying text mining to the processing of survey data. This approach significantly increases the efficiency of extracting information from textual data and improves its quality. In the future, neural networks could be used to process large volumes of textual data—both unstructured and semi-structured—that may come from administrative sources.

Thank you for your attention!

ZarovaEV@develop.mos.ru