

Методы машинного обучения и технологии искусственного интеллекта в семантическом анализе результатов анкетного наблюдения в официальной статистике

*Елена Зарова, д.э.н., профессор,
Аналитический центр Москвы*



1. ПОНЯТИЕ СЕМАНТИЧЕСКОГО АНАЛИЗА

Происхождение
слова

- ❑ Семантика – от греческого слова **σημαντικός** «обозначающий»

Суть метода

- ❑ Семантический анализ – это процесс извлечения **явного и скрытого** смысла из текста. В отличие от традиционных методов анализа текстов, которые фокусируются на их прочтении и систематизации по некоторым признакам, семантический анализ стремится понять **глубокие смыслы текстов** и их наборов (**корпусов**), анализируя синтаксические структуры, психо-лексические связи, контекст и другие факторы, влияющие на смысл



МЕТОДЫ СТАТИСТИКИ И МАШИННОГО ОБУЧЕНИЯ В ИНТЕЛЛЕКТУАЛЬНОМ АНАЛИЗЕ ТЕКСТОВЫХ ДАННЫХ (*text mining*) – РЕЗУЛЬТАТОВ АНКЕТНОГО ОПРОСА

Кластеризация, классификация

Группировка документов по сходству, например, по темам или стилям, обучение моделей для автоматической классификации документов по заданным категориям

1

2

Извлечение признаков

Автоматическое выделение ключевых слов, их комбинаций и других признаков, для "глубокого" выявления смысла текста

3

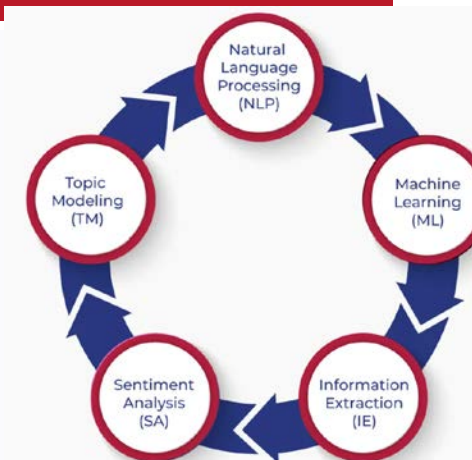
Сентимент-анализ

Оценка эмоциональной тональности скрытого смысла текста

4

Предсказание

Использование моделей машинного обучения для предсказания будущих событий или тенденций на основе анализа текстовых данных



Пример -дневниковая форма регистрации времени. Это структурированная форма саморегистрации, где допускается **некоторая вариативность в описании**, но основная информация должна быть внесена по установленным правилам/**ПОЛУСТРУКТУРИРОВАННЫЕ ДАННЫЕ**

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

Представляют:

интервьюеры выборочного наблюдения использования суточного фонда времени населением

- территориальному органу Росстата по установленному им адресу

Сроки представления

до 3 октября 2024 г.

ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ

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

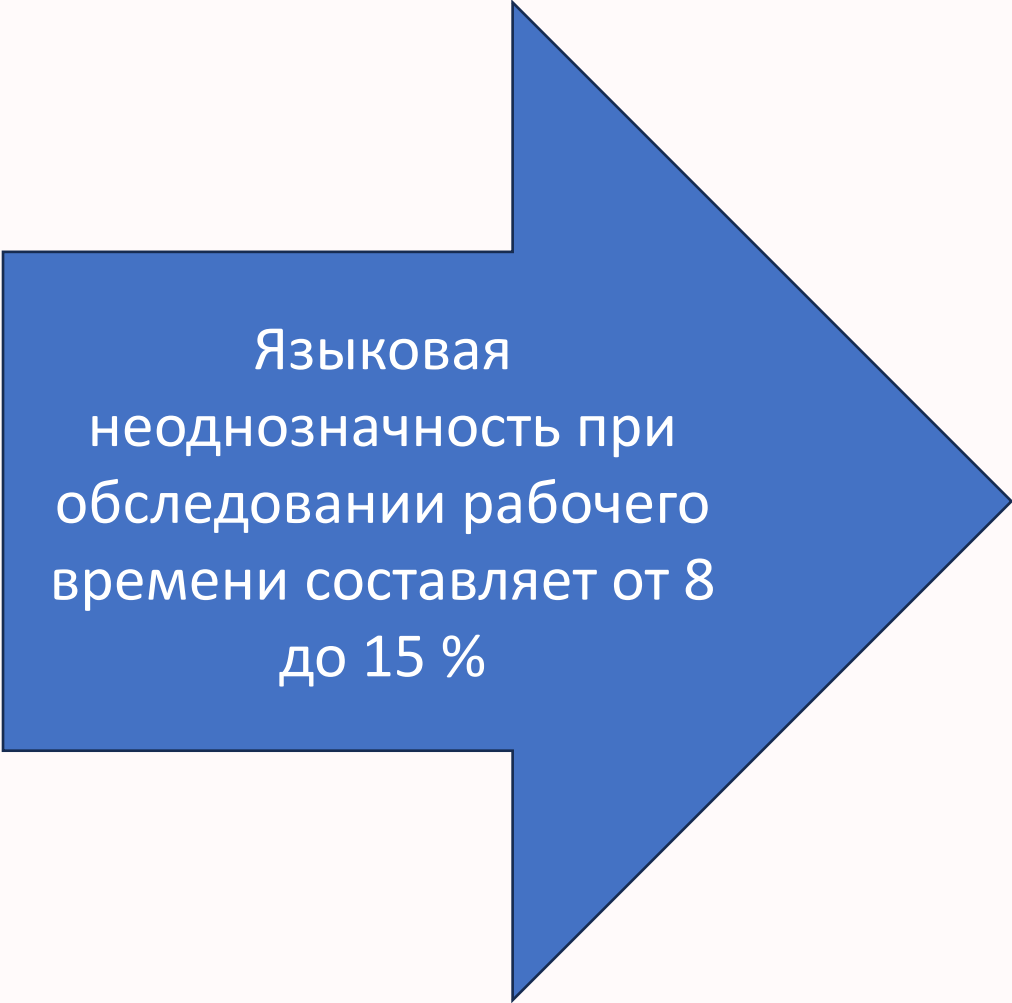
Территория

Время	Что Вы делали?		Что еще Вы делали? Запишите самое важное параллельное занятие		Использование сети Интернет	Где вы были?		Вы были один (одна) или с кем-то из знакомых Вам людей?			
	Запишите свое основное занятие в каждый 10-минутный интервал с 16.00 до 19.00	КОД вида основной деятельности		КОД вида параллельной деятельности		Запишите расположение или вид транспорта например дома, дома у друзей, в школе, на рабочем месте, в магазине, пешком, в автомобиле, автобусе	КОД места нахождения	Одна/ (один)	С детьми 0-9 лет, живущими в Вашем домо-хозяйстве	С другими членами Вашего домохо-зяйства	С другими знако-мыми 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 смарт-фоне		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	

Семантический анализ для обработки результатов анкетного опроса



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.

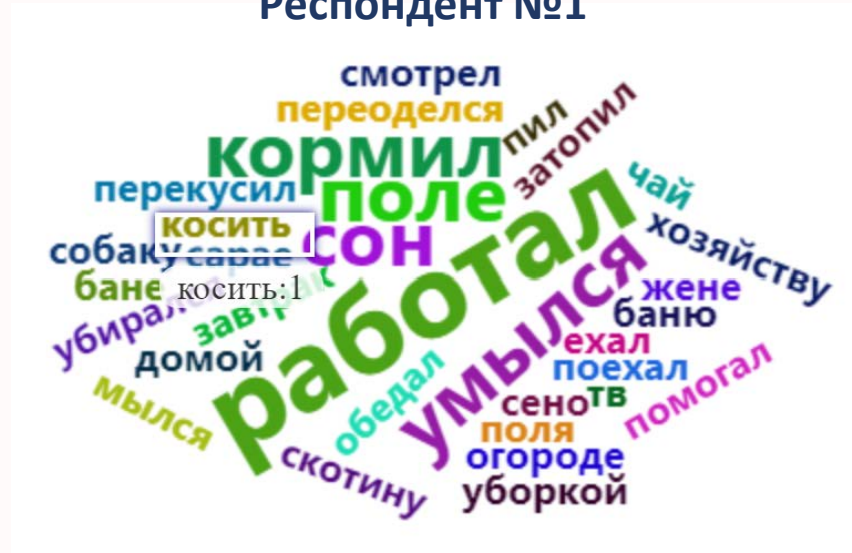
В результате часть ответов оказывается "размытыми", двусмысленными или слишком общими для автоматической категоризации.

Примеры:

- Ответ: "Работаю дома" — может означать как удалённую работу, так и работу по дому (домашние обязанности).
- Ответ: "Занимался делами" — не раскрывает конкретный вид деятельности.

Примеры частотного анализа записей дневников использования времени пятью респондентами - ОБЛАКА СЛОВ

Респондент №1



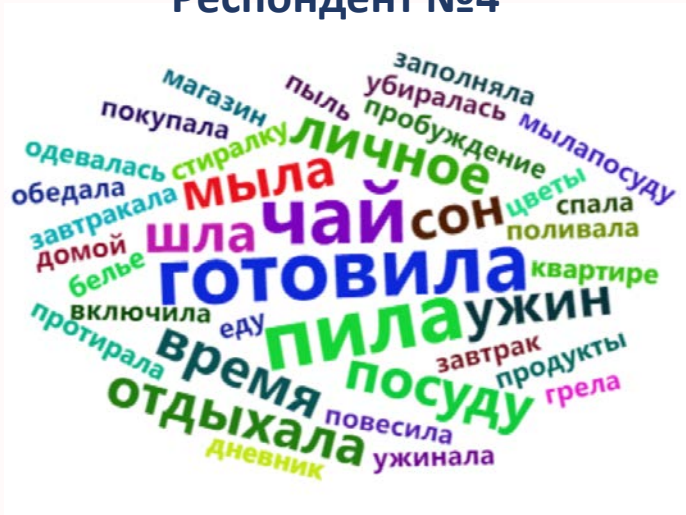
Респондент № 2



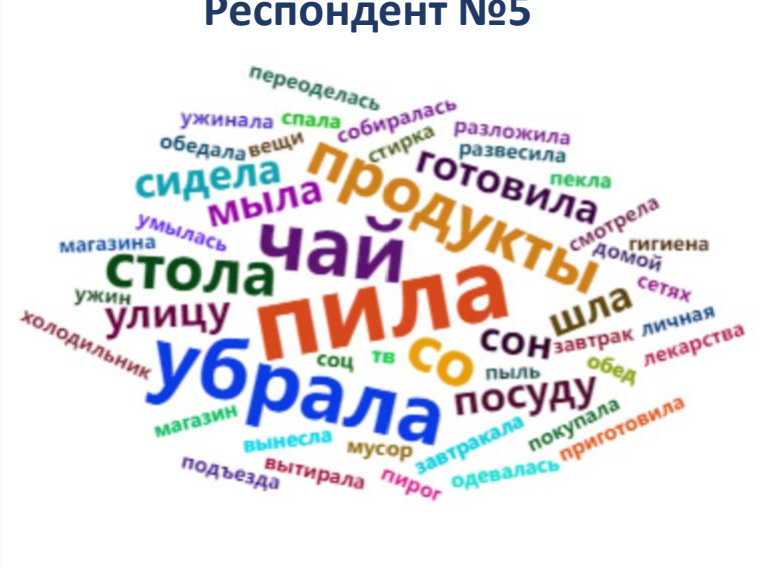
Респондент №3



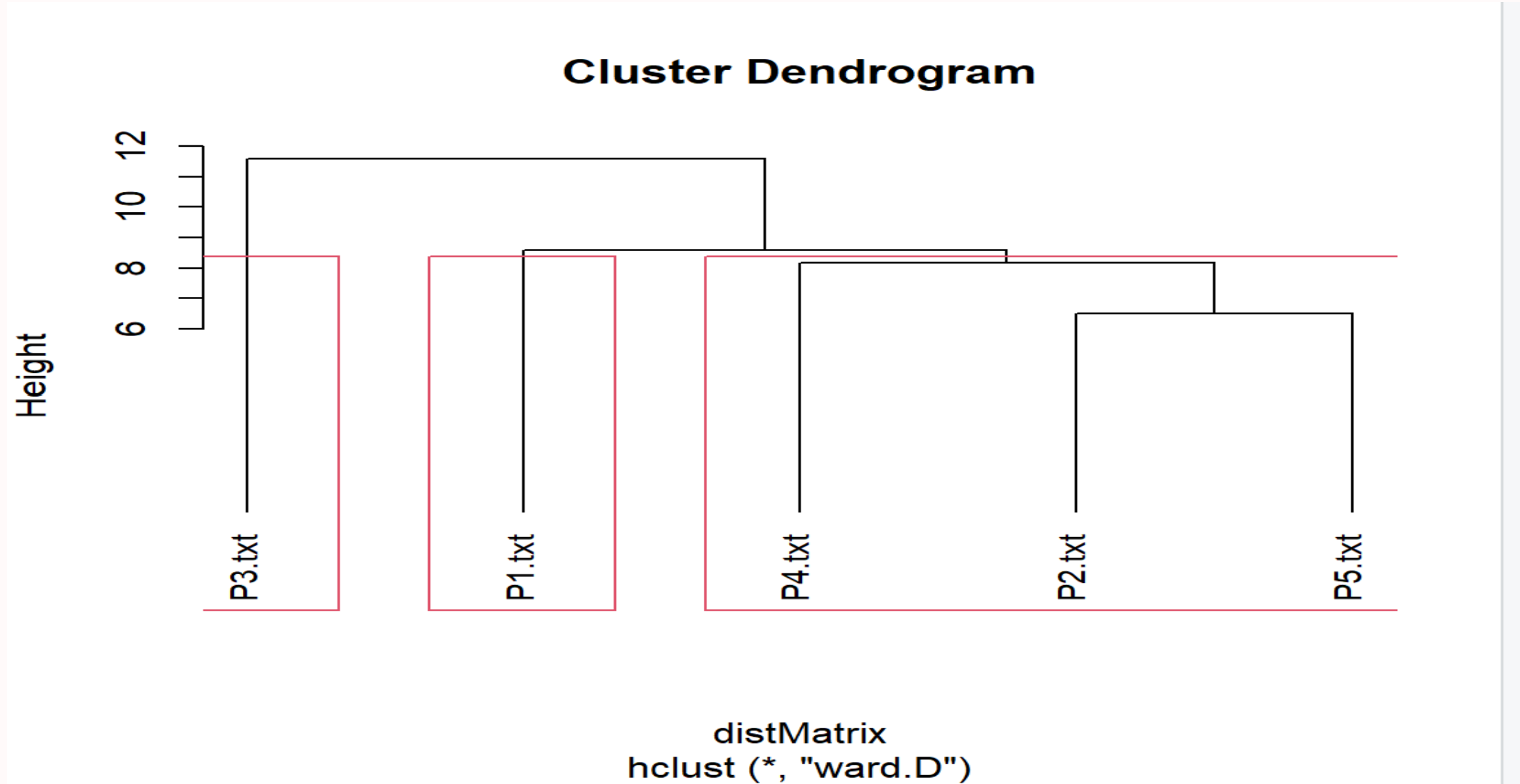
Респондент №4



Респондент №5



Кластеризация дневников использования времени по содержанию



Респонденты 2 и 5 дали **весьма схожие ответы** (надо проверить интервьюера)

Оценка ассоциаций показывает избыточность слова «сарае» по первому респонденту и наименований приема пищи – по второму. Надо усовершенствовать указания с типовыми ответами

Респондент №1

Ассоциации к слову «работал»

\$работал			
	сарае	хозяйству	поле
	0.55	0.55	0.33

Респондент № 2

Ассоциации к слову «готовила»

\$готовила			
	завтрак	обед	ужин
	0.56	0.56	0.56

Выводы.

Полученные результаты – это «проба пера», но она подтвердила необходимость применения текст майнинга в обработке анкетных данных. Это реально повышает эффективность извлечения информации из текстовых данных, обеспечивает повышение ее качества. В перспективе – нейронные сети для обработки огромного объема текстовых данных (неструктурированных и полуструктурированных, которые могут «прибыть» из административных источников)

Спасибо за внимание!

ZarovaEV@develop.mos.ru