РГП на ПХВ «Институт информационных и вычислительных технологий» КН МОН РК

Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана

Научный руководитель: Мусабаев Р.Р., к.т.н.

Соруководитель: Касымжанов Б.К., с.н.с.

Лаборатория «Анализа и моделирования информационных процессов»

Группы проекта

I. **Мусабаев Р.Р.**: Уалиева И.М., Красовицкий А.М., Мейрамбеккызы Ж., Аманбай А., Козбагаров О.Б., Төлеу А., Төлеген Г., Сейтқали Д., Нурзакова Ж.

II. **Мухамедиев Р.И.**: Якунин К.О., Кучин Я.И., Сымагулов А., Мурзахметов С.Б., Мустакаев Р.Р., Шалқарбайұлы А.

III. **Техническая**: Касымжанов Б.К., Ибраева В.М., Мукашев А.Ш., Меркебаев А.Г., Шахмаев Р.А., Кулемзин А.А., Айтмухамбетова Г.А.

IV. **АО «ИАЦ»**: Булдыбаев Т. – руководитель проекта соисполнителя

V. **Иностранные ученые:** Барахнин В.Б., Кожемякина О.Ю., Хорошилов А.А., Младенович Н.

Цель проекта

Разработка методических и технологических основ применения информационной системы социального доверия с целью стимулирования устойчивого развития личности с использованием технологий «Больших данных».

Задача. Создание необходимых технических и экспертно-аналитических условий для разработки информационной системы оценки влияния открытых текстовых информационных источников на социум

- Внедрение документов на основе вариационного автоэнкодера с рекуррентной нейронной сетью
- Реферирование текстового документа с помощью Word Mover’s Distance и извлеченных ключевых слов документа
- Группировка новостных публикаций по инфоповодам с помощью методов кластеризации
- Разработаны технологии создания декларативных средств для кластеризации документов СМИ (на основе методов семантического анализа текстов)
- Разработаны методики для автоматического формирования тематических словарей социально-значимых понятий
- Разработан метод декомпозиций в кластеризации

Using Centroid Keywords and WMD for Single Document Extractive Summarization - Использование центроидных ключевых слов и WMD для обобщения извлечения одного документа

- **Extractive** – формируются из имеющихся предложений в тексте
- **Single Document** – используется информация только одного документа
- **Dataset:** DUC 2002 – 567 новостей и их суммаризации
- Метрика оценки качества **ROUGE**

Описание метода

1. **Centroid word embedding:**
   Встраивание центроидного слова

$$C = \sum_{w \in S} E[idx(w)]$$

2. Cosine distance to C: Косинусное расстояние до C

$$Q(w_i) = \frac{E(w_i) \cdot C}{|E(w_i)||C|}$$

3. Sentence scoring with WMD:

$$U(S_i) = D_{WMD}(S_i, K)$$

Что уже есть:

- Есть методы где используются centroid embeddings предложений и документов.
- Есть работы где берут WMD между предложениями в документе.

В чем новизна?

- В этой работе предлагается использовать преимущества обеих методов в комбинации.

Результаты и замечания

System	Recall	Precision	F-measure
S28	.47813	.45779	.46729
S19	.45563	.47748	.46309
S21	.47543	.44635	.46029
Baseline	.47788	.44680	.46172
Our system	**0.45747**	**0.45453**	**0.45582**
TextRank	.46165	.43234	.44640
S29	.46100	.44557	.45269
S23	.43188	.47585	.45018
S27	.45485	.44808	.45014
MEAD	.44506	.45290	.44729
s15	.44805	.43323	.44014

System	Recall	Precision	F-measure
gold to gold	0.50576	0.50566	0.50549

Table 2: ROUGE-1 evaluation scores for our system, top 7 DUC02 systems, MEAD, TextRank, and the baseline.

Выводы:

- По результатам ROUGE предложенный метод может конкурировать с state of the art системами суммаризаций.
- Максимально объективный score который может достигнуть системы это 50% F-меры, выше этой отметки можно считать overfitting-ом.

Замечания:

- Использовать tf-idf.
- Обосновать почему 25% ближайших слов к центру являются ключевыми словами.

Word mover's distance

Идея: Расстояние между текстами, ***D*** *– это минимальная потраченная работа для транспортировки одного текста в другую. Чем меньше затрачено работы тем больше схожи два текста между собой.*

Работа = (вес слова) x (дистанция)

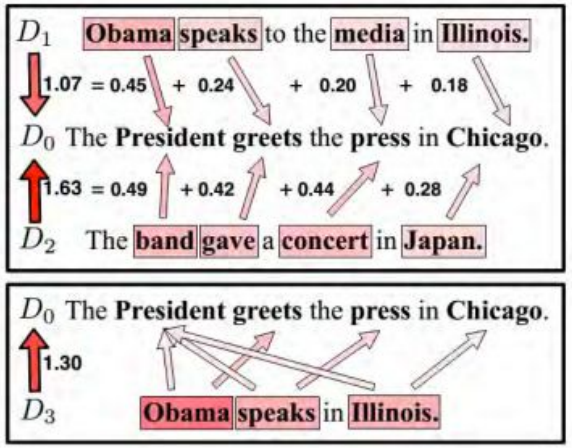
TF-IDF
Машинное обучение

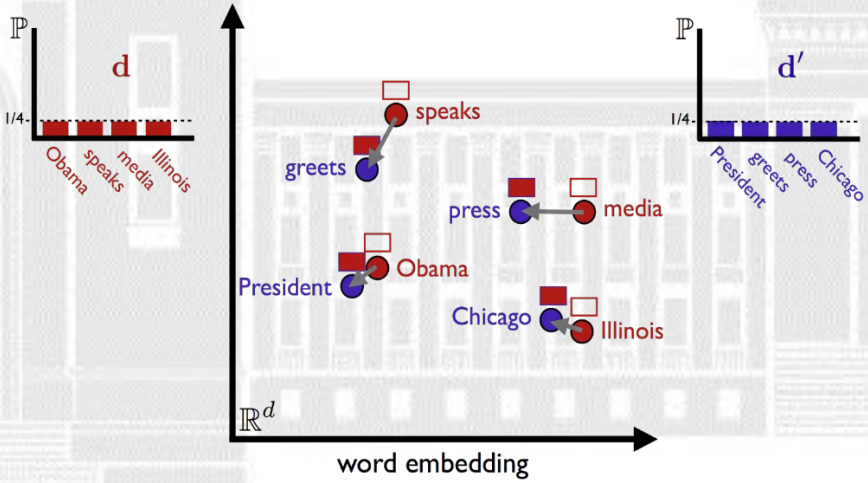
Word2Vec:
Нейронные сети

Формула:

$$D = \sum_{i,j=1}^{n} T_{ij} c(i,j) \to \min_{T \geq 0} \quad \text{где,} \quad \sum_{j=1}^{n} T_{ij} = d_i, \ \sum_{i=1}^{n} T_{ij} = d'_j \quad \text{при,} \quad d_i \geq 0, d'_j \geq 0 \ \sum_i d_i = 1, \sum_j d'_j = 1$$

Пример:





Группировка новостных публикаций по инфоповодам с помощью методов кластеризации

Постановка задачи:

Разработать подходы к группировке текстовой информации по инфоповодам на основе их семантического содержания с помощью методов кластеризации

Область применения – разрабатываемая информационная система для анализа новостных статей, публикуемые в казахстанском сегменте средств массовой информации на русском языке.

Инфоповод – это одно событие, происшествие или заявление, которое тиражируется в СМИ.

Комбинированный подход: Мера Жаккара + WMD

Разработанная функция расстояния между публикациями:

$$D(i,j) = 1 - \frac{\sqrt{J(Title_i, Title_j) + W(Text_i, Text_j)}}{\sqrt{2}}$$

$J(Title_i, Title_j)$ - близость Жаккара между заголовками двух статей

$W(Text_i, Text_j)$ - близость WMD между текстами статьи

Мера Жаккара :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Мера, основанная на Word Mover's Distance:

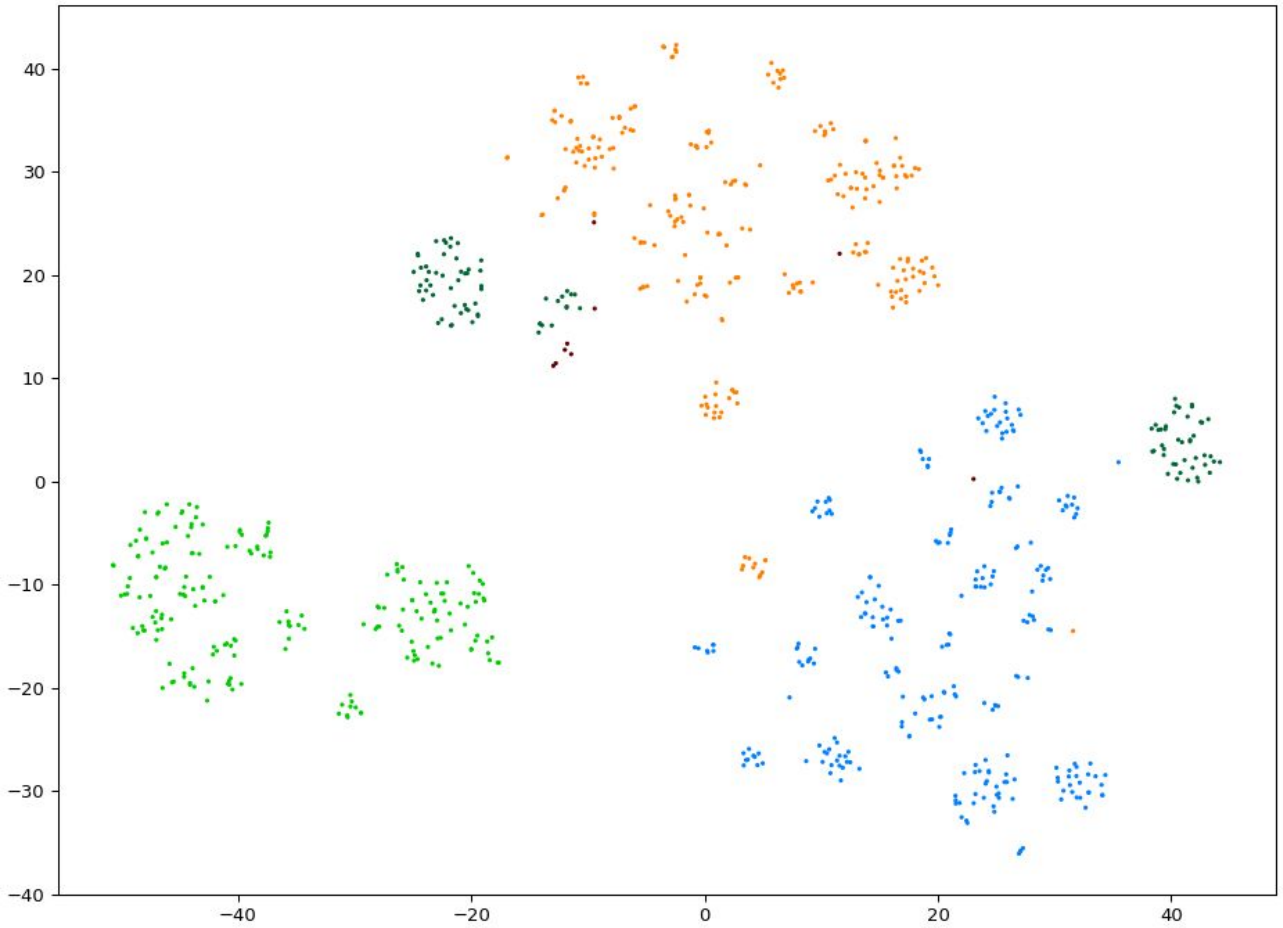
$$W(A, B) = max\left(0, 1 - \frac{d_w(A,B)}{median\, d_w(.,.)}\right)$$

Комбинированный подход: Мера Жаккара + WMD

Функция расстояния	Мера Жаккара (заголовки) + WMD (тексты)		WMD (тексты)	
Параметры	по умолчанию	подбор оптимальных: preference = -1.6; остальные по умолчанию	по умолчанию	подбор оптимальных (Adjusted_rand) preference = -1.6; остальные по умолчанию
Число кластеров	88	84	100	76
V-measure	91.79%	91.80%	89.43%	89.89%
Adjusted mutual information Скорректированная взаимная информация	77.57%	77.09%	73.67%	72.24%
Adjusted_rand	70.83%	71.76%	60.86%	64.96%
Fowlkes-Mallows	71.29%	72.34%	61.29%	66.27%

Комбинированный подход: Мера Жаккара + WMD

t-SNE (t-distributed Stochastic Neighbor Embedding)



Светло-зеленым цветом - новости раздела финансы

Темно-зеленым – спорт (футбол)

синие - происшествия

оранжевые - политика

темно-коричневые - уникальные новости - это новости спорта (кроме футбола) и новости культуры и военного дела.

Применимость разработанного подхода к “большим данным”

- Время вычисления матрицы дистанций WMD 822 x 822 составило около **130 минут** (16 процессов было задействовано).

- Если корпус состоит из **1 000 000** статей, то время вычисления матрицы дистанции WMD составит примерно 130*10^6 минут или 36 111 дней или **99 лет**.

- Таким образом, требуется модифицировать подход с целью применения к “большим данным”.

Виды представления публикаций

Во всех случаях за исключением stopwords	NOUNS	NOUNS + VERBS	ALL WORDS
	wmd около _45 минут	wmd около _85 минут	wmd около _130_ минут
Число кластеров	106	104	108
V-measure	0.8929	0.8927	0.8920
Adjusted mutual information	0.7403	0.7376	0.7394
Adjusted_rand	0.6044	0.6082	0.6028
Fowlkes-Mallows	0.6086	0.6123	0.6071

Первые k предложения новостной публикации

Во всех случаях за исключением stopwords	K=6	K=5	K=4	K=3	K=2	K=1
	wmd около 53 минуты	wmd около 43 минуты	wmd около 30 минут	wmd около 22 минут	wmd около 16 минут	wmd около 12 минут
Число кластеров	112	112	115	116	115	121
V-measure	0.8880	0.8951	0.8897	0.8808	0.8520	0.8128
Adjusted mutual information	0.7352	0.7546	0.7402	0.7205	0.6527	0.5612
Adjusted_rand	0.5898	0.6010	0.5908	0.5545	0.5020	0.3912
Fowlkes-Mallows	0.5952	0.6071	0.5972	0.5601	0.5078	0.3976

Комбинированный подход: Мера Жаккара + Word's Average

Функция расстояния:

$$D(i,j) = 1 - \frac{\sqrt{J(Title_i, Title_j) + W(Text_i, Text_j)}}{\sqrt{2}}$$

$J(Title_i, Title_j)$ - близость Жаккара между заголовками двух статей

$W(Text_i, Text_j)$ –близость между текстами статьи согласно формуле ниже

Мера Жаккара :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Мера, основанная на евклидовом расстоянии:

$$W(A,B) = max\left(0, 1 - \frac{d_w(A,B)}{median\, d_w(.,.)}\right)$$

Комбинированный подход: Мера Жаккара + Word's Average

Функция расстояния	Мера Жаккара + WMD	Мера Жаккара + Word's Average Texts	Word's Average Titles + Word's Average Texts
Параметры	по умолчанию	по умолчанию	по умолчанию
Число кластеров	88	86	86
V-measure	91.79%	91.67%	88.77%
Adjusted mutual information	77.57%	76.96%	70.74%
Adjusted_rand	70.83%	70.83%	58.06%
Fowlkes-Mallows	71.29%	71.39%	59.09%

Применимость разработанного подхода к “большим данным”

- Рассмотрен корпус из **10 000** новостей. Время вычисления матрицы евклидова расстояния данного корпуса (1 процесс было задействован) составило **72 минуты**.

- Если корпус состоит из **1 000 000** статей, то время вычисления матрицы дистанции составит примерно 720 000 минут или **200 дней**.

Технологии создания декларативных средств для кластеризации документов СМИ (на основе методов семантического анализа текстов)

Задачи исследования

1. *Разработать новые методы, алгоритмы и технологии решения задачи создания декларативных средств для автоматической кластеризации текстовых документов СМИ.*
2. *Исследовать и разработать методы и алгоритмы выделения из текстов сущностей (значимых понятий) для задачи кластеризации.*
3. *Исследовать и разработать алгоритмы формирования частотных словарей слов и словосочетаний и представить их в табличном виде.*
4. *Исследовать и разработать технологии и процедуры назначение элементам формализованного представления документа весовых коэффициентов их смысловой значимости.*
5. *Выполнить анализ полученных результатов при различных исходных данных.*
6. *Разработать общую технологическую схему процесса создания декларативных средств для автоматической кластеризации текстовых документов СМИ.*

Теоретическая концепция фразеологического концептуального анализа текстов

Основной идеей этой концепции *является обоснование использования в качестве основных единиц смысла устойчивых фразеологических и терминологических словосочетаний, обозначающих понятия и отношения между понятиями, представленные в предметной области.*

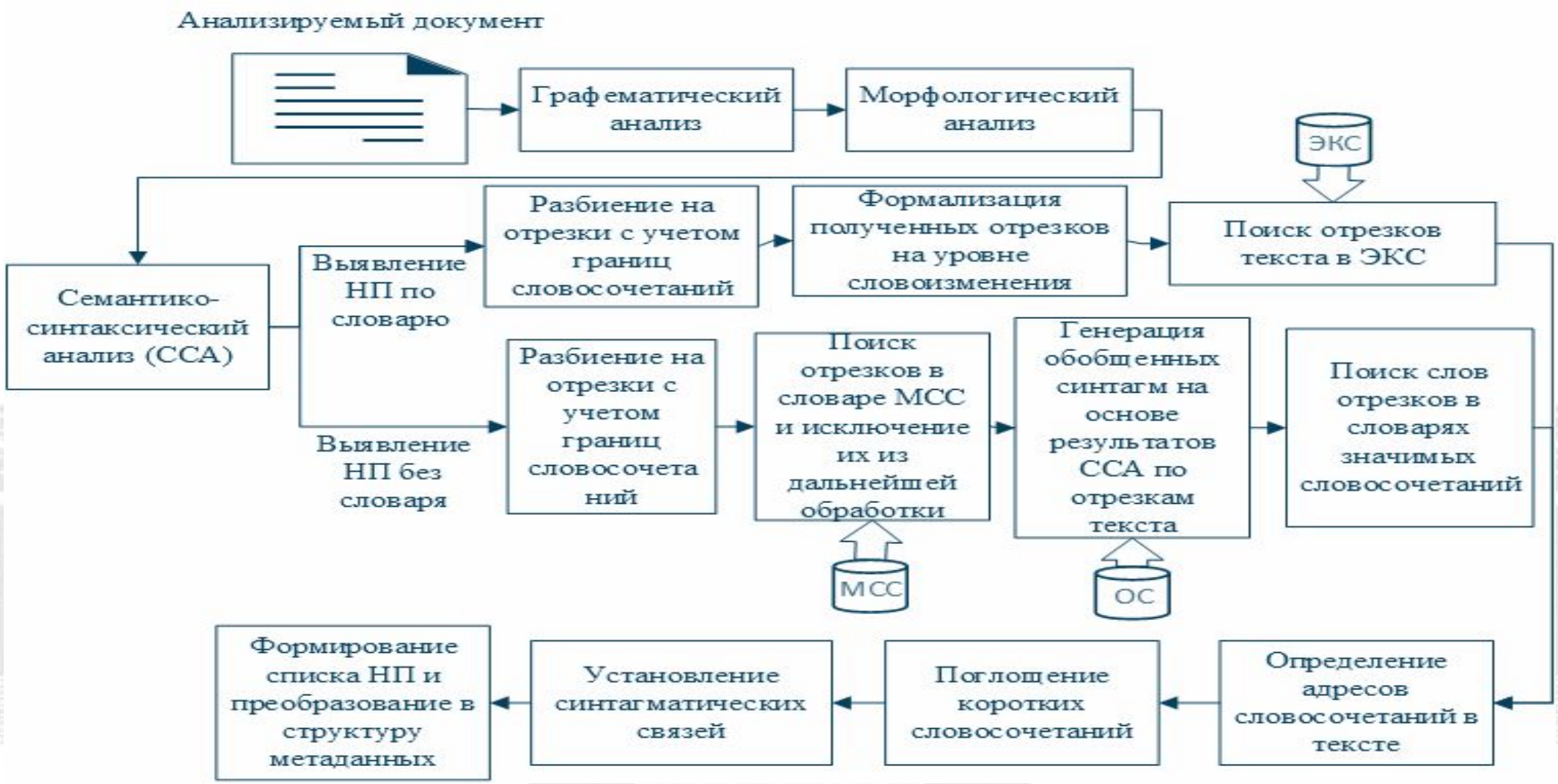
Иерархия единицы смысла:

- *Наименование понятия (сущность) – выражено словом или словосочетанием*
- *Предложение – его смысловой структурой является предикатно-актантная структура*
- *Сверхфразовое единство –фрагмент текста, объединенный общей темой*

Смысловое представление содержания текста *- концептуальный образ документа (КОД) - совокупность взаимосвязанных наименований понятий текста, расположенных в нем строго определенном порядке)*

Семантическая карта документа *– концептуальный граф, в котором вершины – нормализованные наименования понятий, дуги – унифицированные смысловые отношения между понятиями*

Гибридный алгоритм №5 выявления наименований понятий в текстах документов



Исходные статистические данные по массиву сообщений СМИ

- *Кол. Документов в массиве = 3 004 документов*
- *Всего слов в массиве документов= 523 810 слов*
- *Разных слов (на уровне словоизменения) = 88 925*
- *Среднее число слов в документе = 174.4 слов/док*
- *Среднее число разных слов в документе = 29.5 слов/док*

- *Всего словосочетаний в массиве (по словарю ЭКС)= 1 106 355 словосоч.*
- *Разных словосочетаний (на уровне словоизменения слов) = 67 571 словосоч.*
- *Кол. разных главных слов (на уровне словоизменения слов) = 5 577слов*
- *Среднее число словосочетаний в документе = 368.3 словосоч./док*
- *Среднее число разных словосочетаний в документе = 22.5 словосоч./док*

Результаты выполненных исследований

- *Разработаны новые методы, алгоритмы и технологии решения задачи создания декларативных средств для автоматической кластеризации текстовых документов СМИ.*
- *Исследованы и разработаны методы и алгоритмы выделения из текстов сущностей (значимых понятий) для задачи кластеризации.*
- *Разработаны алгоритмы формирования частотных словарей слов и словосочетаний и представления их в табличном виде.*
- *Разработан алгоритм формирования смыслового представления документов.*
- *Разработаны технологии и процедуры назначение элементам формализованного представления документа весовых коэффициентов их смысловой значимости.*
- *Выполнен предварительный анализ полученных результатов при различных исходных данных.*

Автоматическое формирование тематических словарей социально-значимых понятий

Распознавание социально значимых тем во множестве разнотематических новостных данных.

Какие темы можно отнести к социально значимым?

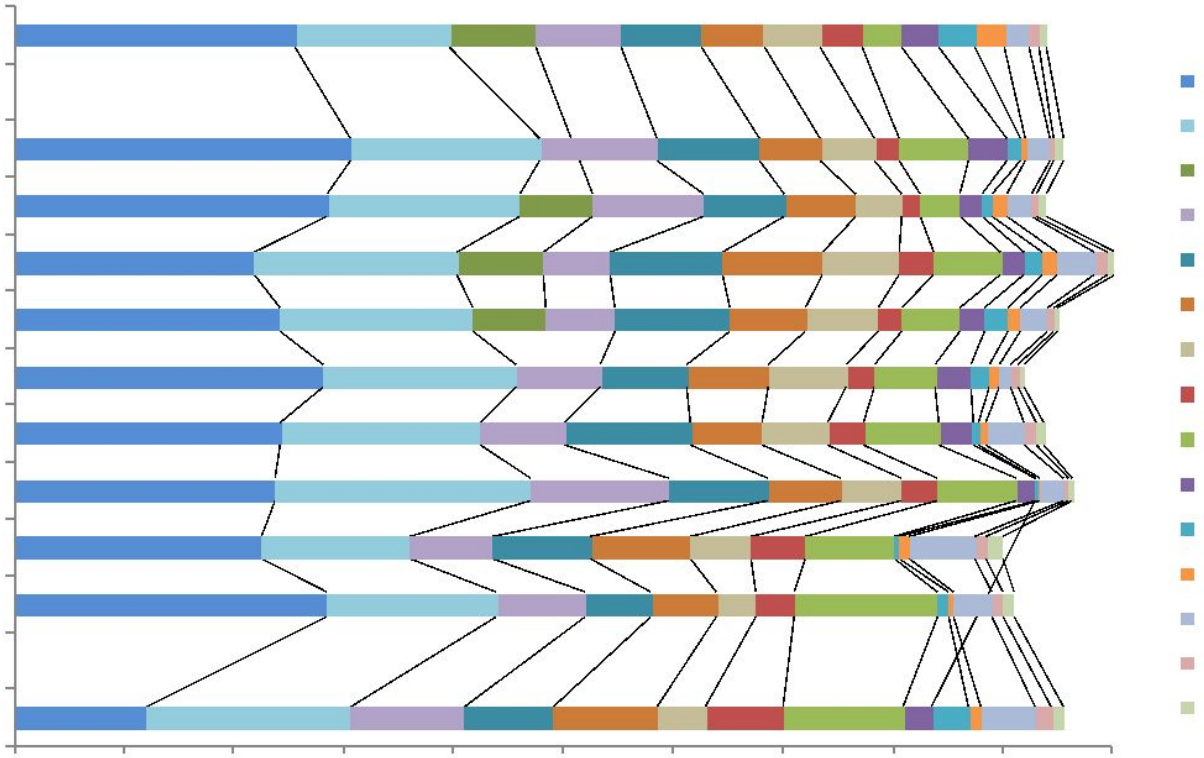


ТАБЛИЦА 1. Статистика по данным социологических исследований ЦСПИ «Стратегия»

Алгоритм выявления социально значимых новостей из кластеров новостных статей

1.Резонансные

Самые частотные темы (размер кластера пропорционален резонансности)

2. Скандально-событийные (шоу-бизнес)

1.Словарь именованных сущностей (например, самые популярные имена из шоу-бизнеса) 2. Словарь именованных сущностей из социально значимых тем

3. Социально значимые

Словарь статических тем

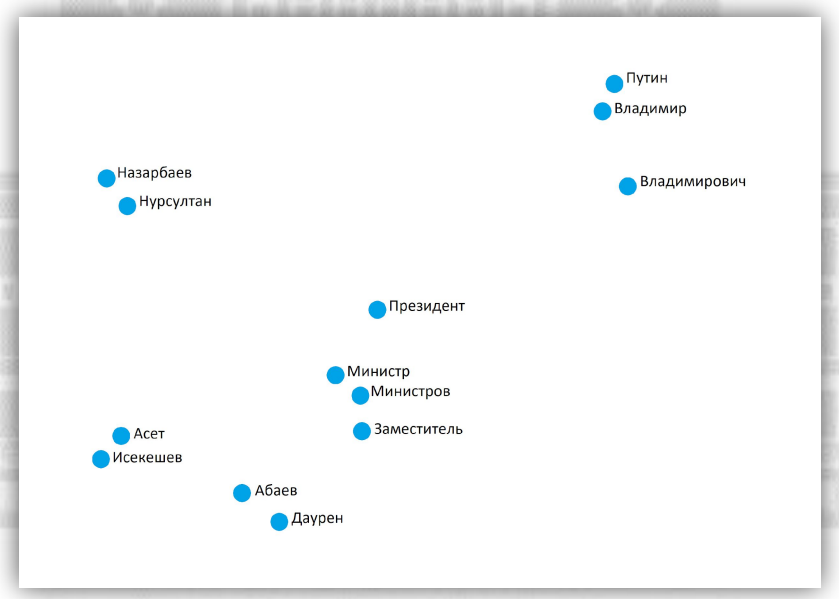
*относительно большого корпуса новостей 2,3 млн статей

I ФОРМИРОВАНИЕ ТЕМАТИЧЕСКИХ СЛОВАРЕЙ НА ОСНОВЕ CO-OCCURRENCE МАТРИЦЫ

	спорт	бокс	аэропорт	театр	Казахстан	Головкин	Геннадий
Головкин	8	9	1	0	6	0	10
бокс	3	0	0	0	2	8	7

Матрица смежности слов

II ТЕМАТИЧЕСКИЕ СЛОВАРИ НА ОСНОВЕ WORD2VEC



Метод декомпозиций в кластеризации

Мотивация

- Кластеризация на больших наборах данных. В задачах NLP актуальна для тематической кластеризации текстов, составления тематических словарей, других задачах с набором данных в метрическом пространстве
- Хорошее качество кластеризации за разумное/приемлемое время
- ‘Рейтинговые’ соревнования на разных алгоритмах / и на разных наборах данных UCI

Оценки качества алгоритмов кластеризации

- Оценка на известных наборах данных с (частичной/полной) классификацией. Если данные размечены, например получены из UCI, то можем использовать скорректированный Рэнд индекс (adjusted Rand index)
- С помощью внутри- и меж- кластерных эвристик
- С помощью ***SSD*** (Sum of Square Distance) критерия
  - Не требует размеченных данных
  - Имеет статистический смысл
  - Оценка вычислимая быстро

Идея нашего метода

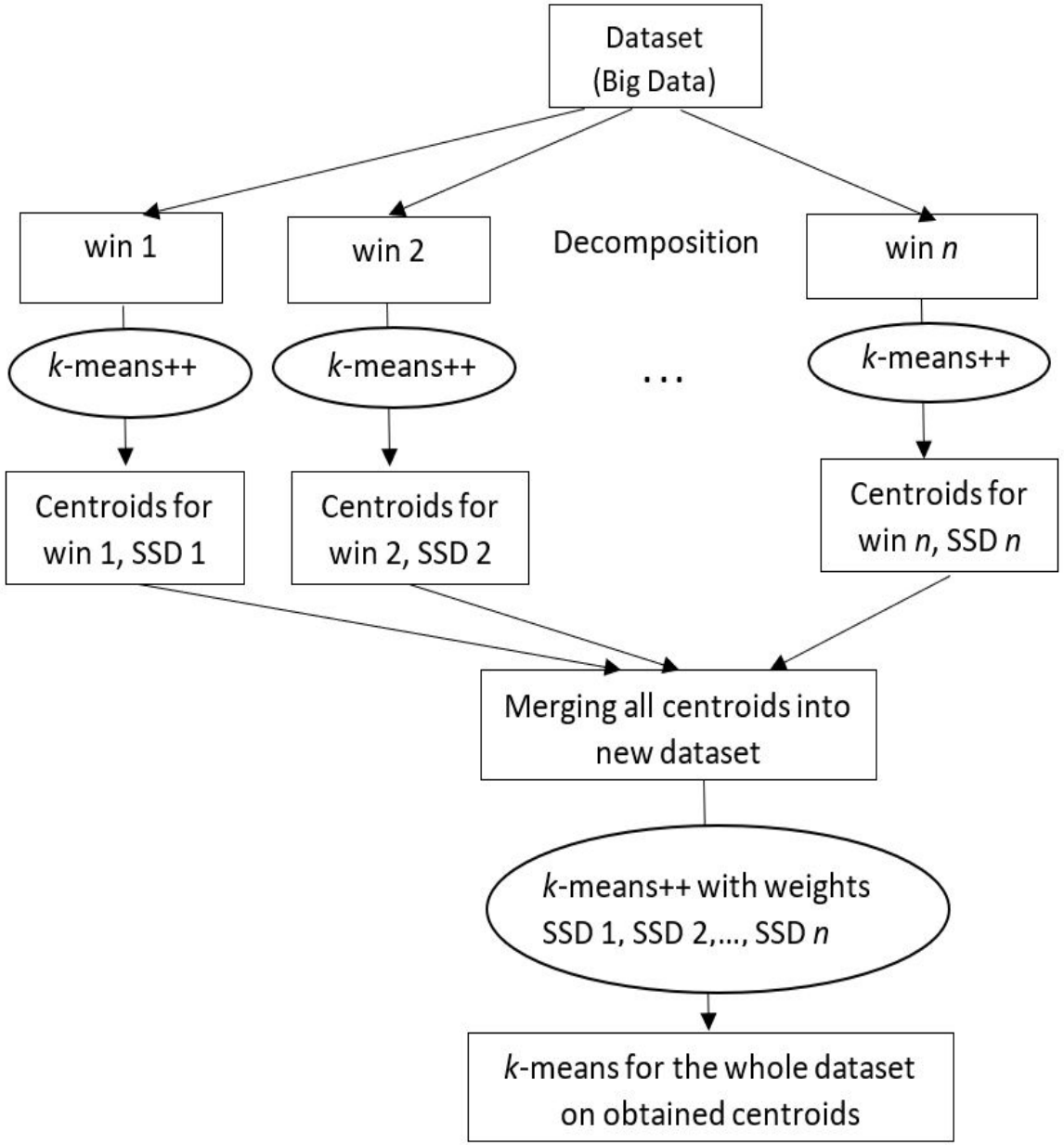
- Получать кластеризацию на сравнительно небольших подмножествах (выборках) исходных данных – окнах используя k-means++
- Найденные центроиды и их соответствующие значения SSD использовать для поиска улучшенной инициализации. Для этого используем взвешенную оценку.

- Преимущества подхода:
  - За счет сокращения числа вычислений с большей вероятностью находим оптимальную кластеризацию
  - Менее чувствителен к шумовым выбросам

Параллельная декомпозиция Phase 1

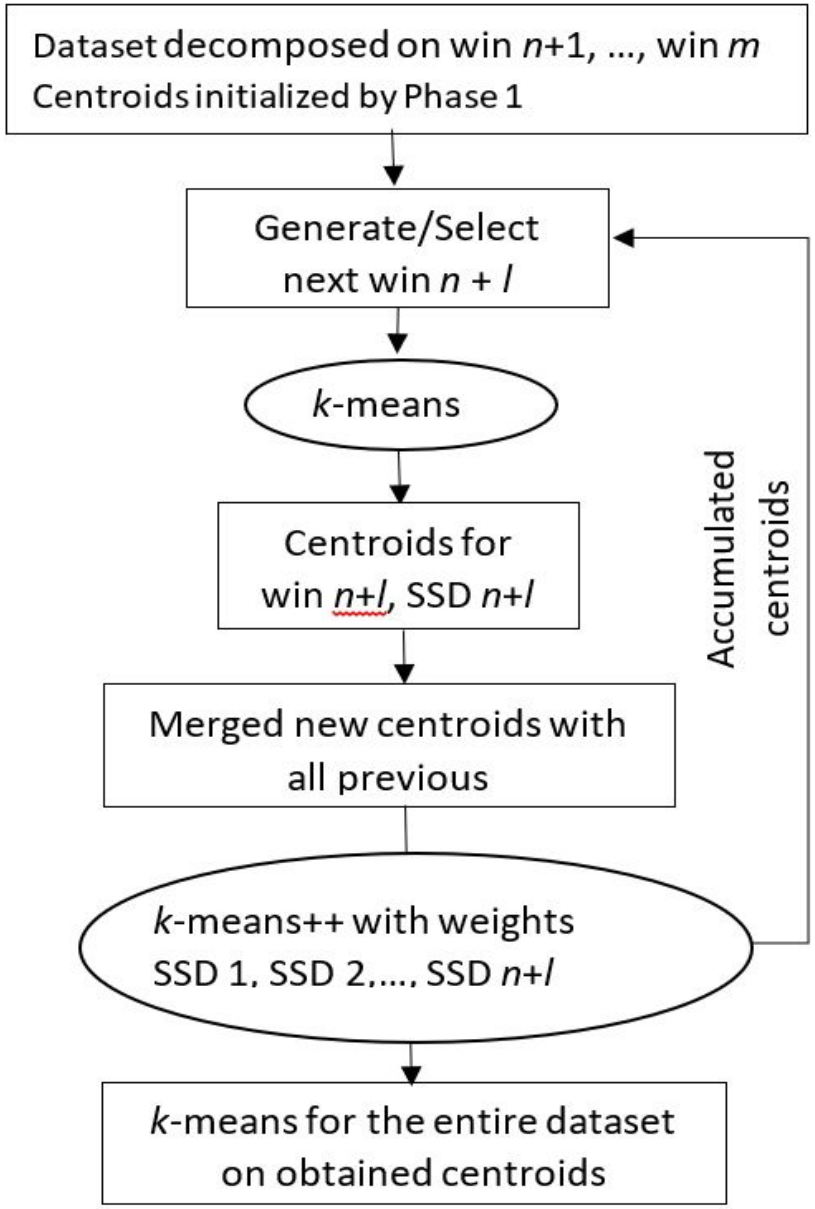
- Win 1,…, win n независимые выборки(окна) из полного набора данных

- SSD 1,…,SSD n соответствующие оценки



Последовательная декомпозиция Phase 2

- Используется предыдущий алгоритм для инициализации
- Добавление следующего окна *win n+l* вносит вклад в общее расположение начальных центроидов в соответствии с полученным SSD *n+l*
- Останов по заданному времени/числу итераций



*Результаты экспериментов на синтетических наборах данных и данных UCI***

dataset size	num. of experiments	improves k-means++		num. of overtime	
	1590	33 (2.08%)	0.964	0	0.330
	1186	35 (2.95%)	0.989	17	0.433
	312	14 (4.49%)	0.992	15	0.051
	1140	345 (30.26%)	1.000	10	0.386
	300	160 (53.33%)	1.000	12	0.394
	82	40 (48.78%)	0.996	41	0.971
	30	26 (86.67%)	1.000	30	1.650
	260	165 (63.46%)	1.000	0	0.160

Обобщение метода декомпозиций на другие алгоритмы кластеризации

- Заменить *k-means++* любым кластерным алгоритмом для которого критерий SSD имеет смысл, как например для
  - Mini batch k-means
  - J-means
  - H-means
  - Hybrid algorithms
  - etc . . .
- Все остальные шаги алгоритма поиска центроидов остаются неизменными
- Таким образом предлагаем обобщенную мета-эвристику для ускорения кластеризации на больших наборах данных

Благодарю за внимание