Методы определения семантической близости документов

Области применения:

1. Текстовый поиск в интернете.
2. Поиск «близких» документов.
3. Классификация текстов.
4. Устранение многозначности.

Методы:

1. По тексту
2. По связям

Методы:

1. По тексту
2. По связям

Латентно-семантический анализ

Задача: кластеризовать новости по заголовкам.

Британская полиция знает о местонахождении основателя WikiLeaks

В суде США начинается процесс против россиянина, рассылавшего спам

Церемонию вручения Нобелевской премии мира бойкотируют 19 стран

В Великобритании арестован основатель Wikileaks Джулиан Ассандж

Украина игнорирует церемонию вручения Нобелевской премии

Шведский суд отказался рассматривать апелляцию основателя Wikileaks

НАТО и США разработали планы обороны стран Балтии против России

Полиция Великобритании нашла основателя WikiLeaks, но, не арестовала

Подготовка:

1. Удаление стоп-слов
2. Стемминг
3. Удаление слов в единст- венном экземпляре

Британская **полиц**ия знает о местонахождении **основател**я **WikiLeaks**

В **суд**е **США** начинается процесс **прот**ив россиянина, рассылавшего спам

Церемонию **вручен**ия **Нобелевск**ой **прем**ии мира бойкотируют 19 **стран**

В **Великобритан**ии **арестова**н **основател**ь **Wikileaks** Джулиан Ассандж

Украина игнорирует **церемон**ию **вручен**ия **Нобелевск**ой **прем**ии

Шведский **суд** отказался рассматривать апелляцию **основател**я **Wikileaks**

НАТО и **США** разработали планы обороны **стран** Балтии **прот**ив России

Полиция **Великобритан**ии нашла **основател**я **WikiLeaks**, но, не **арестова**ла

Считаем количество раз вхождения каждого слова в документы и заносим в матрицу.

	T1	T2	T3	T4	T5	T6	T7	T8	T9
wikileaks	1	0	0	1	0	1	0	1	0
арестова	0	0	0	1	0	0	0	1	0
великобритан	0	0	0	1	0	0	0	1	0
вручен	0	0	1	0	1	0	0	0	1
нобелевск	0	0	1	0	1	0	0	0	1
основател	1	0	0	1	0	1	0	1	0
полиц	1	0	0	0	0	0	0	1	0
прем	0	0	1	0	1	0	0	0	1
прот	0	1	0	0	0	0	1	0	0
стран	0	0	1	0	0	0	1	0	0
суд	0	1	0	0	0	1	0	0	0
сша	0	1	0	0	0	0	1	0	0
церемон	0	0	1	0	1	0	0	0	0

Сингулярное разложение матрицы:

$M = U*W*V^t$

U и V^t – ортогональные

W – диагональная (элементы в порядке неубывания)

wikileaks	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	-0.64
арестова	0.34	-0	0.07	0.41	-0.42	-0.02	0.1	0.17	0.01
великобритан	0.34	-0	0.07	0.41	-0.42	-0.02	0.1	0.17	-0.01
вручен	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.07
нобелевск	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	0.32
основател	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	0.64
полиц	0.31	-0	0.05	0.07	0.57	-0.6	0.29	0.37	-0
прем	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.25
прот	0.02	0.03	-0.61	0.13	-0.05	-0.22	0	-0.25	0
стран	0.01	0.22	-0.31	0.39	0.41	0.56	-0.22	0.4	-0
суд	0.12	0.01	-0.38	-0.62	-0.3	0.12	0.21	0.55	-0
сша	0.02	0.03	-0.61	0.13	-0.05	-0.22	0	-0.25	0
церемон	0	0.38	0.03	0.02	0.08	0.31	0.82	-0.29	0

·

3.41	0	0	0	0	0	0	0	0
0	3.3	0	0	0	0	0	0	0
0	0	2.27	0	0	0	0	0	0
0	0	0	1.49	0	0	0	0	0
0	0	0	0	1.19	0	0	0	0
0	0	0	0	0	0.98	0	0	0
0	0	0	0	0	0	0.71	0	0
0	0	0	0	0	0	0	0.43	0
0	0	0	0	0	0	0	0	0

·

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.43	0.05	0.01	0.54	0	0.37	0.01	0.63	0
-0	0.02	0.65	-0.01	0.59	-0	0.09	-0.01	0.47
0.03	-0.7	-0.04	0.06	0.1	-0.16	-0.67	0.09	0.09
-0.22	-0.24	0.15	0.28	-0.11	-0.68	0.44	0.33	-0.13
0.69	-0.32	0.22	-0.49	-0.12	-0.03	0.27	-0.02	-0.19
-0.27	-0.34	0.44	0.29	-0.13	0.45	0.12	-0.31	-0.45
-0.03	0.3	0.14	-0.17	0.44	-0.15	-0.3	0.24	-0.71
-0.3	0.12	0.4	-0.39	-0.53	0.12	-0.23	0.46	0.13
0.35	0.35	0.35	0.35	-0.35	-0.35	-0.35	-0.35	0

Строки и столбцы с меньшим сингулярным числом дают меньший вклад в произведение.

Оставим только 2 самых весомых.

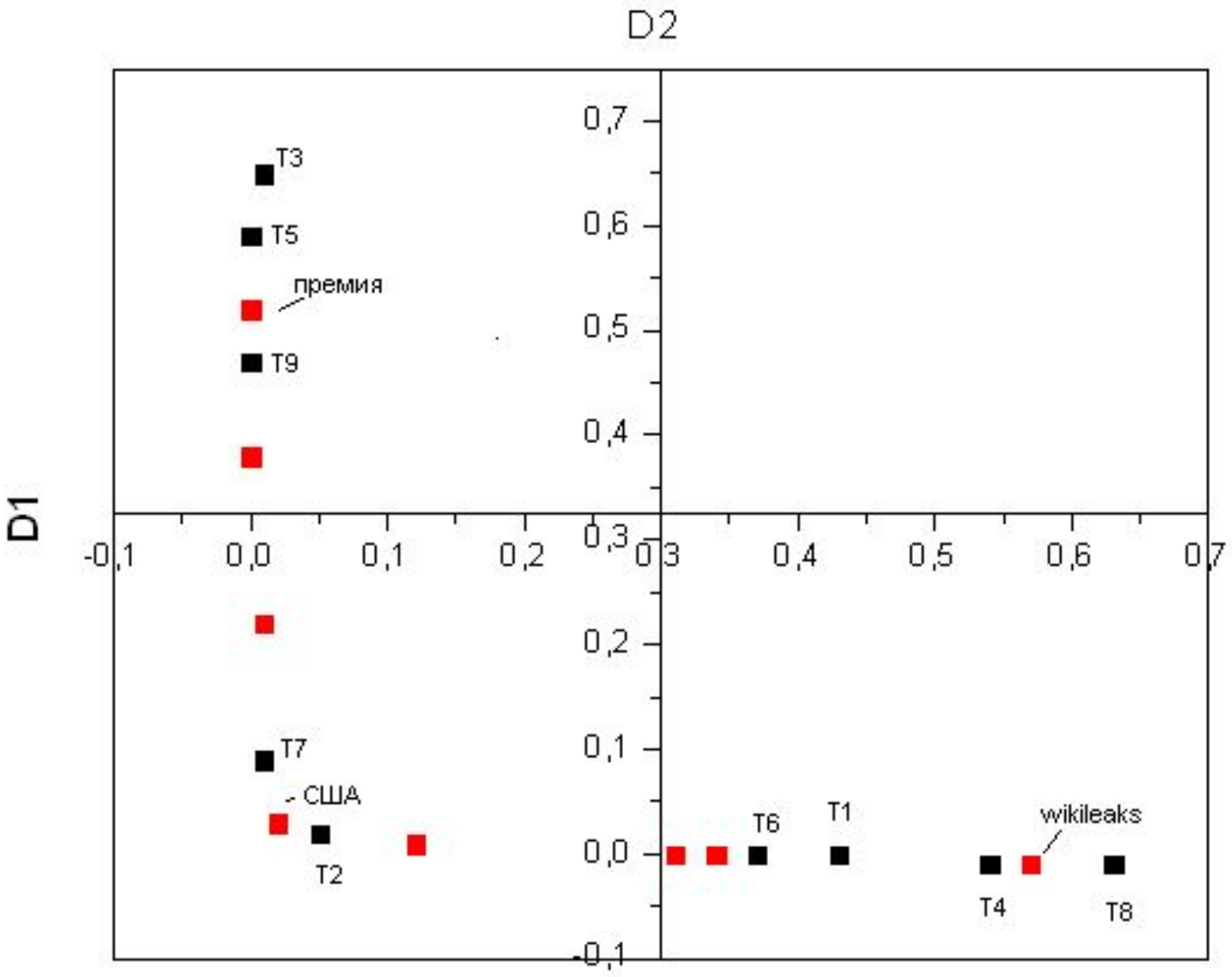
wikileaks	0.57	-0.01
арестова	0.34	-0
великобритан	0.34	-0
вручен	0	0.52
нобелевск	0	0.52
основател	0.57	-0.01
полиц	0.31	-0
прем	0	0.52
прот	0.02	0.03
стран	0.01	0.22
суд	0.12	0.01
сша	0.02	0.03
церемон	0	0.38

*

3.41	0
0	3.3

*

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.43	0.05	0.01	0.54	0	0.37	0.01	0.63	0
-0	0.02	0.65	-0.01	0.59	-0	0.09	-0.01	0.47


D2
D1
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
-0,1
-0,1
0,0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
T3
T5
премия
T9
T7
США
T2
T6
T1
wikileaks
T4
T8

Методы:

1. По тексту
2. По связям

Методы, использующие связи: абстрагируемся от текста, важны только связи между документами.

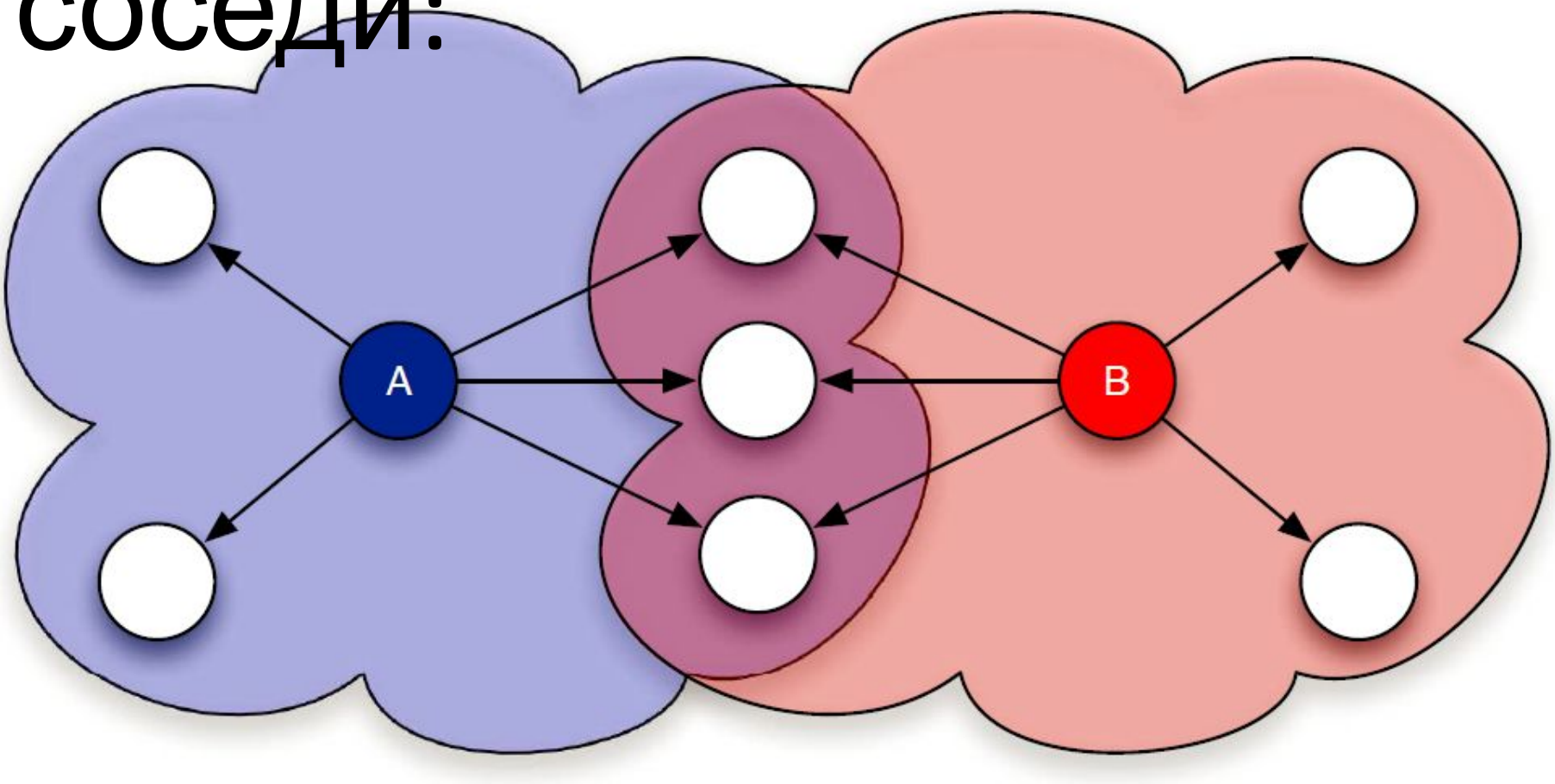
Унификация.

1. Локальные
2. Глобальные

1. Локальные
2. Глобальные

Локальные: близость определяется для пары вершин и не затрагивает большинство вершин.

Ближайшие соседи:



N(a) – множество ближайших соседей узла a

СимКо

$$sim_{cos}(a,b) = \frac{|N(a) \cap N(b)|}{\sqrt{|N(a)|^2 + |N(b)|^2}}$$

Коэффициент Жаккара:

$$sim_{Jaccard}(a,b) = \frac{|N(a) \cap N(b)|}{|N(a) \cup N(b)|}$$

Коэффициент Дайса:

$$sim_{Dice}(a,b) = \frac{2\,|N(a) \cap N(b)|}{|N(a)| + |N(b)|}$$

Для направленных графов:

- Со-цитирование
- Библиографическое сочетание

1. Локальные
2. Глобальные

Глобальные: вычисляют близость между всеми вершинами графа.

SimRank: два объекта похожи, если на них ссылаются похожие объекты

$$s(a,a) = 1, s(a,b) = \frac{C}{|I(a)|\,|I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

C – коэффициент затухания.

Метод итеративен.

$$R_0(a,b) = \begin{cases} 0 & (\text{if} \quad a \neq b) \\ 1 & (\text{if} \quad a = b) \end{cases}$$

$$R_{k+1}(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} R_k(I_i(a), I_j(b))$$

$$\lim_{k \to \infty} R_k(a,b) = s(a,b)$$

Затраты времени и памяти.

Базовый подход.

$O(n^2)$ памяти.

$O(Kn^2d_2)$ времени, где:

K – количество итераций

d_2 – среднее значение $|I(a)||I(b)|$ по всем (a, b)

Затраты времени и памяти.

Улучшенный подход: рассматриваем только близкие вершины в графе.

Пусть r – радиус в котором рассматриваются соседи.

d_r – среднее количество соседей в r.

$O(d_r n)$ памяти

$O(Knd_r d_2)$ времени

??