Лабораторная работа №3

Кластеризация текстов по тематикам

Цель работы: научиться реализовывать на выбранном языке программирования алгоритмы кластеризации текстов по тематикам на основе латентно-семантического анализа.

Краткие теоретические сведения

Латентно-семантический анализ отображает документы и отдельные слова в так называемое «семантическое пространство», в котором и производятся все дальнейшие сравнения. При этом делаются следующие предположения:

1) Документы это просто набор слов. Порядок слов в документах игнорируется. Важно только то, сколько раз то или иное слово встречается в документе.

2) Семантическое значение документа определяется набором слов, которые как правило идут вместе. Например, в биржевых сводках, часто встречаются слова: «фонд», «акция», «доллар»

3) Каждое слово имеет единственное значение. Это, безусловно, сильное упрощение, но именно оно делает проблему разрешимой.

Пример

Для примера возьмем несколько заголовков из различных новостей. На первом шаге из этих заголовков исключаем, так называемые, стоп-символы. Это слова которые встречаются в каждом тексте и не несут в себе смысловой нагрузки, это, прежде всего, все союзы, частицы, предлоги и множество других слов. Полный список использованных стоп-символов можно посмотреть в Приложении 1.

Далее необходимо выполнить операцию стемминга. Она не является обязательной, если набор текстов достаточно большой, или если тексты на английском языке, в силу того, что количество вариаций той или иной словоформы в английском языке существенно меньше чем в русском. В нашем же случае, пропускать этот шаг не стоит т.к. это приведет к существенной деградации результатов.

Дальше исключаем слова встречающиеся в единственном экземпляре. Это тоже необязательный шаг, он не влияет на конечный результат, но сильно упрощает математические вычисления. В итоге у нас остались, так называемые, индексируемые слова, они выделены жирным шрифтом:

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

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

3. **Церемон**ию **вручен**ия **Нобелевск**ой **прем**ии мира бойкотируют 19 **стран**

4. В **Великобритан**ии **арестова**н **основател**ь сайта **Wikileaks** Джулиан Ассандж

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

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

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

8. **Полиц**ия **Великобритан**ии нашла **основател**я **WikiLeaks**, но, не **арестова**ла

9.В Стокгольме и Осло сегодня состоится **вручен**ие **Нобелевск**их **прем**ий

Латентно семантический анализ

На первом шаге требуется составить частотную матрицу индексируемых слов. В этой матрице строки соответствуют индексированным словам, а столбцы — документам. В каждой ячейке матрицы указано какое количество раз слово встречается в соответствующем документе (рис. 1).

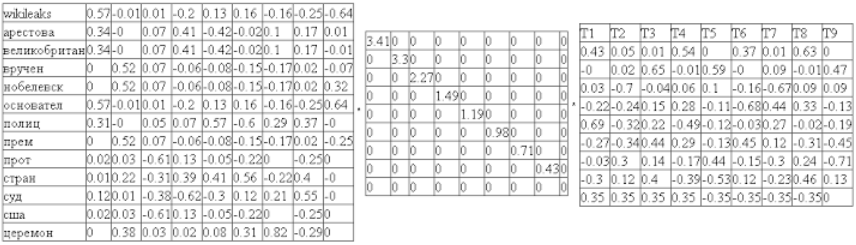
	T1	T2	T3	T4	T5	T6	T7	T8	T9
wikileaks	1	0	0	1	0	1	0	1	0
арестова	0	0	0	1	0	0	0	1	0
великобритан	0	0	0	1	0	0	0	1	0
вручен	0	0	1	0	1	0	0	0	1
нобелевск	0	0	1	0	1	0	0	0	1
основател	1	0	0	1	0	1	0	1	0
полиц	1	0	0	0	0	0	0	1	0
прем	0	0	1	0	1	0	0	0	1
прот	0	1	0	0	0	0	1	0	0
стран	0	0	1	0	0	0	1	0	0
суд	0	1	0	0	0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

Рисунок 1. Частотная матрица индексируемых слов

Следующим шагом необходимо провести сингулярное разложение полученной матрицы (рис. 2). Т.е. исходную матрицу M мы представляем в виде:

M = U*W*Vt

где U и Vt – ортогональные матрицы, а W – диагональная матрица. Причем диагональные элементы матрицы W упорядочены в порядке убывания. Диагональные элементы матрицы W называются сингулярными числами.



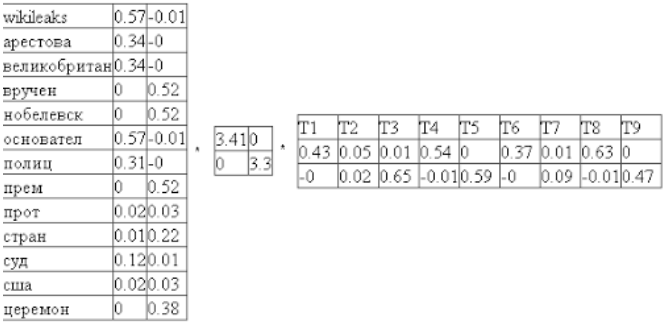
wikileaks	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	-0.64
арестова	0.34	-0	0.07	0.41	-0.42	-0.02	0.1	0.17	0.01
великобритан	0.34	-0	0.07	0.41	-0.42	-0.02	0.1	0.17	-0.01
вручен	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.07
нобелевск	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	0.32
основател	0.57	-0.01	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
0	3.3	0	0	0	0	0	0	0	0
0	0	2.27	0	0	0	0	0	0	0
0	0	0	1.49	0	0	0	0	0	0
0	0	0	0	1.19	0	0	0	0	0
0	0	0	0	0	0.98	0	0	0	0
0	0	0	0	0	0	0.71	0	0	0
0	0	0	0	0	0	0	0.43	0	0
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	

Рисунок 2. Сингулярное разложение частотной матрицы индексируемых слов

Согласно простым правилам произведения матриц, видно, что столбцы и строки соответствующие меньшим сингулярным значениям дают наименьший вклад в итоговое произведение. Например, мы можем отбросить последние столбцы матрицы U и последние строки матрицы V^t, оставив только первые 2. Важно, что при этом гарантируется, оптимальность полученного произведения. Разложение такого вида называют двумерным сингулярным разложением (рис. 3):



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

Рисунок 3. Двумерное сингулярное разложение

На практике, конечно, количество групп будет намного больше, пространство будет не двумерным а многомерным, но сама идея остается той же. Можно определять местоположения слов и статей в нашем пространстве и использовать эту информацию для, например, определения тематики статьи.

В полученном факторном пространстве документы и термины концентрируются областями, имеющими общий семантический и латентный смысл.

Применительно к кластеризации, получаемые области и есть кластеры. С помощью математических преобразований можно определить центры кластеров.

Задание к лабораторной работе

Написать программу на выбранном языке программирования, реализующую описанный выше алгоритм для кластеризации заголовков по темам. Количество заголовков задать не более 10 (оптимально от 7 до 10), количество кластеров не более 3. Программа должна запрашивать заголовки (они могут храниться в файлах). После получения двумерного сингулярного разложения осуществить прямую кластеризацию самостоятельно выбранным и изученным методом (иерархические алгоритмы, поиск k-средних и т.д.). Результатом работы программы должен быть файл, содержащий заголовки, разбитые по кластерам.

Контрольные вопросы

Приложение 1

Стоп-символы русского языка

Стоп символы, они же стоп-слова, это слова, встречающиеся практически во всех текстах и не несущие специальной смысловой нагрузки. В русском языке, к стоп символам относятся предлоги, суффиксы, причастия, междометия, частицы и т.п. Неполный список стоп-слов представлен ниже:

-	еще	него	сказать
а	ж	нее	со
без	же	ней	совсем
более	жизнь	нельзя	так
больше	за	нет	такой
будет	зачем	ни	там
будто	здесь	нибудь	тебя
бы	и	никогда	тем
был	из	ним	теперь
была	из-за	них	то
были	или	ничего	тогда
было	им	но	того
быть	иногда	ну	тоже
в	их	о	только
вам	к	об	том
вас	кажется	один	тот
вдруг	как	он	три
ведь	какая	она	тут
во	какой	они	ты
вот	когда	опять	у
впрочем	конечно	от	уж
все	которого	перед	уже
всегда	которые	по	хорошо
всего	кто	под	хоть
всех	куда	после	чего
всю	ли	потом	человек
вы	лучше	потому	чем
г	между	почти	через
где	меня	при	что
говорил	мне	про	чтоб
да	много	раз	чтобы
даже	может	разве	чуть
два	можно	с	эти
для	мой	сам	этого
до	моя	свое	этой

другой	мы	свою	этом
его	на	себе	этот
ее	над	себя	эту
ей	надо	сегодня	я
ему	наконец	сейчас	
если	нас	сказал	
есть	не	сказала	

Помимо указанных слов имеет смысл еще фильтровать цифры, отдельные буквы и знаки препинания.