Анализ данных в бизнесе

Текстовая аналитика: семинар

Алексей Пятов // Руководитель группы текстовой аналитики SAS Russia
Константин Дудников // Эксперт по текстовой аналитике SAS Russia

210130

План

1. Тематическое моделирование на пальцах
2. Практикум по тематическому моделированию на SAS Viya

Тематическое моделирование

На пальцах

Тематическое моделирование

Примеры бизнес-задач

На что в основном жалуются пользователи?
За что проводят и получают платежи клиенты банка?

Тематическое моделирование

Определимся в терминах

Что такое текст и что такое тема?

Тематическое моделирование

Определимся в терминах

Что такое текст и что такое тема?

Текст – набор слов.
Тема – набор ключевых слов.

Тематическое моделирование

Разберем на примере

Что такое текст и что такое тема?

Текст – набор слов.
Тема – набор ключевых слов.

1. Британская полиция знает о местонахождении основателя WikiLeaks.
2. В суде США начинается процесс против россиянина, рассылавшего спам.
3. Церемонию вручения Нобелевской премии мира бойкотируют 19 стран.
4. В Великобритании арестован основатель сайта Wikileaks Джулиан Ассандж.
5. Украина игнорирует церемонию вручения Нобелевской премии.
6. Шведский суд отказался рассматривать апелляцию основателя Wikileaks.
7. НАТО и США разработали планы обороны стран Балтии против России.
8. Полиция Великобритании нашла основателя WikiLeaks, но не арестовала.
9. В Стокгольме и Осло сегодня состоится вручение Нобелевских премий.

Тематическое моделирование

Подготовка корпуса

Токенизация – делаем из строки набор слов.

1. [британская], [полиция], [знает], [о], [местонахождении], [основателя], [wikileaks]
2. [в], [суде], [сша], [начинается], [процесс], [против], [россиянина], [рассылавшего], [спам]
3. [церемонию], [вручения], [нобелевской], [премии], [мира], [бойкотируют], [19], [стран]
4. [в], [великобритании], [арестован], [основатель], [сайта], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорирует], [церемонию], [вручения], [нобелевской], [премии]
6. [шведский], [суд], [отказался], [рассматривать], [апелляцию], [основателя], [wikileaks]
7. [нато], [и], [сша], [разработали], [планы], [обороны], [стран], [балтии], [против], [россии]
8. [полиция], [великобритании], [нашла], [основателя], [wikileaks], [но], [не], [арестовала]
9. [в], [стокгольме], [и], [осло], [сегодня], [состоится], [вручение], [нобелевских], [премий]

Тематическое моделирование

Подготовка корпуса

Лемматизация – делаем количество слов в корпусе меньше.

1. [британская], [полиция], [знает], [о], [местонахождении], [основателя], [wikileaks]
2. [в], [суде], [сша], [начинается], [процесс], [против], [россиянина], [рассылавшего], [спам]
3. [церемонию], [вручения], [нобелевской], [премии], [мира], [бойкотируют], [19], [стран]
4. [в], [великобритании], [арестован], [основатель], [сайта], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорирует], [церемонию], [вручения], [нобелевской], [премии]
6. [шведский], [суд], [отказался], [рассматривать], [апелляцию], [основателя], [wikileaks]
7. [нато], [и], [сша], [разработали], [планы], [обороны], [стран], [балтии], [против], [россии]
8. [полиция], [великобритании], [нашла], [основателя], [wikileaks], [но], [не], [арестовала]
9. [в], [стокгольме], [и], [осло], [сегодня], [состоится], [вручение], [нобелевских], [премий]

Тематическое моделирование

Подготовка корпуса

Лемматизация – делаем количество слов в корпусе меньше.

1. [британский], [полиция], [знать], [о], [местонахождение], [основатель], [wikileaks]
2. [в], [суд], [сша], [начинаться], [процесс], [против], [россиянин], [рассылавший], [спам]
3. [церемония], [вручение], [нобелевский], [премия], [мир], [бойкотировать], [19], [страна]
4. [в], [великобритания], [арестованный], [основатель], [сайт], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорировать], [церемония], [вручение], [нобелевский], [премия]
6. [шведский], [суд], [отказаться], [рассматривать], [апелляция], [основатель], [wikileaks]
7. [нато], [и], [сша], [разработать], [план], [оборона], [страна], [балтия], [против], [россия]
8. [полиция], [великобритания], [найти], [основатель], [wikileaks], [но], [не], [арестовать]
9. [в], [стокгольм], [и], [осло], [сегодня], [состояться], [вручение], [нобелевский], [премия]

Тематическое моделирование

Подготовка корпуса

Удаление стоп-слов – никакого смысла не несут.

1. [британский], [полиция], [знать], [о], [местонахождение], [основатель], [wikileaks]
2. [в], [суд], [сша], [начинаться], [процесс], [против], [россиянин], [рассылавший], [спам]
3. [церемония], [вручение], [нобелевский], [премия], [мир], [бойкотировать], [19], [страна]
4. [в], [великобритания], [арестованный], [основатель], [сайт], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорировать], [церемония], [вручение], [нобелевский], [премия]
6. [шведский], [суд], [отказаться], [рассматривать], [апелляция], [основатель], [wikileaks]
7. [нато], [и], [сша], [разработать], [план], [оборона], [страна], [балтия], [против], [россия]
8. [полиция], [великобритания], [найти], [основатель], [wikileaks], [но], [не], [арестовать]
9. [в], [стокгольм], [и], [осло], [сегодня], [состояться], [вручение], [нобелевский], [премия]

Тематическое моделирование

Подготовка корпуса

Удаление стоп-слов – никакого смысла не несут.

1. [британский], [полиция], [знать], [местонахождение], [основатель], [wikileaks]
2. [суд], [сша], [начинаться], [процесс], [россиянин], [рассылавший], [спам]
3. [церемония], [вручение], [нобелевский], [премия], [мир], [бойкотировать], [страна]
4. [великобритания], [арестованный], [основатель], [сайт], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорировать], [церемония], [вручение], [нобелевский], [премия]
6. [шведский], [суд], [отказаться], [рассматривать], [апелляция], [основатель], [wikileaks]
7. [нато], [сша], [разработать], [план], [оборона], [страна], [балтия], [россия]
8. [полиция], [великобритания], [найти], [основатель], [wikileaks], [арестовать]
9. [стокгольм], [осло], [состояться], [вручение], [нобелевский], [премия]

Тематическое моделирование

Подготовка корпуса

Удаление редких слов – либо опечатки, либо не помогают описать тему.

1. [британский], [полиция], [знать], [местонахождение], [основатель], [wikileaks]
2. [суд], [сша], [начинаться], [процесс], [россиянин], [рассылавший], [спам]
3. [церемония], [вручение], [нобелевский], [премия], [мир], [бойкотировать], [страна]
4. [великобритания], [арестованный], [основатель], [сайт], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорировать], [церемония], [вручение], [нобелевский], [премия]
6. [шведский], [суд], [отказаться], [рассматривать], [апелляция], [основатель], [wikileaks]
7. [нато], [сша], [разработать], [план], [оборона], [страна], [балтия], [россия]
8. [полиция], [великобритания], [найти], [основатель], [wikileaks], [арестовать]
9. [стокгольм], [осло], [состояться], [вручение], [нобелевский], [премия]

Тематическое моделирование

Латентно-семантический анализ

Строим терм-документную матрицу

	d1	d2	d3	d4	d5	d6	d7	d8	d9
апелляция	0	0	0	0	0	1	0	0	0
арестованный	0	0	0	1	0	0	0	0	0
арестовать	0	0	0	0	0	0	0	1	0
балтия	0	0	0	0	0	0	1	0	0
бойкотировать	0	0	1	0	0	0	0	0	0
британский	1	0	0	0	0	0	0	0	0
великобритания	0	0	0	1	0	0	0	1	0
вручение	0	0	1	0	1	0	0	0	1
…									

Тематическое моделирование

Латентно-семантический анализ

Взвешиваем термы:

- tf-idf
- Mutual information (MI)
- Entropy

	d1	d2	d3	d4	d5	d6	d7	d8	d9
апелляция	0	0	0	0	0	0,46	0	0	0
арестованный	0	0	0	0,18	0	0	0	0	0
арестовать	0	0	0	0	0	0	0	0,61	0
балтия	0	0	0	0	0	0	0,38	0	0
бойкотировать	0	0	0,37	0	0	0	0	0	0
британский	0,19	0	0	0	0	0	0	0	0
великобритания	0	0	0	0,18	0	0	0	0,23	0
вручение	0	0	0,26	0	0,38	0	0	0	0,52
…									

Тематическое моделирование

Латентно-семантический анализ

Выбираем количество тем и выполняем сингулярное разложение матрицы.

апелляция	0,57	-0,01	0,01
арестованный	0,34	0	0,07
арестовать	0,34	0	0,01
балтия	0	0,12	-0,62
бойкотировать	0	0,52	-0,21
британский	0,57	-0,01	-0,18
великобритания	0,31	0,05	0,15
вручение	0,02	0,67	0,09
…			

матрица U

3,41	0	0
0	3,3	0
0	0	2,27

матрица Σ

d1	d2	d3	d4	d5	d6	d7	d8	d9
0,63	0,05	0,01	0,54	0	0,47	0,01	0,63	0
0	0,02	0,65	-0,01	0,59	0	0,09	-0,01	0,48
0,03	-0,7	-0,04	0,06	0,1	-0,16	-0,67	0,09	0,09

матрица V^T

Тематическое моделирование

Латентно-семантический анализ

Интерпретируем результаты

	Темы		
	1	2	3
апелляция	0,57	-0,01	0,01
арестованный	0,34	0	0,07
арестовать	0,34	0	0,01
балтия	0	0,12	-0,62
бойкотировать	0	0,52	-0,21
британский	0,57	-0,01	-0,18
великобритания	0,31	0,05	0,15
вручение	0,02	0,67	0,09
…			

матрица U

3,41	0	0
0	3,3	0
0	0	2,27

матрица Σ

d1	d2	d3	d4	d5	d6	d7	d8	d9	
0,63	0,05	0,01	0,54	0	0,47	0,01	0,63	0	Тема 1
0	0,02	0,65	-0,01	0,59	0	0,09	-0,01	0,48	Тема 2
0,03	-0,7	-0,04	0,06	0,1	-0,16	-0,67	0,09	0,09	Тема 3

матрица V^T

Тематическое моделирование

Латентно-семантический анализ

Проверяем

	d1	d2	d3	d4	d5	d6	d7	d8	d9
апелляция, арестованный, арестовать, британский, великобритания	1	0	0	1	0	1	0	1	0
бойкотировать, вручение	0	0	1	0	1	0	0	0	1
балтия, США	0	1	0	0	0	0	1	0	0

1. Британская полиция знает о местонахождении основателя WikiLeaks.
2. В суде США начинается процесс против россиянина, рассылавшего спам.
3. Церемонию вручения Нобелевской премии мира бойкотируют 19 стран.
4. В Великобритании арестован основатель сайта Wikileaks Джулиан Ассандж.
5. Украина игнорирует церемонию вручения Нобелевской премии.
6. Шведский суд отказался рассматривать апелляцию основателя Wikileaks.
7. НАТО и США разработали планы обороны стран Балтии против России.
8. Полиция Великобритании нашла основателя WikiLeaks, но не арестовала.
9. В Стокгольме и Осло сегодня состоится вручение Нобелевских премий.

Тематическое моделирование

Латентное размещение Дирихле

А что если распределение тем по документам и слов по темам смоделировать распределением Дирихле?..

Получится смоделировать много тем с условием: в каждом документе не более 3-5 тем.

- Немного о работе LDA на пальцах
- Чуть больше и с примерами кода – здесь есть ссылки на подробные статьи с формулами

Тематическое моделирование

Библиотеки

gensim – topic modelling for humans.
sklearn – topic modelling with LDA.

Практикум на SAS Viya

От сырых данных до разметки за 10 минут