Topic Modeling Report

Шер Артём Владимирович

4 января 2023

1 Информация о разбиении данных по числу документов на партиции

Разбиение на 2 партиции: [8038, 7991].
Разбиение на 5 партиций: [3206, 3206, 3207, 3205, 3205].

2 Эксперименты с количеством тем

Результаты экспериментов с $num\_topics = 10, 20, 50$ представлены в таблице 1.

Таблица 1: Результаты экспериментов по времени с разным количеством тем. (sec)

Broadcast state	**Num Topics**		
	10	**20**	**50**
False	1200	1953	5161
True	1464	2461	4760

Можно сделать вывод о том, что использование broadcast не ускоряет алгоритм для маленького количества тем, но возможно есть небольшой прирост когда тем 50.

3 Эксперименты с количеством проходов по документу

Результаты экспериментов с $num\_document\_passes = 1, 2, 5, 10$ представлены в таблице 2.

Таблица 2: Результаты экспериментов с разным количеством проходов по документу.

	Num Document Passes			
	1	**2**	**5**	**10**
Time (sec)	693	1095	1945	3609
Perplexity	11588	10531	5789	5153
Words in theme variety	album	king	church	king
	club	emperor	saint	roman
	station	film	town	father
	game	empire	village	cause
	king	roman	french	opera
	system	album	trust	mother
	president	china	community	son
	park	province	paris	man
	show	show	institute	empire
	church	award	museum	daughter

Можно сделать ожидаемый вывод о том, что время работы увеличивается, а перплексия снижается с увеличением числа проходов по документу. Что касается интерпретируемости тем, то для 1 и 2 проходов

её оказывается недостаточно, хотя для 2 проходов уже можно сделать вывод об одной теме, но бывают выбросы. Разница между 5 и 10 проходами не существенна, поэтому оптимальным числом проходов будет 5.

4 Эксперименты с параметром регуляризации

Результаты экспериментов с $beta = 0.0, -0.1, -1.0, -25.0$ представлены в таблице 3.

Таблица 3: Результаты экспериментов с различным параметром реугляризации.

	Beta			
	0.0	**-0.1**	**-1.0**	**-25.0**
Sparsity (%)	92.0	91.9	91.6	98.7
Perplexity	5722	5747	6443	5e15

Можно сделать вывод о том, что коэффициент $beta = -1.0$ увеличивает разреженность матрицы на 1%. Перплексия увеличивается с увеличением параметра beta, как и разреженность.