Классификация тематик полученных на основе тематической модели с последовательной регуляризацией для коллекции научных статей из библиотеки OnePetro

Федор Краснов

Эксперт Блока научного Инжиниринга,

Газпромнефть Научно-технический центр.

Ведение

Тематическое моделирование - одно из современных направлений статистической *обработки естественного языка.*

Вероятностная тематическая модель коллекции документов описывает каждую тему дискретным распределением вероятностей слов, а каждый документ – дискретным распределением вероятностей тем.

При таком подходе можно построить бесконечное количество различных тематических моделей для коллекции документов.

На практике для задач *Text Mining* нужны «хорошие» тематические модели:

Темы должны быть однородными и нести в себе уникальный смысл.

Теория

Существует несколько десятков различных методик для построения тематических моделей. Наиболее популярной является методика построения на основе распределения Дирихле - Latent Dirichlet Allocation (LDA), предложенная Дэвидом Блеем в 2003.

$$p(d,w) = \sum_{\{t \in T\}} p(d)p(w|t)p(t|d),$$

T – множество тем, p(t) – неизвестное распределение тем во всей коллекции, p(d) – априорное распределение документов (оценка $^{n_d}/_n$), p(w) – априорное распределение на множестве слов ($^{n_w}/_n$).

$\theta_d = (p(t|d): t \in T)$ - векторы документов, $|T| \sim Dir(\theta, \alpha), \alpha \in \mathbb{R}^{|T|}$

$\phi_t = (p(w|t): w \in W)$ - векторы тем, $|W| \sim Dir(\phi, \beta), \beta \in \mathbb{R}^{|W|}$

Практическая проблема

С помощью методики LDA получаются непригодные для использования темы. Примеры:

Плохая Тема 1:

(нефть:0.0051, или:0.082, газ:0.0005, но:0.0003, если:0.0004, …)

Плохая Тема 2:

(карбонаты:0.0001, скважина:0.0001, Россия:0.0001, Запад:0.0001, …)

Хорошая Тема 3:

(геофизика:0.003, сейсмика:0.004, s-волна:0.002, импеданс:0.001, …)

Аддитивная регуляризация (ARTM)

ARTM делает темы более информативными при помощи добавления на этапе обучения к функции правдоподобия регуляризаторов вида:

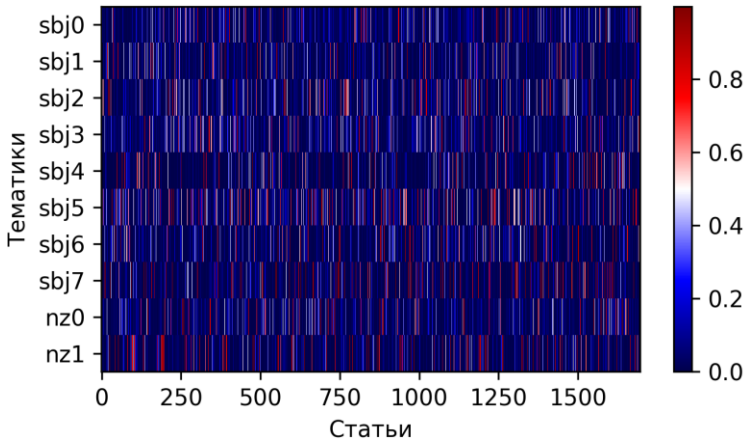
$$R(\Phi,\Theta) = \sum_{t\in \mathrm{T}} \sum_{w\in W} \beta_{wt} ln(\phi_{wt}) + \sum_{d\in \mathrm{D}} \sum_{t\in T} \alpha_{td} \ln(\theta_{td})$$

Данный подход управления качеством тематической модели предложен проф. Воронцовым К.В.

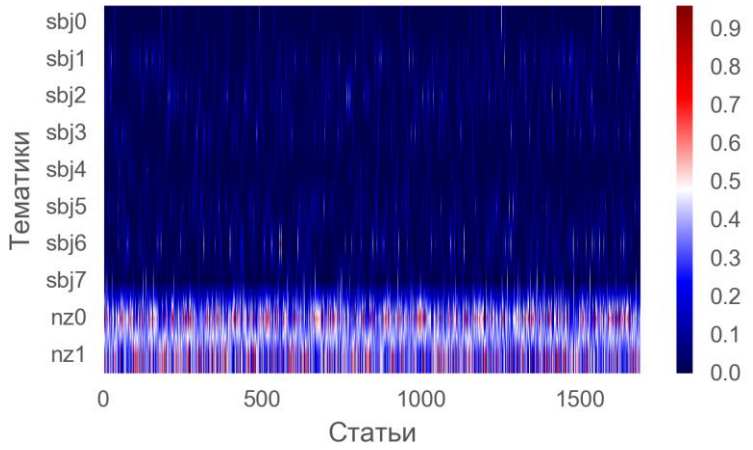
ARTM позволяет выделять предметные и фоновые темы.

Практические результаты 1

LDA



ARTM



Тема	sbj0	sbj1	sbj2	sbj3	sbj4	sbj5	sbj6	sbj7
Терм1	liquid	shale	fracturing	injection	corrosion	casing	recovery	safety
Терм2	pipeline	porosity	proppant	fractures	nace	cement	injection	management
Терм3	pipe	logging	hydraulic	shale	concentration	mud	steam	risk
Терм4	velocity	sand	stress	matrix	samples	hole	core	human
Терм5	multiphase	pore	fractures	hydraulic	inhibitor	mpd	viscosity	health
Терм6	slug	samples	stage	recovery	acid	bit	flooding	business
Терм7	friction	core	shale	fractured	ph	drill	solvent	assessment
Терм8	bhr	spwla	treatment	bakken	steel	string	heavy	training
Терм9	group	clay	conductivity	porosity	houston	pipe	saturation	company
Терм10	holdup	symposium	stages	unconventional	iron	liner	surfactant	activities

Практические результаты 2

Фрагмент исследования:

Например, документ №555 обладает самым большим весом тематики *0.72* (**sbj6**). Вероятности других основных тематик для этого документа равны нулю. Таким образом этот документ согласно модели, полностью посвящен тематике **sbj6**, представленной словами:

recovery, injection, steam, core, viscosity, flooding, solvent, heavy, saturation, surfactant

При помощи эксперта тематике sbj6 дано название:

«*Chemical enhanced oil recovery*»

Заключение

- Методика ARTM позволяет выделять предметные и фоновые темы.
- Библиотека *BigARTM (http://bigartm.org/)* позволяет выстраивать последовательно несколько регуляризаторов и управлять группами тематик.
- Широко используемые методы построения тематических моделей на основе LDA не дают таких возможностей.