Относительная перплексия как мера качества тематических моделей

Нижибицкий Евгений Алексеевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: nizhibitsky@cs.msu.ru

Тематическое моделирование — это способ построения модели коллекции текстовых документов, при котором каждая тема описывается дискретным распределением на множестве терминов, а каждый документ — дискретным распределением на множестве тем.

Пусть для каждого документа d из коллекции задано число n_{dw} вхождений слова w в d. Тематическая модель описывает вероятность появления слов, опираясь на гипотезу условной независимости $p(w|t) = p(w|d,t)$ и формулу полной вероятности:

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

Для нахождения распределений $p(t|d)$ и $p(w|t)$ по исходным данным (n_{dw}) будем использовать модель LDA [2].

Существует несколько способов оценки качества построенной модели. Наиболее распространённым критерием является перплексия, равная экспоненте от минус усреднённого логарифма правдоподобия:

$$P = \exp\left(-\frac{1}{n}\sum_{d \in D}\sum_{w \in d} n_{dw} \ln p(w|d)\right),$$

где n — длина коллекции в словах. Перплексия зависит от мощности словаря и распределения частот слов в коллекции $p(w) = n_w/n$. Поэтому с её помощью невозможно оценивать качество удаления стоп-слов и нетематических слов, сравнивать методы разреживания словаря, а также униграммные и n-граммные модели.

Данная работа направлена на поиск критерия, также основанного на значении правдоподобия, но нечувствительного к изменению состава словаря. Предлагается *относительная перплексия*, принимающая значения из отрезка $[0, 1]$ (чем меньше, тем лучше):

$$RP = \frac{P - P_{\min}}{P_{\max} - P_{\min}},$$

где $P_{\min}$ — минимальная перплексия униграммной модели докумен-

тов ($p(w|d) = n_{dw}/n_d$), а $P_{\max}$ — максимальная перплексия униграммной модели коллекции ($p(w|d) = n_w/n$, где n_w — число вхождений слова w во всех документах коллекции, n_d — длина документа d). Относительная перплексия уменьшается с ростом числа тем $|T|$, достигая 0 при $T = \min\{W, D\}$, когда тематическая модель вырождается в униграммную модель документа, и 1 при $T = 1$, когда она вырождается в униграммную модель коллекции. Таким образом, относительная перплексия показывает положение модели относительно наилучшего и наихудшего достижимых значений перплексии.

В работе исследуется зависимость относительной перплексии от мощности словаря и числа тем. Для экспериментов использовалась коллекция статей научной конференции NIPS за 1987–1999 гг. на английском языке. В каждом эксперименте при фиксированном числе тем из начального словаря коллекции отбрасывалась его случайно выбранная десятая часть, до полного исчерпания словаря. После каждого отбрасывания производилось обучение модели с помощью библиотеки `gensim` [3]. Полученные модели оценивались с помощью перплексии и относительной перплексии. На Рис. 1 каждой линии соответствует один такой эксперимент, начертания линий соответствуют различному числу тем.

Из правого графика можно сделать вывод, что относительная перплексия слабо зависит от мощности словаря, лучше характеризует способность модели описывать коллекцию. Её численное значение показывает, насколько точность модели близка к предельно достижимому минимуму перплексии.

Можно предполагать, что в коллекции существуют *основные темы*, существенно превышающие по мощности остальные. Они выявляются даже после 7-кратного разреживания словаря. В данном эксперименте относительная перплексия практически не зависит от разреженности словаря при $|T| = 50$. Поэтому можно предположить, что данная коллекция содержит как раз около 50 основных тем.

При большем числе тем $|T|$ относительная перплексия уменьшается по мере разреживания словаря. Это объясняется тем, что темы не одинаковы по мощности. При случайном разреживании словаря малые темы становятся статистически незначимыми и перестают выявляться.

При меньшем числе тем $|T|$ относительная перплексия увеличивается по мере разреживания словаря. Предположительно, это связано с тем, что тематическая модель вынужденно объединяет основные темы, различия между объединёнными темами становятся

незначимыми, темы сближаются и становятся более похожи на униграммную модель коллекции.

Работа выполнена при поддержке гранта РФФИ 14-07-00965.

Иллюстрации

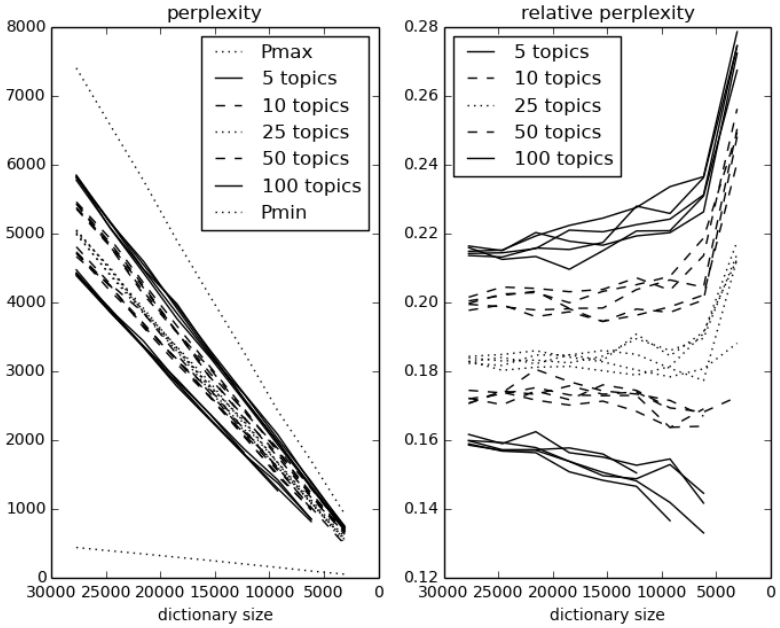

Рис. 1: Изменение функционалов при разреживании словаря.

Литература

1. Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. 2013. Т. 1, № 6. С. 657–686.

2. David Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation // Journal of Machine Learning Research, 2003, P. 993–1022.

3. Radim Řehůřek, Petr Sojka, Software Framework for Topic Modelling with Large Corpora // In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010, P. 45–50, `http://radimrehurek.com/gensim/`.