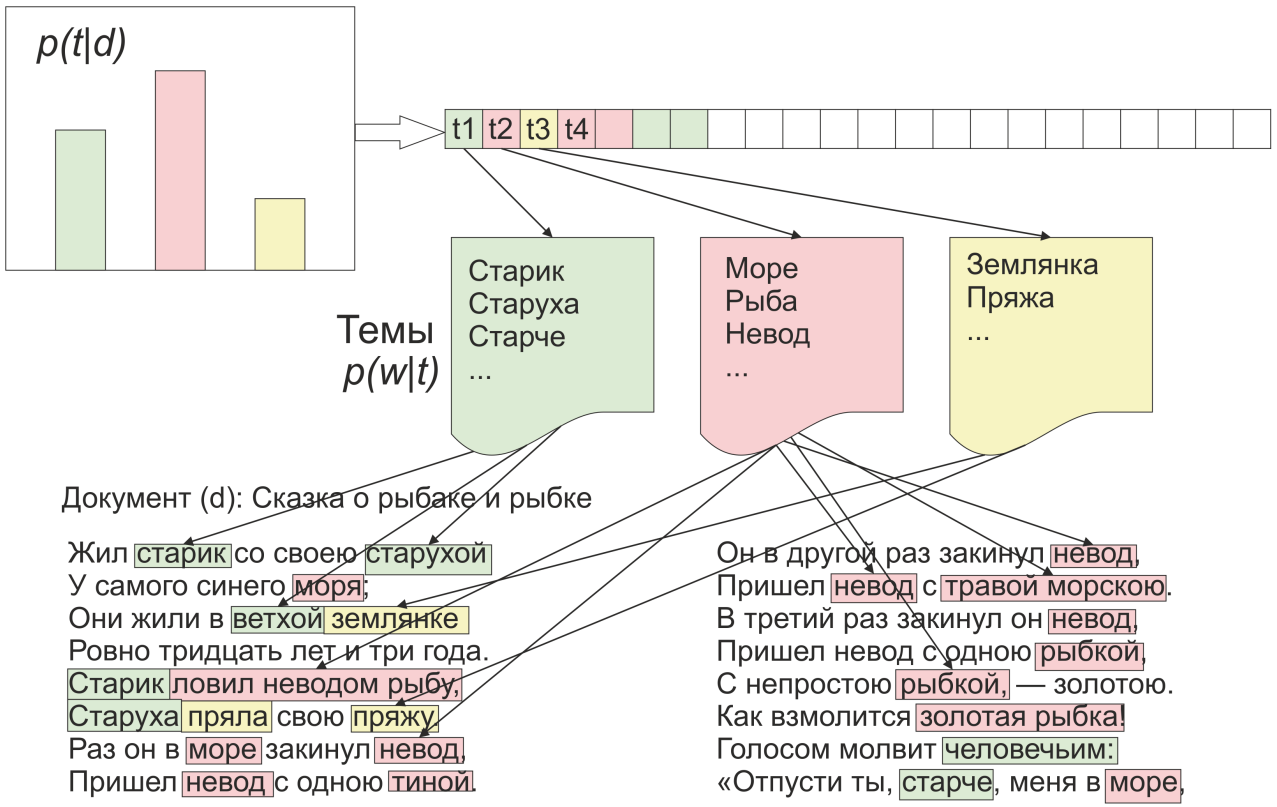
Относительная перплексия как мера качества тематических моделей

Нижибицкий Евгений Алексеевич

Факультет ВМК МГУ имени М. В. Ломоносова

7 апреля 2014 г.



Строим модели коллекции текстовых документов, темы описываются дискретным распределением на множестве терминов, а документы — дискретными распределениями на множестве тем.

- для каждого документа d из коллекции задано число $\mathrm{n}_{\mathrm{d}w}$ вхождений слова w в d.
- опираемся на гипотезу условной независимости $\mathrm{p}(w|\mathrm{t}) = \mathrm{p}(w|\mathrm{d},\mathrm{t})$
- по формуле формулу полной вероятности:

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

- необходимо найти распределения $\mathrm{p}(\mathrm{t}|\mathrm{d})$ и $\mathrm{p}(w|\mathrm{t})$ по исходным данным $(\mathrm{n}_{\mathrm{d}w})$.

Используем модель online LDA [Matthew D. Hoffman, David M. Blei – Online Learning for Latent Dirichlet Allocation, 2010]. Готовая реализация — библиотека gensim под Python (http://radimrehurek.com/gensim/).

Насколько хорошо модель описывает данные:

- правдоподобие L
- перплексия $\mathrm{P} = exp(-\mathrm{L}/\mathrm{N})$
- information rate: $\mathrm{R} = \frac{-\log_2 \mathrm{L}}{\mathrm{N}}$
- критерий Акаике: $\mathrm{AIC} = -2\mathrm{L} + W\mathrm{T}$ и др.

Интерпретируемость тем:

- метод пристального взгляда
- когерентность тем

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 $p(w) = n_w/n$, отсюда получаем ее недостатки:

- невозможно оценивать качество удаления стоп-слов и нетематических слов
- нельзя сравнивать методы разреживания словаря
- нельзя сравнивать униграммные и n-граммные модели.

Необходим критерий, основанный на значении правдоподобия, но нечувствительный к изменению состава словаря. Предлагается *относительная перплексия*, принимающая значения из отрезка $[0, 1]$ (чем меньше, тем лучше):

$$\mathrm{RP} = \frac{\mathrm{P} - \mathrm{P}_{\mathsf{min}}}{\mathrm{P}_{\mathsf{max}} - \mathrm{P}_{\mathsf{min}}},$$

где $\mathrm{P}_{\mathsf{min}}$ — минимальная перплексия униграммной модели документов $(\mathrm{p}(w|\mathrm{d}) = \mathrm{n}_{\mathrm{d}w}/\mathrm{n}_{\mathrm{d}}$), а $\mathrm{P}_{\mathsf{max}}$ — максимальная перплексия униграммной модели коллекции $(\mathrm{p}(w|\mathrm{d}) = \mathrm{n}_w/\mathrm{n}$, где n_w — число вхождений слова w во всех документах коллекции, n_{d} — длина документа $\mathrm{d})$.

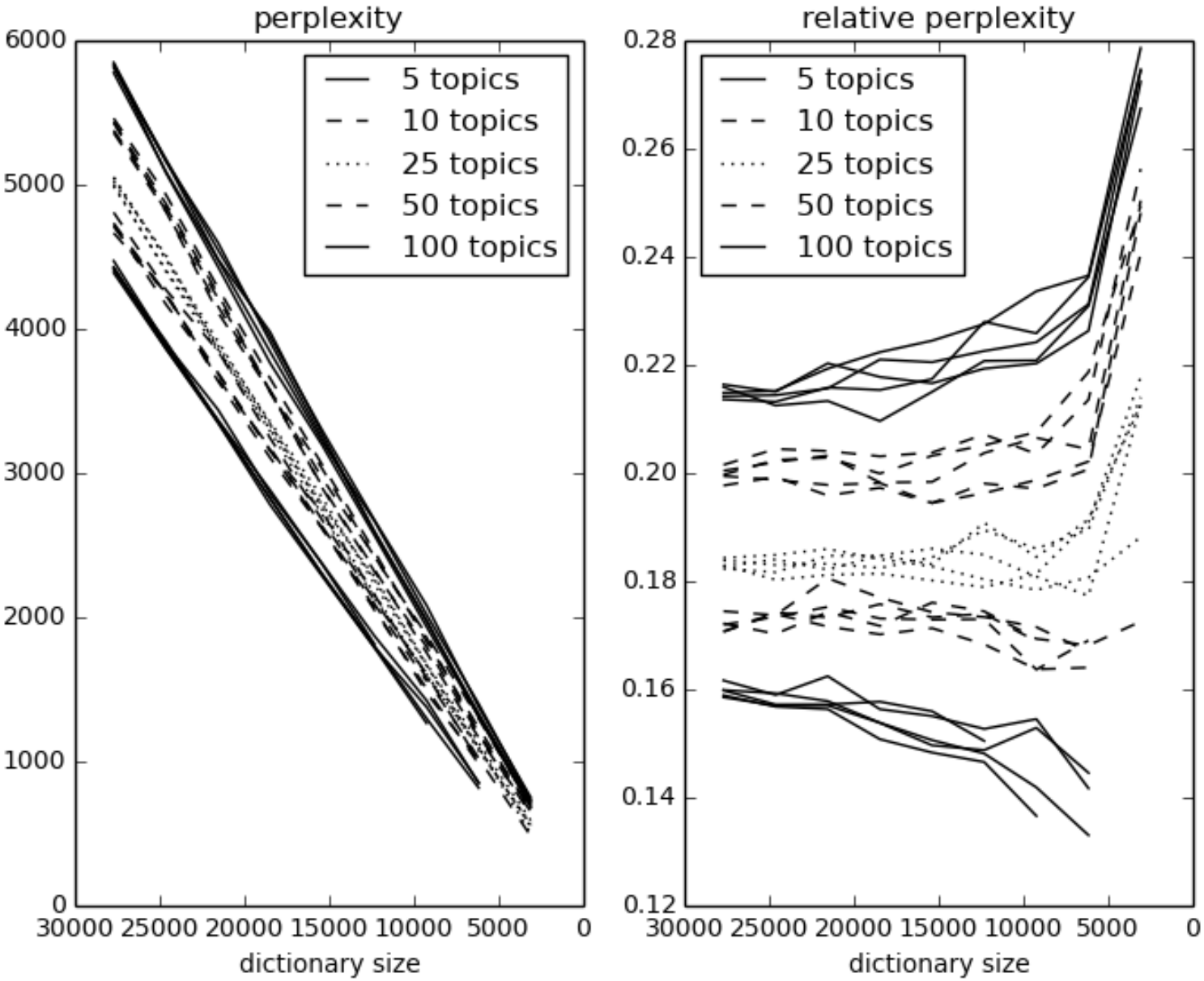
Относительная перплексия уменьшается с ростом числа тем $|\mathrm{T}|$, достигая 0 при $\mathrm{T} = \min\{W, \mathrm{D}\}$, когда тематическая модель вырождается в униграммную модель документа, и 1 при $\mathrm{T} = 1$, когда она вырождается в униграммную модель коллекции.

Данные для экспериментов — коллекция статей научной конференции NIPS за 1987–1999 гг. на английском языке.
В каждом эксперименте

1. при фиксированном числе тем из начального словаря коллекции отбрасывалась его случайно выбранная десятая часть до полного исчерпания словаря;
2. после каждого отбрасывания производилось обучение модели (`gensim`);
3. полученные модели оценивались с помощью перплексии и относительной перплексии.

Эксперименты

Результат



Выводы

- Можно предполагать, что в коллекции существуют *основные темы*, существенно превышающие по мощности остальные. Они выявляются даже после отбрасывания 2/3 словаря.
- 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.

Спасибо за внимание!