Оценка ассоциативной связи слов

Word2Vec, PMI score, Collocations

Вальба О.В, Лайко Р.С

Данные

Значения forward/backward показывают ассоциативную силу (free association response probabilitiy) между исходными словами согласно работе Nelson et al[1]

source	target	forward	backward
ability	able	0.077	0.014
tenor	voice	0.053	0.0
proprietor	sale	0.020	0.0
useless	worthless	0.111	0.088
...	...	...	...

Таблица 1: Пример исходных данных (source: Nelson et al.[1])

[1] Behavior Research Methods, Instruments, Computers 2004, 36 (3), 402–40, The University of South Florida free association, rhyme, and word fragment norms

Оценка ассоциативной близости

1. Получить численное представление пар слов
2. Оценить их ”близость”
3. Вычислить корреляцию с исходной ассоциативностью:
   - Коэфф. Пирсона $\rho_{\mathrm{X},\mathrm{Y}} = \frac{\mathrm{E}[(\mathrm{X}-\mu_{\mathrm{X}})(\mathrm{Y}-\mu_{\mathrm{Y}})]}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}$
   - Коэфф. Кендалла $\tau = \frac{1}{\mathrm{n}(\mathrm{n}-1)} \sum_{\mathrm{i}\neq\mathrm{j}} \mathrm{sgn}(\mathrm{x_i} - \mathrm{x_j})\mathrm{sgn}(\mathrm{y_i} - \mathrm{y_j})$

Оценка ассоциативной близости

- Косинусная близость распределенных векторов слов

$$\cos(\theta) = \frac{< \mathrm{w}_1, \mathrm{w}_2 >}{||\mathrm{w}_1|| \cdot ||\mathrm{w}_2||}$$

- GP-метрика[2][3]

$$\mathrm{GP}(\mathrm{w}_1, \mathrm{w}_2) = \begin{cases} \frac{\mathrm{P}(\mathrm{w}_1, \mathrm{w}_2)}{||\mathrm{w}_2||}, \mathrm{P}(\mathrm{w}_1, \mathrm{w}_2) < ||\mathrm{w}_2|| \\ 1 + \frac{\mathrm{P}(\mathrm{w}_1, \mathrm{w}_2)}{||\mathrm{w}_2||}, \mathrm{P}(\mathrm{w}_1, \mathrm{w}_2) \geq ||\mathrm{w}_2|| \end{cases}$$

- Коэффициент PMI (Pointwise Mutual Information)

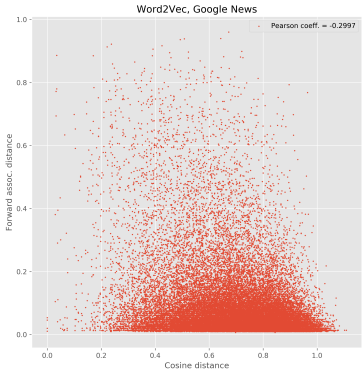
$$\mathrm{pmi}(\mathrm{w}_1; \mathrm{w}_2) = \log \frac{\mathrm{p}(\mathrm{w}_2|\mathrm{w}_1)}{\mathrm{p}(\mathrm{w}_2)}$$

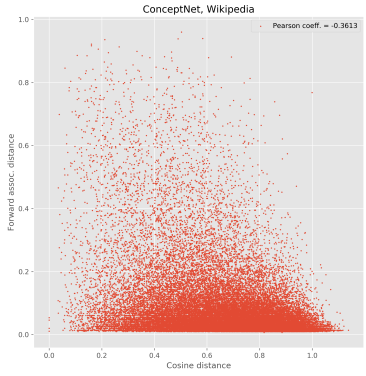
- Вероятность коллокациии n-грамов

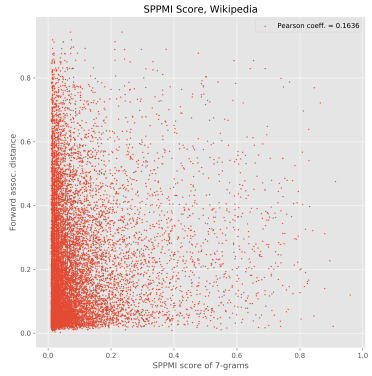
[2] David Galea, Peter Bruza, Deriving Word Association Networks from Text Corpora

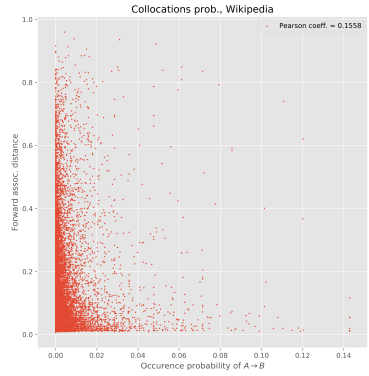
[3] $\mathrm{P}(\mathrm{w}_1, \mathrm{w}_2) = \frac{<\mathrm{w}_1, \mathrm{w}_2>}{||\mathrm{w}_2||}$

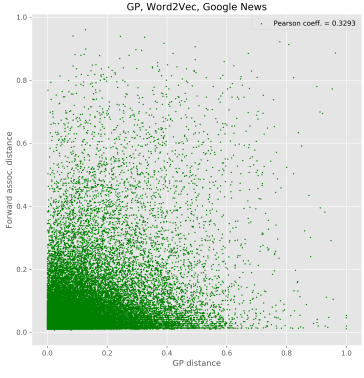
Визуализация результатов

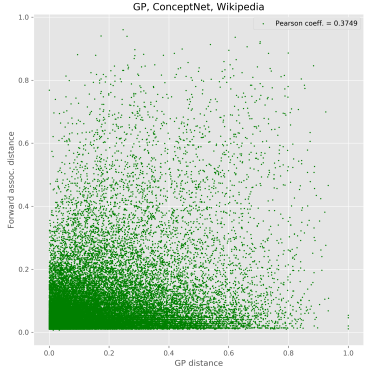


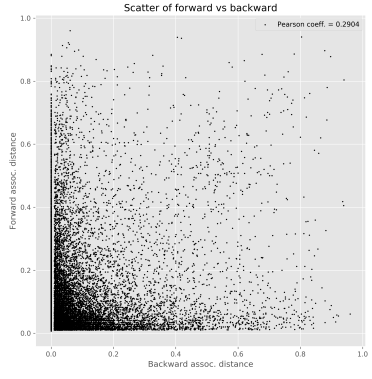


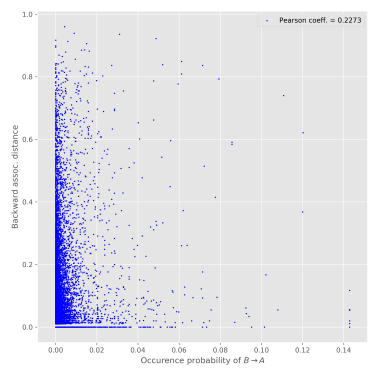












Результаты

Модель	Корр. Пирсона	Корпус
Word2Vec	0.29	Google News
ConceptNet	0.36	Wikipedia + Wiktionary
SPPMI	0.16	Wikipedia (latest dump)
GloVe	0.21	Twitter
fastText	0.28	Wikipedia + IMDB
n-grams occ.$_{\text{forward}}$	0.15	Wikipedia (latest dump)
n-grams occ.$_{\text{backward}}$	0.23	Wikipedia (latest dump)

Пороговые значения

- Выберем порог ассоциативной близости
- Возьмем пары ($\mathrm{source} \to \mathrm{target}$) у которых оно больше этого порога
- Посчитаем корреляцию (Пирсона, Кендалла)

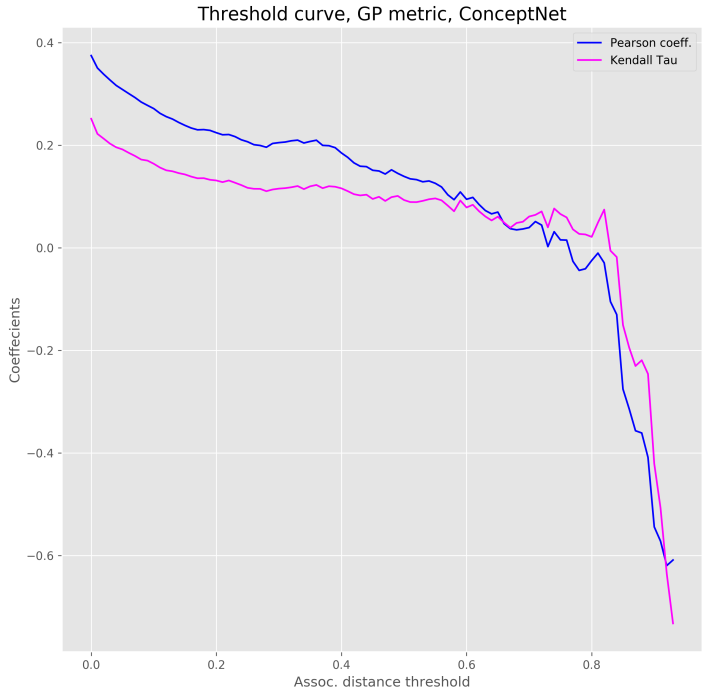


Рис. 1: Зависимость величины корреляции от порогового значения ассоциативной близости в исходном датасете

В перспективе

- Работа с аналогичными датасетами на русском языке
- Распределение Ципфа
- Совмещение сетевых и распределенных подходов