

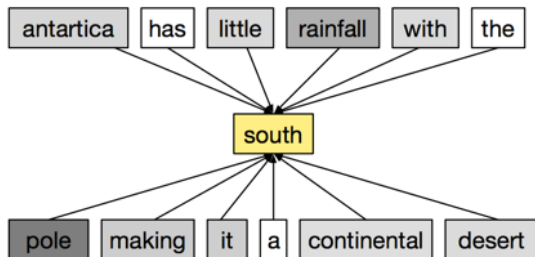
Возможности использования локального контекста в вероятностных тематических моделях

Анна Потапенко

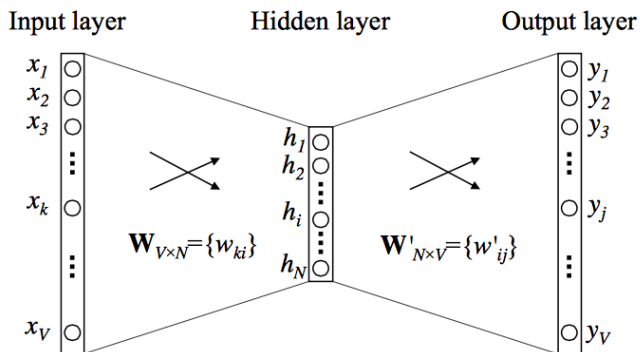
Научный руководитель:
д.ф-м.н. Воронцов К.В.

26 ноября 2015

Задача предсказания контекста по слову



Архитектура нейронной сети word2vec



Mikolov et al. – Distributed Representations of Words and Phrases and their Compositionality, 2013.

Word2vec (Skip-gram model)

Вероятность предсказать слово «O» по слову «I»:

$$p(O|I) = \frac{\exp(\langle w_I, w'_O \rangle)}{\sum_{v=1}^V \exp(\langle w_I, w'_v \rangle)}$$

Функция потерь аддитивна по элементам выборки:

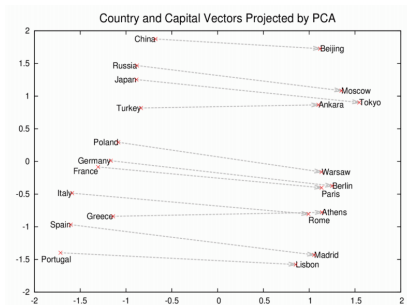
$$E = - \sum_{n=1}^N \log p(O_n|I_n) \rightarrow \min_{W, W'}$$

Эффективные способы подсчета softmax:

- ▶ Hierarchical softmax
- ▶ Negative sampling

Чем полезны обученные векторы (word embeddings):

- ▶ плотные векторы небольшой размерности
- ▶ близость векторов соответствует близости слов:



Completed • Knowledge • 579 teams

Bag of Words Meets Bags of Popcorn

Tue 9 Dec 2014 – Tue 30 Jun 2015 (3 months ago)

Вероятностная тематическая модель

Дано:

D — корпус документов; W — множество слов (словарь);

n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$

Найти:

- ▶ $\varphi_{wt} = p(w|t)$ — распределение слов в темах $t \in T$;
- ▶ $\theta_{td} = p(t|d)$ — распределение тем в документах $d \in D$.

Вероятностная модель порождения документа:

$$p(w|d) = \sum_{t \in T} p(w|t) p(t|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}$$

Задача максимизации правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \log \sum_{t \in T} \varphi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta},$$

Похожи ли модели PLSA и word2vec?

Похожи ли модели PLSA и word2vec?

- ▶ Определим «псевдо-документы»: пусть документ d , порожденный словом w содержит все слова, встречающиеся в окнах ширины $k = 10$ каждого вхождения слова w в коллекцию.
- ▶ Тогда обе модели задаются максимизацией правдоподобия:

$$L = \sum_d \sum_w n_{dw} \log p(w|d) \rightarrow \max$$

- ▶ Но в PLSA: $p(w|d) = \sum_t p(w|t)p(t|d)$,

$$\text{а в word2vec: } p(w|d) = \frac{\exp(\langle w, d \rangle)}{\sum_{d=1}^D \exp(\langle w, d \rangle)}$$

Модель GloVe

Обозначим $p_{ij} = p(w_j | w_i)$.

- ▶ word2vec maximizes cross-entropy:

$$E = - \sum_{i=1}^V \sum_{j=1}^V n_{ij} \log p_{ij} = - \sum_{i=1}^V n_i \sum_{j=1}^V \hat{p}_{ij} \log p_{ij} = \sum_{i=1}^V n_i H(\hat{p}_i, p_i)$$

- ▶ GloVe maximizes L_2 norm:

$$E = \sum_{i,j} f(n_{ij}) (\langle w_i, w'_j \rangle - \log n_{ij})^2,$$

где $f(n_{ij})$ понижает веса для частых слов. Subsampling в word2vec: отбросить слово с вероятностью $1 - \sqrt{\frac{10^{-5}}{n_i}}$

Jeffrey Pennington, Richard Socher, and Christopher D. Manning – GloVe: Global Vectors for Word Representation, 2014.

SGNS: Skip-gram Negative Sampling

Будем предсказывать, встречается ли пара w_j, w_i в корпусе:

$$p = \sigma(\langle w_i, w'_j \rangle) = \frac{1}{1 + \exp(-\langle w_i, w'_j \rangle)}$$

Хотим предсказывать большую вероятность для положительных примеров и маленькую для отрицательных:

$$L = \sum_{i,j} n_{ij} \log \sigma(\langle w_i, w'_j \rangle) + k \mathbb{E}_{p(w_k)} \log \sigma(-\langle w_i, w'_k \rangle),$$

где $p(w_k) = \frac{n_k}{N}$ или $p^{3/4}(w_k)$.

SGNS как разложение shifted PMI-матрицы

Зависимость функционала от одной пары w_j, w_i :

$$l = n_{ij} \log \sigma(\langle w_i, w'_j \rangle) + k n_i \frac{n_j}{N} \log \sigma(-\langle w_i, w'_j \rangle) \rightarrow \max$$

Максимум достигается при:

$$\langle w_i, w'_j \rangle = \log \frac{n_{ij} N}{n_i n_j} - \log k$$

Т.е. SGNS раскладывает shifted PMI-матрицу.

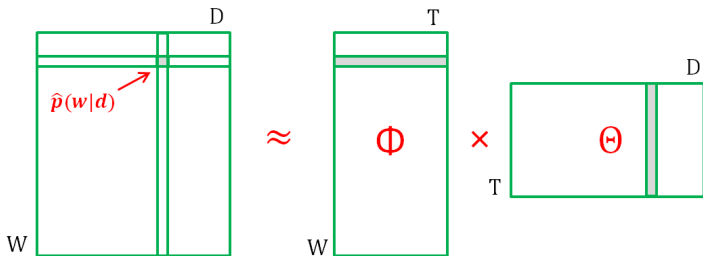
- SVD-разложение для PMI-матрицы, SGNS, GloVe дают похожие результаты при правильном подборе параметров.

Levy et al. – Neural Word Embedding as Implicit Matrix Factorization, 2014;
Improving Distributional Similarity with Lessons Learned from Word
Embeddings, 2015.

Направления работы:

- ▶ Построение тематической модели на «псевдо-документах»:
 - ▶ Как будут отличаться темы от обычной модели?
 - ▶ Вероятностная интерпретация матрицы W' .
 - ▶ Нужна ли регуляризация для такой задачи?
 - ▶ Нужна ли регуляризация в модели SGNS?
- ▶ Включение в модель двух типов документов аналогично мультимодальной тематической модели.
- ▶ Эксперименты на коротких документах: выделение этно-нагруженных тем по сообщениям Вконтакте.
- ▶ Включение rPMI-информации в виде регуляризатора когерентности, эффективная реализация на E-шаге.

Задача построения тематической модели – это задача матричного разложения:



Проблема: матричное разложение неединственно:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

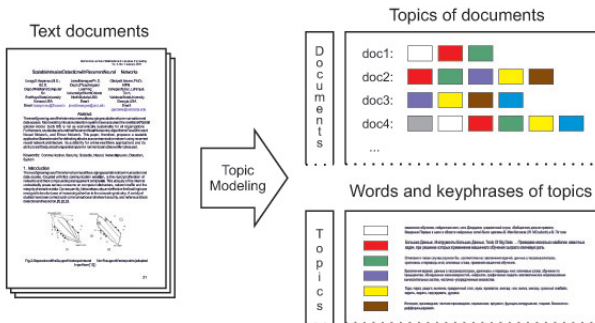
Подход аддитивной регуляризации: учесть дополнительные требования к задаче в виде аддитивных слагаемых.

Направления работы:

- ▶ Построение тематической модели на «псевдо-документах»:
 - ▶ Как будут отличаться темы от обычной модели?
 - ▶ Вероятностная интерпретация матрицы W' .
 - ▶ Нужна ли регуляризация для такой задачи?
 - ▶ Нужна ли регуляризация в модели SGNS?
- ▶ Включение в модель двух типов документов аналогично мультимодальной тематической модели.
- ▶ Эксперименты на коротких документах: выделение этно-нагруженных тем по сообщениям Вконтакте.
- ▶ Включение rPMI-информации в виде регуляризатора когерентности, эффективная реализация на E-шаге.

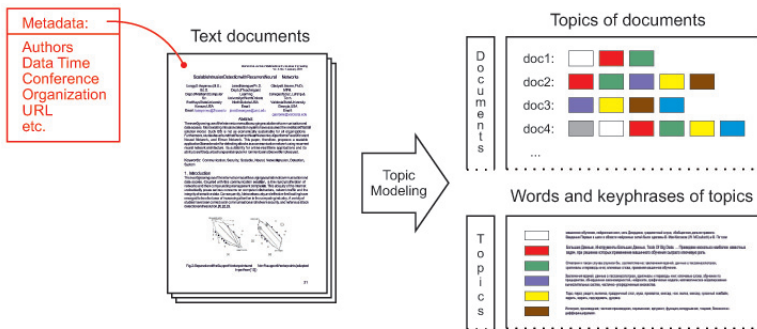
Мультимодальная тематическая модель

Мультимодальная тематическая модель для каждой темы
находит распределение на словах $p(w|t)$,



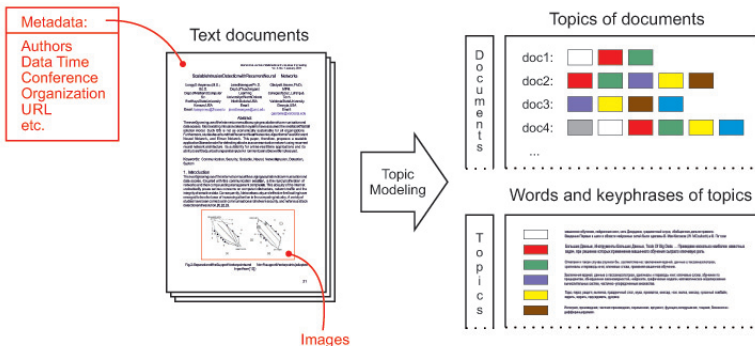
Мультимодальная тематическая модель

Мультимодальная тематическая модель для каждой темы
находит распределение на словах $p(w|t)$, авторах $p(a|t)$,
временных метках $p(y|t)$,



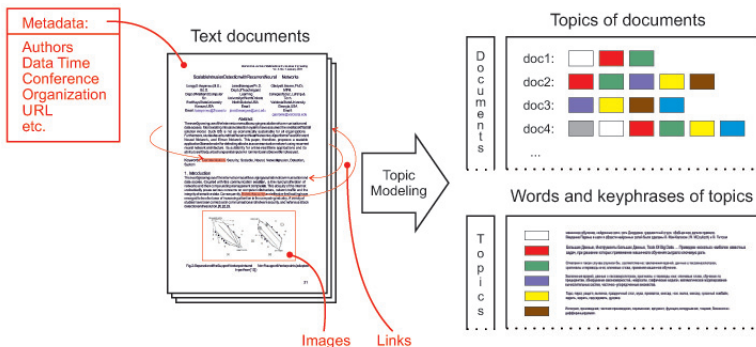
Мультимодальная тематическая модель

Мультимодальная тематическая модель для каждой темы находит распределение на словах $p(w|t)$, авторах $p(a|t)$, временных метках $p(y|t)$, **объектах на изображениях** $p(o|t)$,



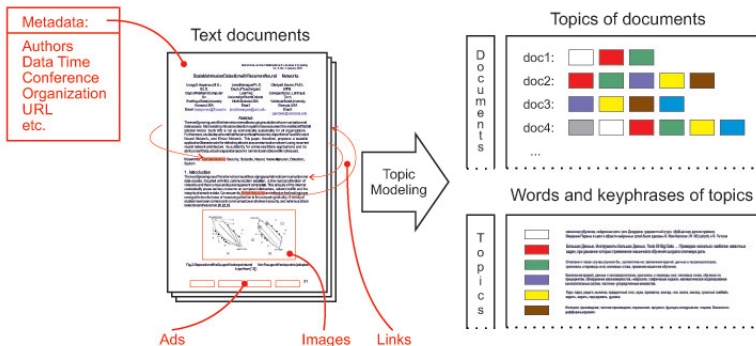
Мультимодальная тематическая модель

emphМультимодальная тематическая модель для каждой темы
находит распределение на словах $p(w|t)$, авторах $p(a|t)$,
временных метках $p(y|t)$, объектах на изображениях $p(o|t)$,
ссылках $p(d'|t)$,



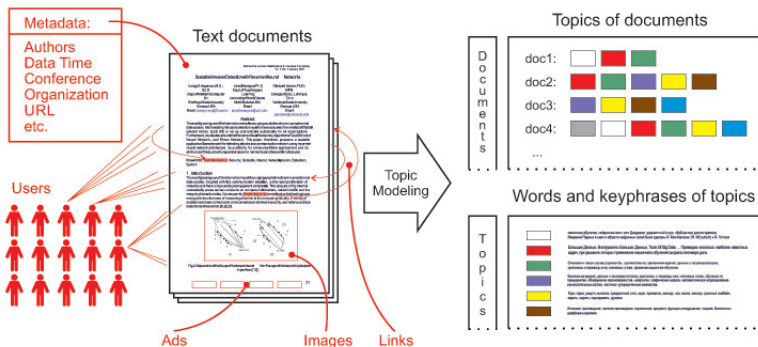
Мультимодальная тематическая модель

Мультимодальная тематическая модель для каждой темы находит распределение на словах $p(w|t)$, авторах $p(a|t)$, временных метках $p(y|t)$, объектах на изображениях $p(o|t)$, ссылках $p(d'|t)$, **рекламных баннерах** $p(b|t)$,



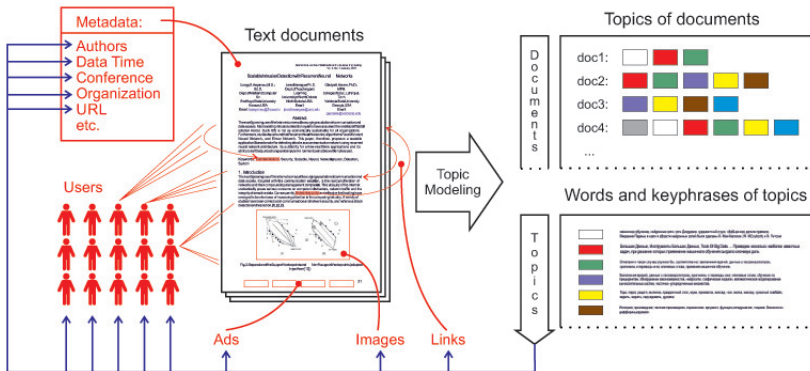
Мультимодальная тематическая модель

Мультимодальная тематическая модель для каждой темы находит распределение на словах $p(w|t)$, авторах $p(a|t)$, временных метках $p(y|t)$, объектах на изображениях $p(o|t)$, ссылках $p(d'|t)$, баннерах $p(b|t)$, **пользователях $p(u|t)$** ,



Мультимодальная тематическая модель

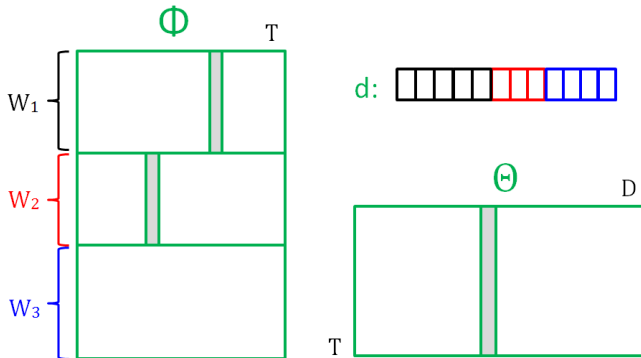
Мультимодальная тематическая модель для каждой темы находит распределение на словах $p(w|t)$, авторах $p(a|t)$, временных метках $p(y|t)$, объектах на изображениях $p(o|t)$, ссылках $p(d'|t)$, баннерах $p(b|t)$, пользователях $p(u|t)$.



Мультимодальная тематическая модель

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \sum_{i=1}^n \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta},$$

где $\lambda_m > 0$, $\tau_i > 0$ – коэффициенты регуляризации.



Направления работы:

- ▶ Построение тематической модели на «псевдо-документах»:
 - ▶ Как будут отличаться темы от обычной модели?
 - ▶ Вероятностная интерпретация матрицы W' .
 - ▶ Нужна ли регуляризация для такой задачи?
 - ▶ Нужна ли регуляризация в модели SGNS?
- ▶ Включение в модель двух типов документов аналогично мультимодальной тематической модели.
- ▶ Эксперименты на коротких документах: выделение этно-нагруженных тем по сообщениям Вконтакте.
- ▶ Включение PMI-информации в виде регуляризатора когерентности, эффективная реализация на E-шаге.

Этно-нагруженные темы

(русские): русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, век, московская, екатерина, москва,

(русские): акция, организация, митинг, движение, активный, мероприятие, совет, русский, участник, москва, оппозиция, россия, пикет, протест, проведение, националист, поддержка, общественный, проводить, участие,

(сирийцы): сирийский, асад, боевик, район, террорист, уничтожать, группировка, дамаск, оружие, алесио, оппозиция, операция, селение, сша, нусра, турция,

(американцы): американский, американка, война, россия, военный, страна, вашингтон, америка, армия, конгресс, сирия, союзный, российский, обама, войска, русский, оружие, операция,

(китайцы): китайский, россия, производство, китаи, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,

(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,

(канадцы): команда, игра, игрок, канадский, сезон, хоккей, сборная, играть, болельщик, победа, кубок, счет, забирать, хоккейный, выигрывать, хоккеист, чемпионат, шайба,

(немцы): армия, война, войска, советский, военный, дивизия, немец, фронт, немецкий, генерал, борт, операция, оборона, русский, бог, победа

Тематическая модель для коротких текстов

- ▶ Каждое слово представляется вектором из PMI-оценок. Составляется матрица из косинусных расстояний между такими векторами.
- ▶ Производится симметричное неотрицательное матричное разложение (SNMF) методом ALS.
- ▶ Подбирается матрица описания документов, соответствующая найденной матрице описания слов.

$$\begin{array}{c} \boxed{S} \\ \\ \boxed{X} \end{array} = \begin{array}{c} \boxed{U} \\ \\ \boxed{U} \end{array} \times \begin{array}{c} \boxed{U^T} \\ \\ \boxed{V} \end{array}$$

Xiaohui Yan et al. – Learning Topics in Short Texts by Non-negative Matrix Factorization on Term Correlation Matrix, 2015.

Направления работы:

- ▶ Построение тематической модели на «псевдо-документах»:
 - ▶ Как будут отличаться темы от обычной модели?
 - ▶ Вероятностная интерпретация матрицы W' .
 - ▶ Нужна ли регуляризация для такой задачи?
 - ▶ Нужна ли регуляризация в модели SGNS?
- ▶ Включение в модель двух типов документов аналогично мультимодальной тематической модели.
- ▶ Эксперименты на коротких документах: выделение этно-нагруженных тем по сообщениям Вконтакте.
- ▶ Включение PMI-информации в виде регуляризатора когерентности, эффективная реализация на E-шаге.

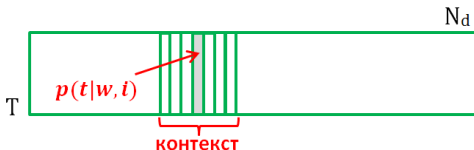
Когерентность темы: тема хорошая, если ее слова статистически часто встречаются «рядом».

$$TCPMI = \sum_{j=2}^{10} \sum_{i=1}^j \log \frac{n_{ij} N}{n_i n_j},$$

где суммирование ведется по всем парам в топ-10 слов темы.

Регуляризация локального контекста:

$$\tilde{p}_{tdw} = p_{tdw} \frac{1}{n_{dw}} \sum_{i=1}^{n_d} [w_i = w] \left(1 + \frac{\partial R_{di}}{\partial p_{tdw}} - \sum_{s \in T} p_{s dw} \frac{\partial R_{di}}{\partial p_{s dw}} \right)$$



Эксперимент: модель PLSA на коллекции пресс-релизов

President and Secretary Honor Ambassador Karen Hughes at Swearing-In Ceremony

people attack terrorist u american state war world life terrorism intelligence united
innocent country one ...

democracy woman freedom human right world nation must today year citizen
democratic liberty country free ...

million assistance u refugee united state humanitarian international support
contribution relief information program food organization ...

Тематическая сегментация последовательного текста по $p(t|d, w)$

- ▶ PLSA предполагает, что тема встречается равномерно по ходу документа. Невозможно обнаружить смены тем.
- ▶ Общая лексика относится к наиболее вероятным темам документа. Поэтому ключевая терминология зашумляется.

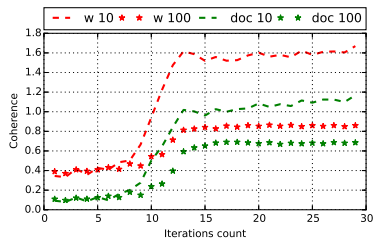
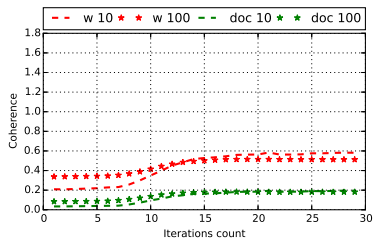
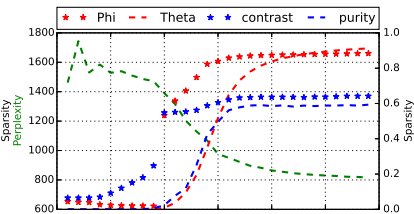
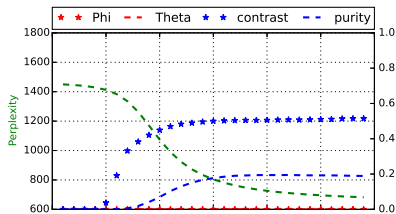
Air Canada's planes assisted in the evacuation. Israel sent tents and mineral water and medical supplies. Italy has sent beds and sheets and blankets and inflatable rafts to help with rescue efforts. Kuwait has pledged \$400 million in oil and a hundred million dollars in humanitarian aid. Qatar and the UAE has pledged \$100 million each. Sri Lanka, one of the world's most impoverished nations that is struggling to overcome the effects of the tsunami, has sent a donation of \$25,000. In all, more than a hundred countries have stepped forward with offers of assistance, and additional pledges of support are coming in every day. To every nation in every province and every local community across the globe that is standing with the American people, and with those who hurt on the Gulf Coast, our entire nation thanks you for your support. Four years ago, the American people saw a similar outpouring of sympathy and support when another tragedy struck our nation, the terrorist attacks of September the 11th, 2001. This Sunday, Americans will mark the fourth anniversary of that terrible day when nearly 3,000 innocent people were murdered. The attacks took place on American soil, yet they left grieving families on virtually every continent. Citizens from dozens of nations were killed on September the 11th. Innocent men and women and children of every race and every religion. And in the four years since the September the 11th attacks, the terrorists have continued to kill -- in Madrid and Istanbul and Jakarta and

Улучшение сегментации регуляризацией локального контекста

- ▶ Ослабление влияния Θ – тематического профиля документа (например, сглаживанием).
- ▶ Использование локальных контекстов, чтобы настроить $p(t|d, w)$ и фильтрация слов, не вписывающихся в контекст, в фоновые темы.

Air Canada's planes assisted in the evacuation. Israel sent tents and mineral water and medical supplies. Italy has sent beds and sheets and blankets and inflatable rafts to help with rescue efforts. Kuwait has pledged \$400 million in oil and a hundred million dollars in humanitarian aid. Qatar and the UAE has pledged \$100 million each. Sri Lanka, one of the world's most impoverished nations that is struggling to overcome the effects of the tsunami, has sent a donation of \$25,000. In all, more than a hundred countries have stepped forward with offers of assistance, and additional pledges of support are coming in every day. To every nation in every province and every local community across the globe that is standing with the American people, and with those who hurt on the Gulf Coast, our entire nation thanks you for your support. Four years ago, the American people saw a similar outpouring of sympathy and support when another tragedy struck our nation, the terrorist attacks of September the 11th, 2001. This Sunday, Americans will mark the fourth anniversary of that terrible day when nearly 3,000 innocent people were murdered. The attacks took place on American soil, yet they left grieving families on virtually every continent. Citizens from dozens of nations were killed on September the 11th. Innocent men and women and children of every race and every religion. And in the four years since the September the 11th attacks, the terrorists have continued to kill -- in Madrid and Istanbul and Jakarta and

Сравнение LDA и модели локальных контекстов



Gaussian LDA: другой способ комбинирования LDA и word2vec

- ▶ Тема – многомерное нормальное распределение над векторными представлениями слов.
- ▶ Документ – мультиномиальное распределение над темами.

