

К ТЕМАТИЧЕСКОЙ СТРУКТУРЕ ЖИВОГО ЖУРНАЛА

Олеся Кольцова

Апрель 2012

Москва

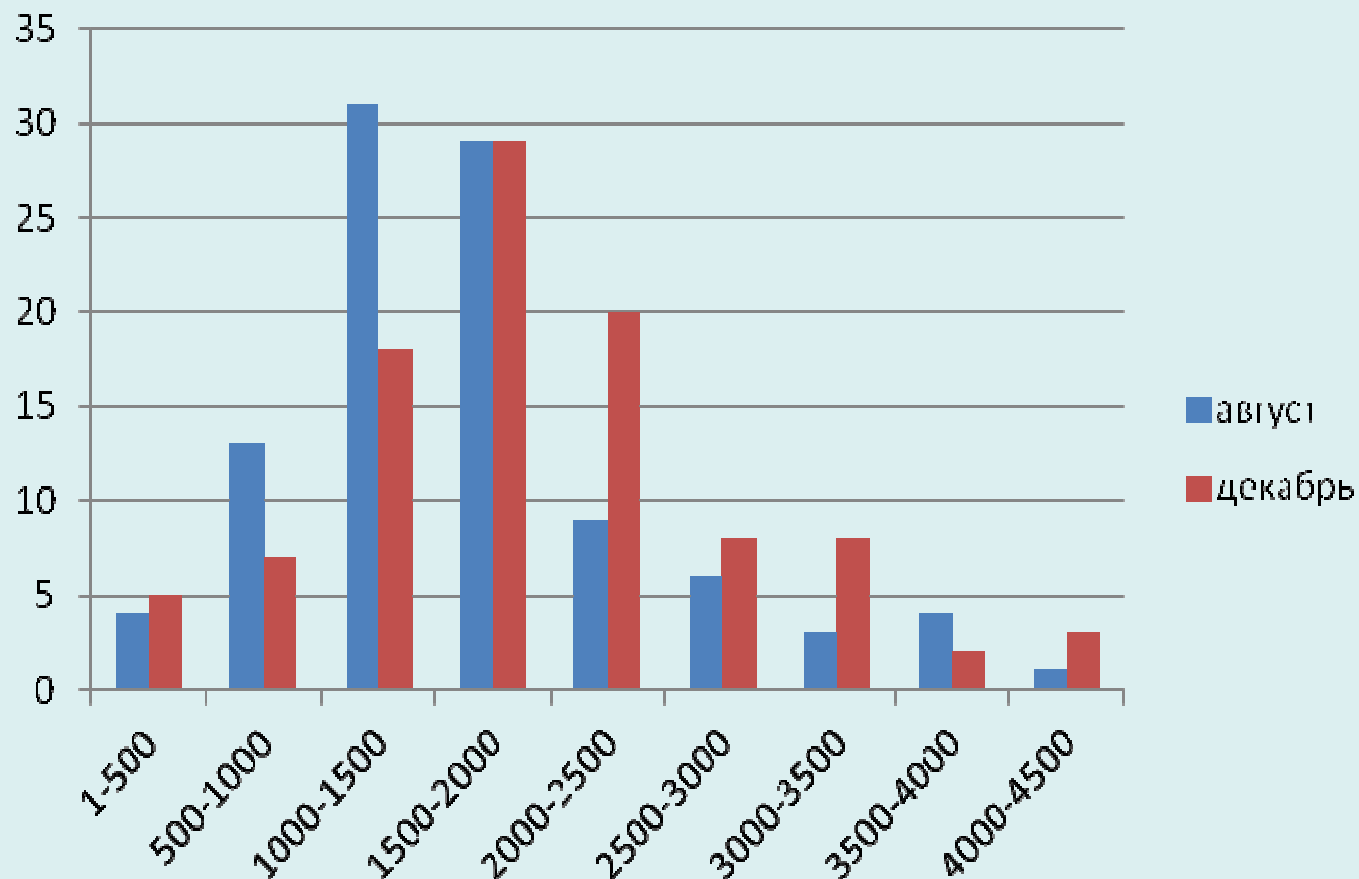
МЕТОДОЛОГИЯ

- Тематическое моделирование
- Stanford Topic Modeling Toolbox
- Алгоритм латентного размещения Дирихле
- Предполагается, что каждая коллекция текстов содержит n скрытых тем
- Каждая тема характеризуется смесью слов, принадлежащих к ней с разной вероятностью.
- Каждый текст представляет собой смесь тем, к которым он принадлежит с разной вероятностью

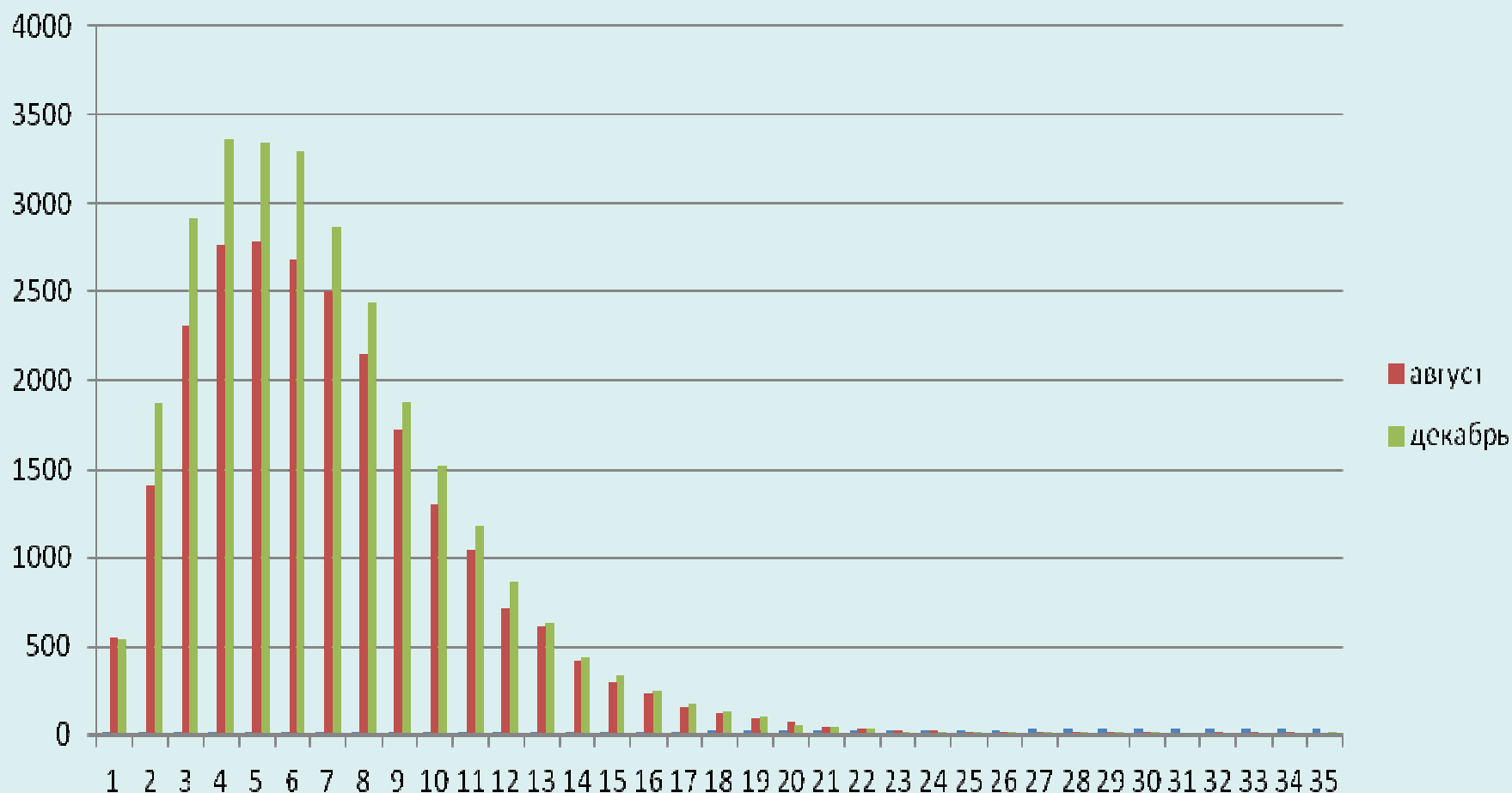
МЕТОДОЛОГИЯ: ССЫЛКИ

- Статья, где впервые вводится алгоритм (много формул)
D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.
<http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf>
- Более популярное изложение на сайте разработчика (почти без формул)
D.M.Blei Introduction to Probabilistic Topic Models
<http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf>
- Обзор вероятностных тематических моделей на русском языке
Daud, A., Li J., Zhou L., Muhammad F. Knowledge Discovery Through Directed Probabilistic Topic Models: a Survey. *Frontiers of Computer Science in China*, Vol. 4, Issue 2, pp. 280-301, 2010. пер. К.В.Воронцов, А.В.Темлянцев и др.
www.machinelearning.ru/wiki/images/9/90/Daud2009survey-rus.pdf

«РАЗМЕРЫ» ТЕМ



ПРИНАДЛЕЖНОСТЬ ТЕКСТОВ К ТЕМАМ



«ТОПОВЫЕ» СЛОВА ТЕМ: ПРИМЕР

Topic 000	25397
украина	1003
президент	611
российский	532
украинский	440
лукашенко	234
глава	229
министр	211
виктор	180
союз	178
янукович	175
государство	170
тимошенко	166
vladimir	164
заявлять	163
беларусь	163
совет	161
ядерный	161
безопасность	158
москва	147
белорусский	147

Topic 001	29420
немцов	905
путин	477
навальный	368
собчак	286
оппозиция	252
разговор	209
болотный	204
божен	200
народ	197
выходить	196
сахаров	195
борис	184
выступать	171
кургинян	166
проспект	163
приходить	161
ксения	158
лидер	152
освистывать	152
называть	139

«ТОПОВЫЕ» ТЕКСТЫ ТЕМ: ПРИМЕР

- 1,0 Из опубликованных ФСБ телефонных разговоров Немцова с Яшиным, Пархоменко, Пономаревым, Панюшкиным и другими я понял только то, что у Бориса Ефимовича какие то недопонимания с Евгенией Чириковой и что он реально не хотел подставлять людей под омоновские дубинки.
- 0,94 Из чужого твиттера: один чувак отнес Навальному LJUSERnavalny в изолятор мандарины, чурчхелы, сулугуни и лаваш. И сопроводил это запиской: "от Кавказа который хватит кормить"
- 0,88 А чо, теперь самые главные революционеры - это собчак, канделака и боженарынска? Суркофф - замечательный разводчик всё-таки!
- 0,87 Белые ленточки. Путин. Выборы. Беспредел. Митинг на Болотной. Посмотрим. Набигут ли спамеры.
- 0,83 После того как выложили записи Немцова - надо точно идти 24-го на митинг. Пусть все знают, что я трусливое офисное хомячье! Ура, товарищи!

НЕКОТОРЫЕ РЕЗУЛЬТАТЫ

- Социально-политические темы в спокойном и в политизированном периоде занимают около трети веса
- Остальное, за исключением 15% «шума», - рекреативные и приватные темы.
- В декабре половина социально-политических тем посвящена выборам и протестам

ВЫБОРЫ И ПРОТЕСТЫ

- **Ретрансмиссия новостей: селективная повестка**
 - 51607 увольнение руководства «Коммерсанта»
 - 43985 реакция властей на протесты (политическая реформа Медведева)
 - 33276 репортажи о протестах
 - 29210 списки выступающих 24 декабря
 - 20152 арест Удальцова
 - 16929 регистрация и отказы в регистрации кандидатов в президенты



ВЫБОРЫ И ПРОТЕСТЫ

- **Мнения**

- 91135 короткие эмоциональные комментарии по кандидатам в президенты
- 54322 за и против митингов – длинные аргум. тексты – Кононеко, lady spring
- 29402 короткие эмоц. комментарии по персонажам протестов, прослушка Немцова
- 11757 длинные аргум. посты о президентских выборах



ВЫБОРЫ И ПРОТЕСТЫ

- **Альтернативные новости**

- 51042 личные отчеты об участии в митингах на Болотной и Сахарова
- 19286 отчеты наблюдателей на выборах



ВЫБОРЫ И ПРОТЕСТЫ



- **Особенная тема:**
50066
предварительные
результаты
голосования,
ЭКЗИТ-Полов,
соц.опросов, явки,
их сопоставление
и рассуждения о
достоверности
выборов

ДРУГИЕ СОЦ.-ПОЛ.ТЕМЫ

- 40023 Американо-ближневосточные отношения
- 25397 российско-украинско-белорусские отношения
- 23974 Российско-Израильские и русско-еврейские отношения
- 20313 История российско-кавказских отношений

- 33452 Китай и мировая экономика
- 28288 мировые финансы

- 58287 криминал и суды
- 46367 региональное управление
- 30537 церковь и православие
- 26740 образование
- 24367 вооруженные силы
- 20588 коррупция

НЕ СОЦ.-ПОЛ. ТЕМЫ

ПРИВАТНАЯ СФЕРА

- 69998 стихи и романтические фотографии
- 57873 любовь и межличностные отношения
- 37799 здоровье
- 36032 дом, семья, работа
- 34850 деньги, работа, зарплаты, кредиты
- 27820 семья и дети
- 24368 секс и красота

НЕ СОЦ.-ПОЛ.ТЕМЫ

ДОСУГ

- 44190 природа
- 41266 архитектура и храмы
- 32182 Новый год
- 29169 книги
- 28024 театры и концерты
- 20366 животные
- 20159 телевидение
- 19921 музеи и живопись
- 18928 футбол
- 18663 фотографирование
- 17118 блоги

НЕ СОЦ.-ПОЛ.ТЕМЫ: ПОТРЕБЛЕНИЕ

- 50501, 33166 товары и услуги онлайн
- 27170 объявления о событиях
- 31586 кулинарные рецепты – основные блюда
- 26596 рестораны
- 25279 кулинарные рецепты - сладкие
- 24775 одежда и тело
- 24127 мобильные устройства и интернет
- 22885 жилье - ремонт
- 22102 жилье - продажа
- 21322 путешествия

НЕ СОЦ.-ПОЛ. ТЕМЫ: ЛЕКСИЧЕСКИЕ ОБЪЕДИНЕНИЯ

- 30814 ненормативная лексика
- 25197 английский язык
- 24989 украинский язык
- 17917 погода
- 14275 календарь