

Девочки, я просто в ШОКЕ 🤯

или анализ корпуса блогов в Instagram



В чем идея?

Посмотреть, как выглядит **язык типичного “неэкологичного” блога** (уезжаю жить в тундру, зарабатываю 20 млн в месяц, я счастливая мамочка и против бодипозитива, записывайтесь на курс).

Сравнить с текстами аналогичного объема и формата - **новостями**, посмотреть как отличается язык.

Правда ли, что все, о чем говорят блоггеры - **сторис, лайки, марафоны, дети и мужья?**

А серьезно?

Исследовательский вопрос:
можно ли выделить тематику блога с помощью корпусного анализа?

Наверное.

+слова, которые на нее указывают будут самыми **частотными**
+язык блогов отличается от просто публицистики
+блоги между собой в общем-то мало различаются, только отдельными частотностями отдельных слов, но все пишут на общем **“инстаграммном языке”**

Как реализовали

- навскидку взяли блоггеров, тематики +- различные. От обзора фикспрайса в Казани до жизни в Дубае.
- спарсили по 50 последних постов со страницы в Инстаграм у каждой(побились с кодировками)
- лемматизировали MyStem'ом
- загрузили в Voyant Tools
- параллельно лемматизировали и загрузили в Voyant маленький кусочек новостного корпуса тайги
- обучили модель на корпусе Инстаграмм-текстов(правда с очень небольшим окном)

Материалы исследования и этапы обработки данных

650 последних постов со страниц в Инстаграм **13** блогеров с разной тематикой (от обзора фикспрайса в Казани до жизни в Дубае)

- парсинг страниц в Instagram
- лемматизация текстов с помощью PyMorphy
- обучение модели

сопоставимый по размеру кусочек новостного сегмента корпуса “Тайга”

- лемматизация текстов с помощью PyMorphy

Анализ корпусов

1

Voyant Tools

2

Stylo

3

Python

Разделы

- 1) Можно ли по корпусу понять тематику блога?
- 2) Чем посты в Инстаграм отличаются от новостей?
- 3) Самые частотные слова
- 4) Бонус: генератор абсурда

@olololnew
@mezenova
@kristitheone
@_madsti_
@maslovaa.a
@nataliamit
@len_club
@mommy_to_sofia1405
@djuelita
@protasovnaa
@lena_kartss
@yana_leventseva
@dariko_kutaladze
@dasha_cher
@nioly



Это список блоггеров



Можно ли по корпусу понять тематику
блога?

Облако слов без стоп-слов

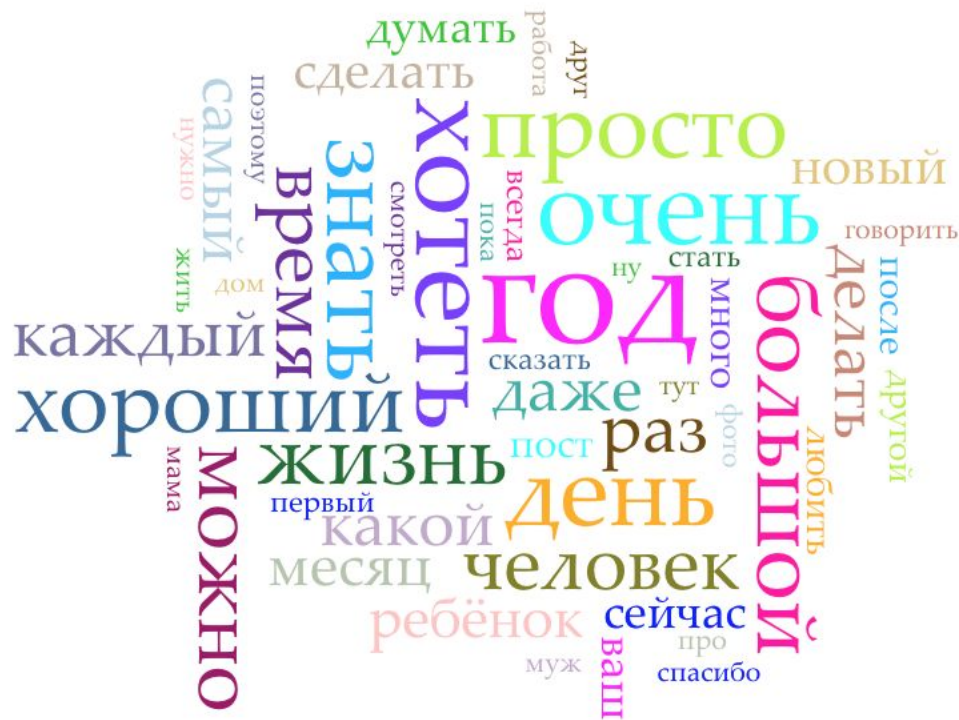


Обратите внимание на

- я
- себя
- свой
- жизнь

очевидно, что это **личные** блоги, и авторы предпочитают писать о себе и своей жизни

А теперь со стоп-словами



Обратите внимание на:

- год
- ребенок
- муж
- мама
- дом
- друг
- работа
- любить

Как искать различие между блогами?



Облако слов

- менеджерить
- менеджержство

Distinctive words

Сравнение distinctive words

интерьер
предмет
икеа

турция
круиз
майами

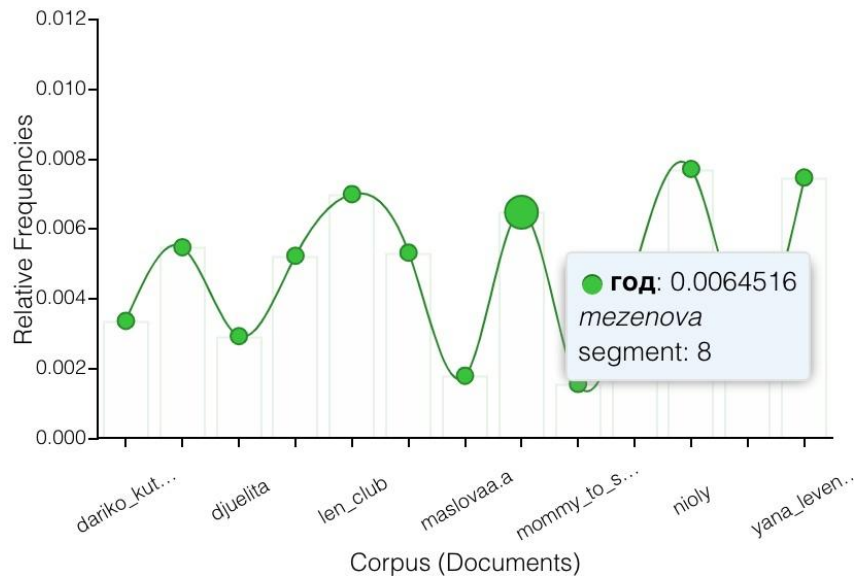
добавлять
курица
увеличивать

любовница,
любовник, жена,
муж, женатый

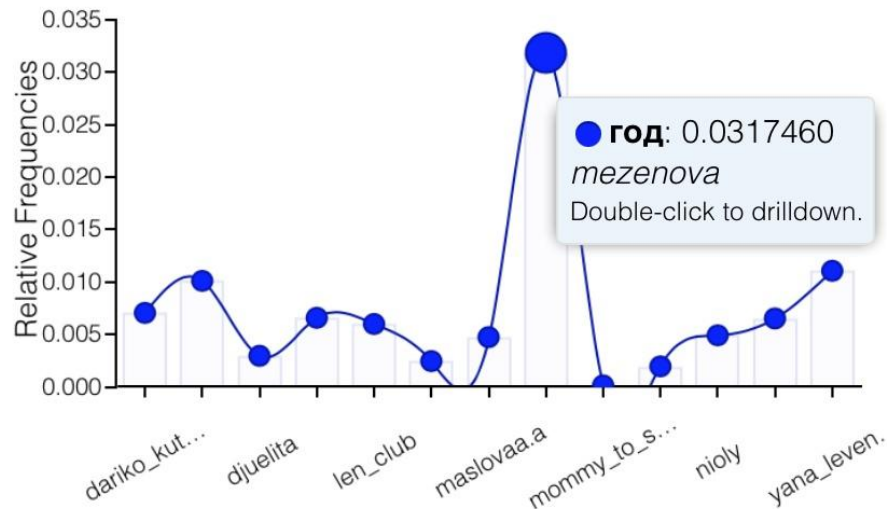
фикспрайс
валдберис
ноготь

шаманизм
коучинг
шаманский

Итоги года




Все посты



Популярные посты

Хотя *обычно* у блоггеров самые залайканные посты *не отличаются* по тематике от постов со средним количеством лайков, но предыдущий слайд показывает **интересный кейс**. У

@mezenova на **2 порядка** различается общая употребимость слова “год” и употребимость слова “год” в самом популярном посте. Видимо, посты с итогами года привлекают людей 🙋



Чем посты в Instagram отличаются от
новостей?

Почему сравниваем с новостями?

1. Похожее время создания

разница максимум в несколько лет. А с классической литературой сравнивать бесполезно

2. В некотором смысле одинаковая тематика

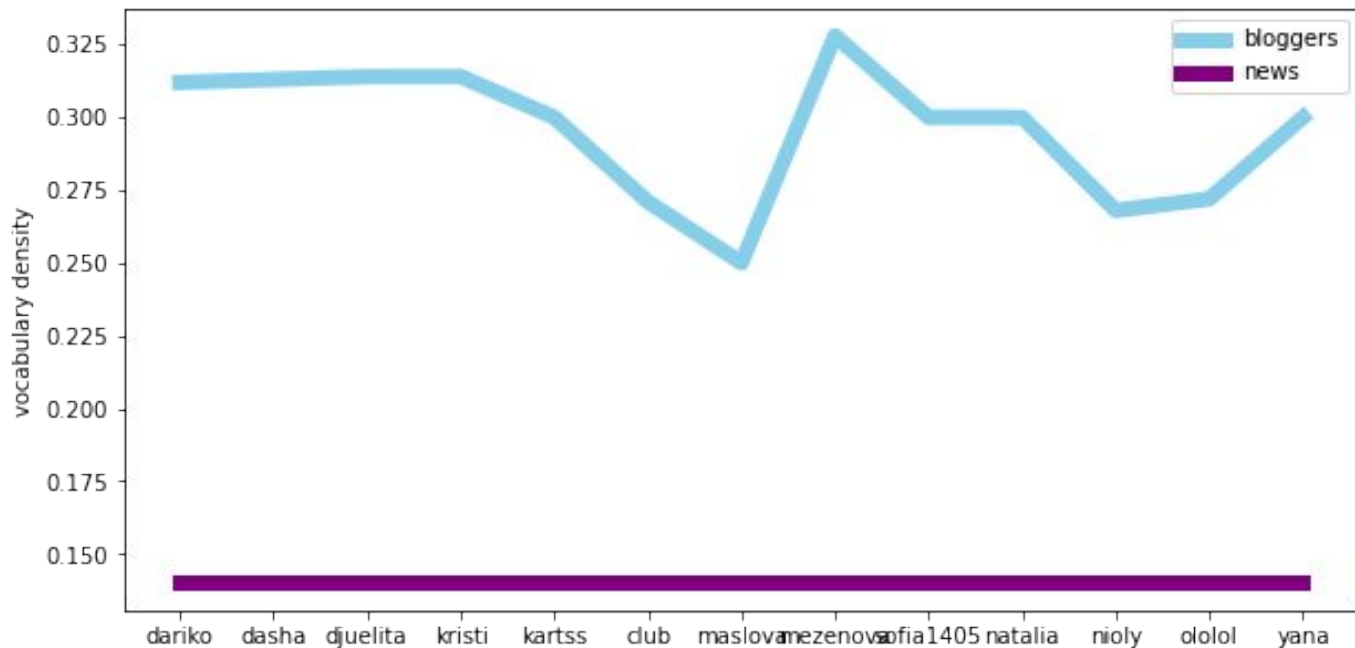
родился, женился, деньги, дома и проч.

3. Одинаковый объем

в средней записи в блоге - 1018 символов, примерно $\frac{2}{3}$ страницы А4

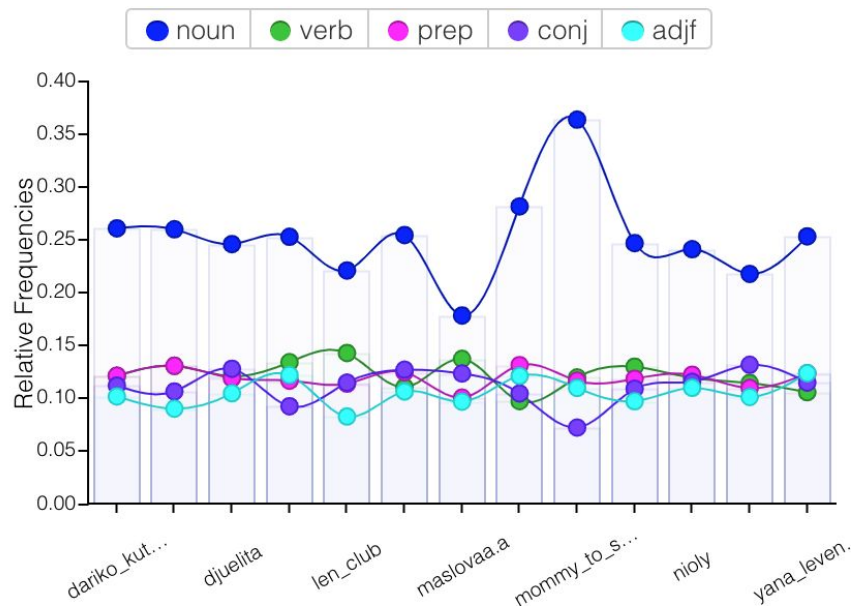
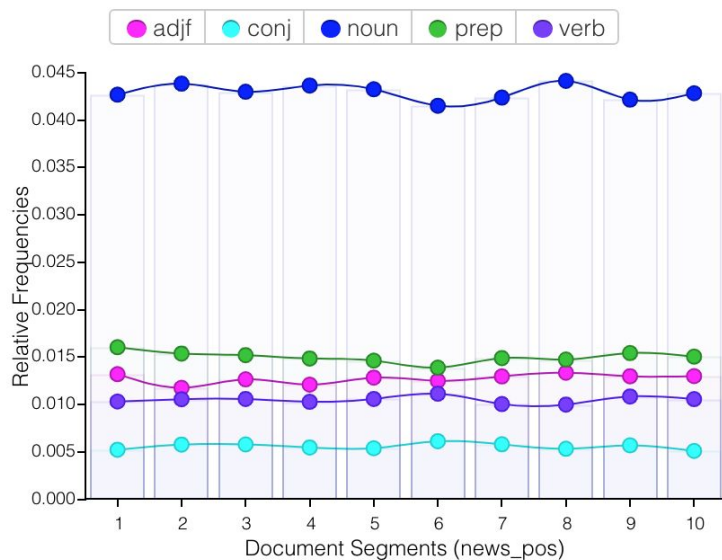
НОВОСТИ VS ПОСТЫ

Газеты сильно уступают блогам в vocabulary density



НОВОСТИ VS ПОСТЫ

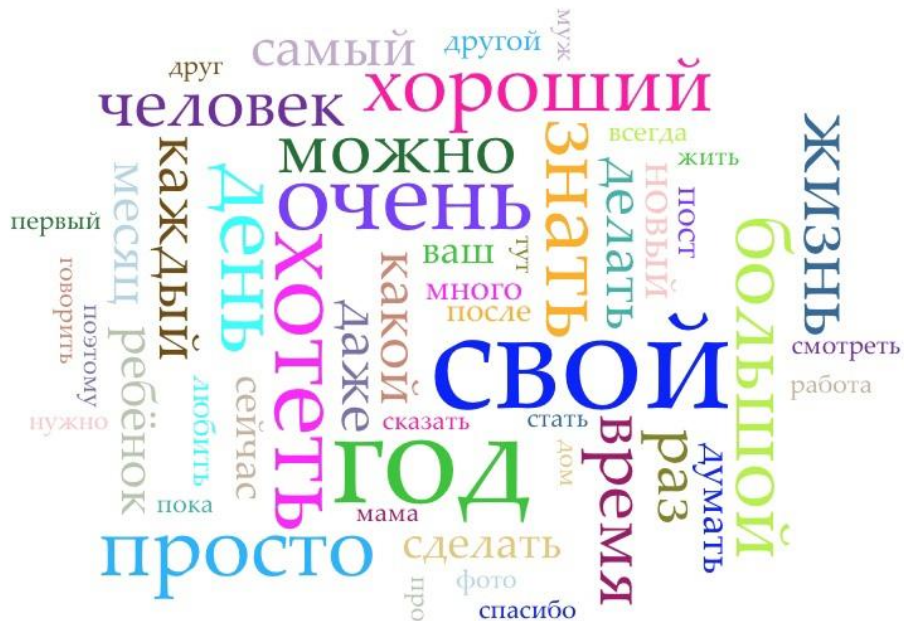
Одинаковое распределение частей речи



НОВОСТИ

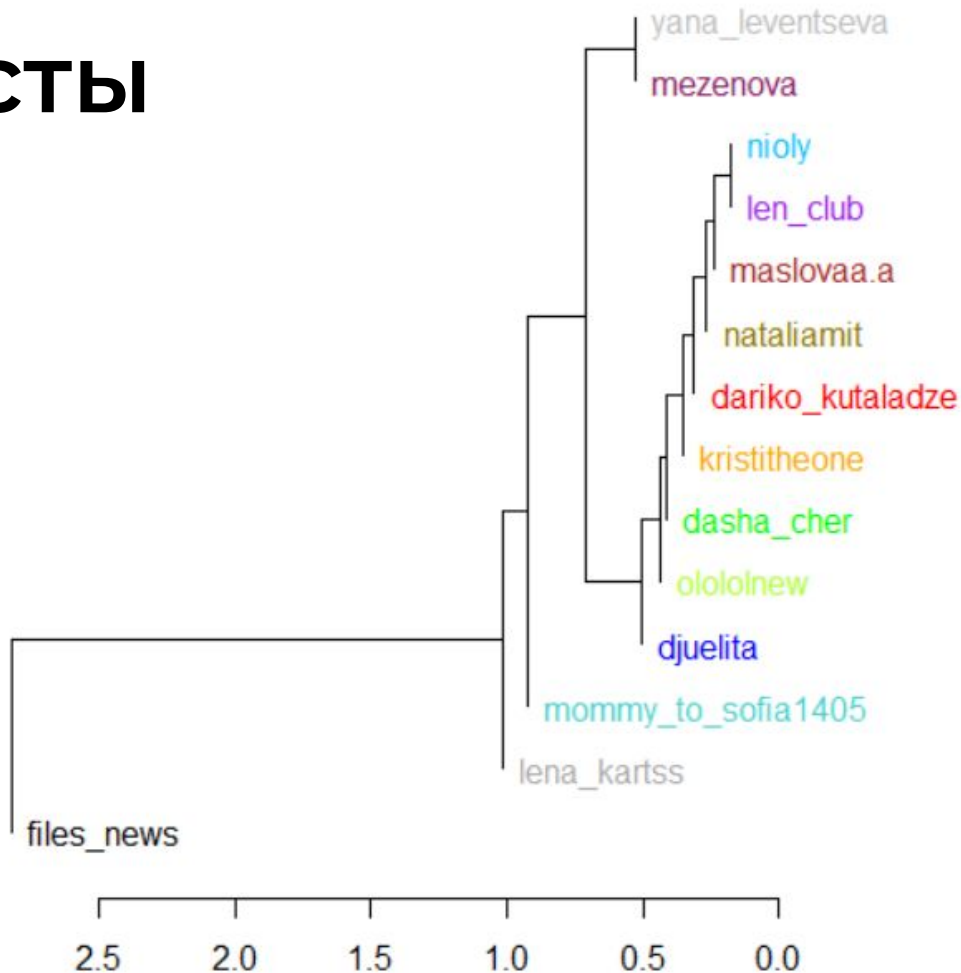
VS

ПОСТЫ



НОВОСТИ VS ПОСТЫ

Новости резко
отличаются в
анализе stylo



3

Самые частотные “блоггерские” слова

Самые популярные прилагательные

1. большой 271

2. ПЕРВЫЙ 151

3. нужный 91

4. красивый 77

5. ЛЮБИМЫЙ 70

6. ?крутой 58

Самые популярные существительные

1. СТОРИ/ИСТОРИЯ	252
------------------	-----

2. ПОСТ	158
---------	-----

3. ВОПРОС	101
-----------	-----

4. ФОТО	99
---------	----

5. КОММЕНТАРИЙ	84
----------------	----



Генератор абсурда*

*в заголовке ссылка на тетрадку

Мы обучили нейросеть
на корпусе и выбрали
самое лучшее из того,
что она выдала.

В сторис будут
подробности про
выбор про
отношения про мужа.
Просто не стоит
ставить отношения в
сториз на страницу.

Ставь.❤️ Сохраняй

Тык-тык ❤️

Помощь от родных
происходит с общения и
просто подарки. И как с
массажа и подарками на
подписчиков, и получаешь
себя. Не могу выставить
всего, поддержки которые в
сториз можно получить

Получила удовольствие от
шапки профиля. В сторис мама
с детьми и подруга.

Выводы

- 1) хотя тему блога можно выделить по distinctive words, в общем-то блоги не особо друг от друга отличаются: они очень близко на стилометрии.
- 2) генератор текста, основанный на частотности символов говорит о том же. Чаще всего он генерирует какой-то поток сознания со словами “сторис”, “подписчик” и тд.
- 3) посты в Инстаграме и новости - разные вещи