Анализ данных с телеграм канала Job for Analysts & Data Scientists



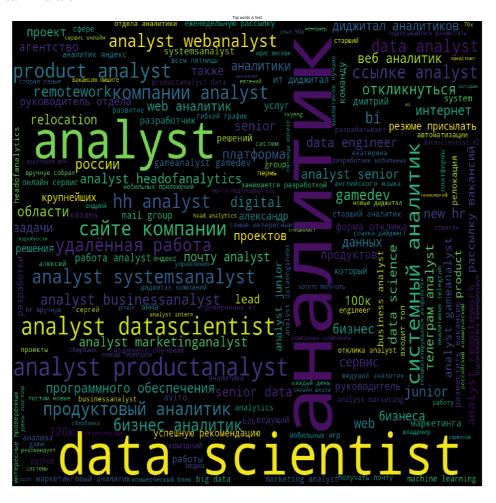
Источник информации — телеграм канал Job for Analysts & Data Scientists (https://t.me/foranalysts) Вакансии для Digital-аналитиков и Data scientists. Канал создан рекрутинговым агентством New.HR

Инструменты — Python:

- telethon для сбора сообщений с канала в телеграме
- gensim для тематического моделирования
- pandas, matplotlib для анализа табличных данных и визуализации
- nltk токенизация текста, работа со стоп словами

Облака тэгов

Наиболее часто встречающиеся слова в описании вакансии:



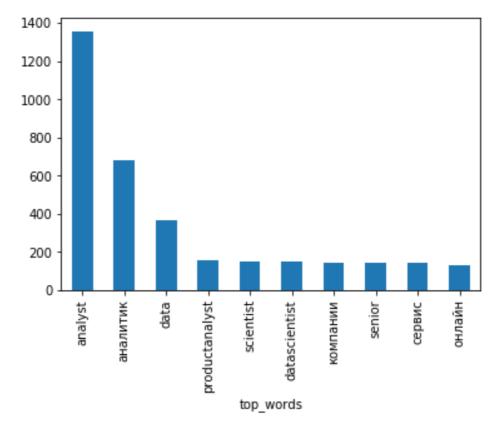
В описании вакансии наиболее часто повторяющиеся слова - это «data scientis», «аналитик», «analyst», «удаленная работа», чуть реже встречаются слова «системный», «продуктувый» аналитик.

Данные результат стал возможен после удаления стоп слов и часто встречаемых слов не имеющих отношения к сути вакансии: 'вакансия','москва','что', 'это', 'так', 'вот', 'быть', 'как', 'в', '—', 'к', 'на', 'за', 'было','требуемые','навыки','ищет', 'удалённо', 'специалиста', 'компания', 'откликаться', 'geekjob','описание','вакансии','ru','eë','пост','санкт-петербург', 'санкт петербург', 'минск','личку','санкт', 'петербург', 'санктпетербург', 'оксана','прутьянова','подробнее'.

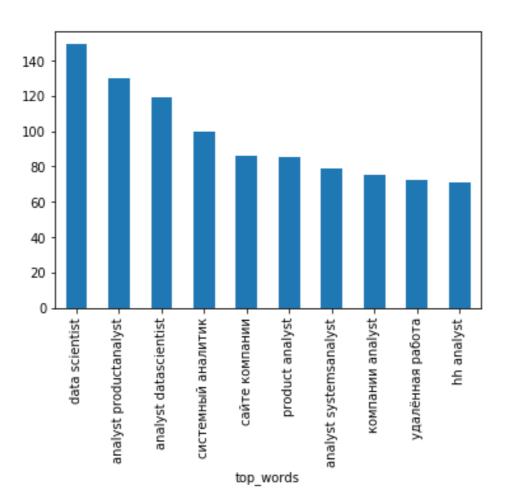
все слова переведены в нижний регистр

города исключены, что бы не вносить много шума в основную картину

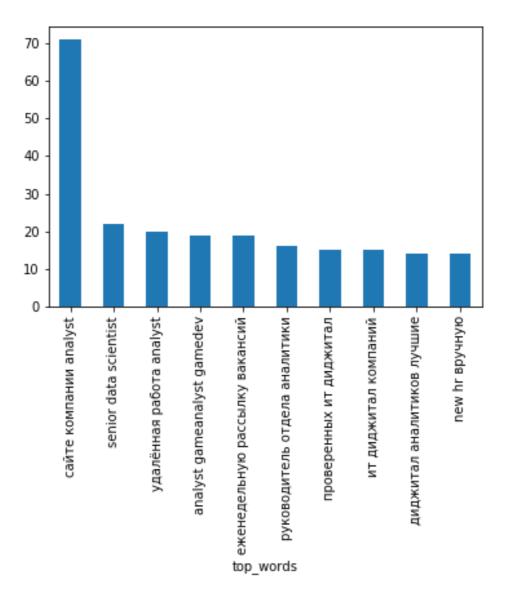
Топ - 10 слов



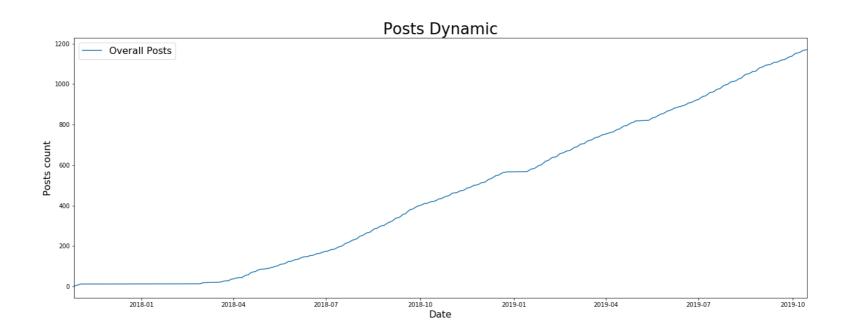
Топ - 10 пар слов



Топ - 10 троек слов

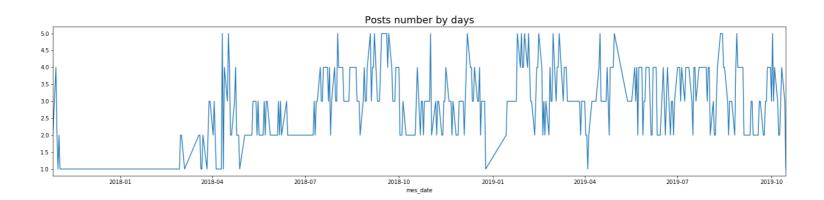


Динамика постов по месяцам

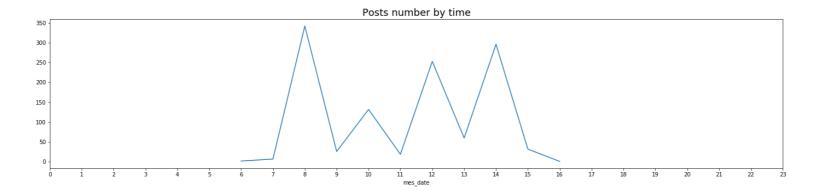


Динамика с накопление, поэтому видим достаточно быстрый подъем с 07/2018 и быстрое увелечение количесва ~ в 6 раз

Динамика постов по дням

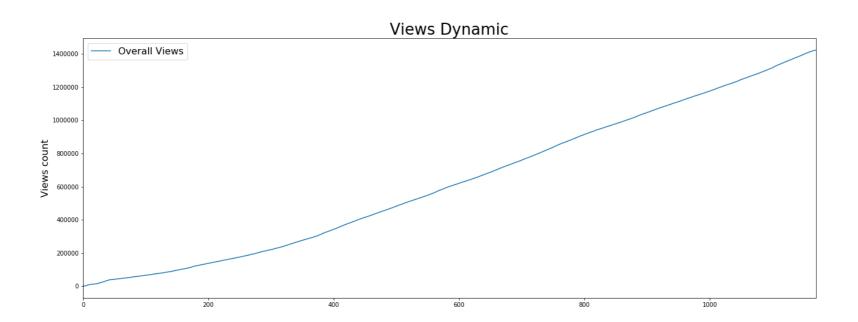


Динамика постов по часам



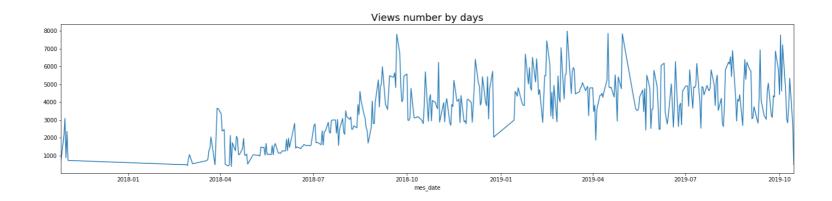
Пики размещения вакансий 8 утра, 12 и 14 часов.

Динамика просмотров



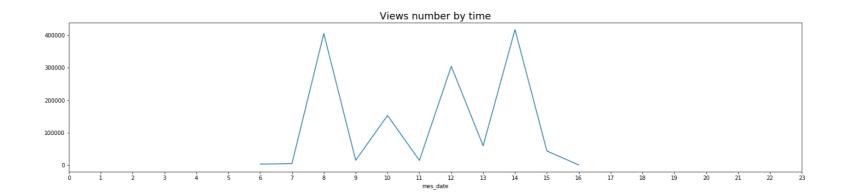
Судя по графику просмотры достаточно линейно, но точно оценить динамику сложно. Посмотрим по дням.

Просмотры по дням



Уже более отчетливо прослеживается тренд, хотя рост не слишком большой и явно замедлился после 07/2019

Просмотры по часам



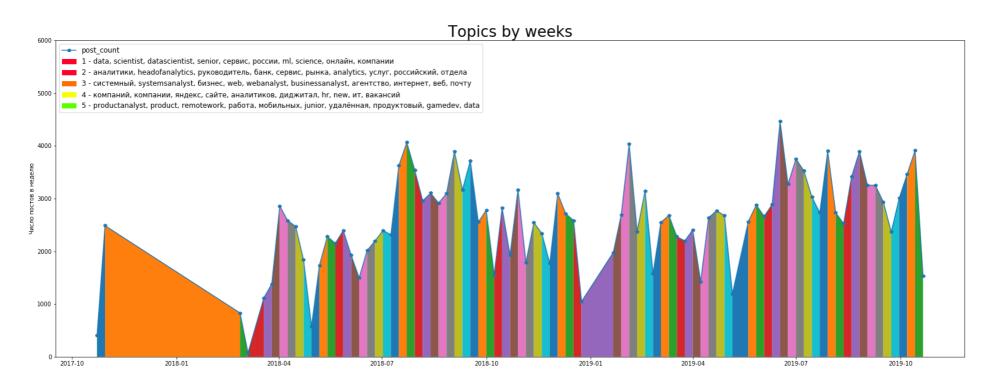
Пики просмотров совпадают с пиковым временем размещения сообщения. Это говорит о том, что подписчики сразу после публикации смотрят пост.

Тематическое моделирования

Построив Латентное размещение Дирихле (LDA) можно увидеть выделение 5 групп, которые я сгруппировал следующим образом: 'data scientist', 'руководители', 'system analyst', 'диджитал', 'product analyst'.

ТемаНаиболее часто встречающиеся слова (14 слов)0 topic : data scientistdata, scientist, datascientist, senior, сервис, россии, ml, science, онлайн, компании, откликнуться, крупнейший, проект, relocation1 topic : руководителианалитики, headofanalytics, руководитель, банк, сервис, рынка, analytics, услуг, российский, отдела, данных, задачи, проект, head2 topic : system analystсистемный, systemsanalyst, бизнес, web, webanalyst, businessanalyst, areнтство, интернет, веб, почту, решений, digital, разработчик, области3 topic : диджиталкомпаний, компании, яндекс, сайте, аналитиков, диджитал, hr, new, ит, вакансий, mail, marketinganalyst, сервис, group4 topic : product analystрroductanalyst, product, remotework, работа, мобильных, junior, удалённая, продуктовый, gamedev, data, разработчик, онлайн, engineer, игр

Распределние тем по неделям



Интересный момент, в самом начале было много вакансия системных аналитиков, а потом доли по тематикам стали стабилизироваться