Влияние новостей на динамику фондового и валютного рынковГоризонтальная линия

# 

# *Введение*

Для решения данной задачи я рассматривал влияние новостей на изменение цены на акции сбербанка и индексов РТС и мос биржи. Задача была сведена к предсказанию бинарной величины (будет ли завтра акция или индекс больше или меньше). Но на будущее я не буду ограничиваться только этой задачей, поскольку решение задачи регрессии или отлова выбросов (сильных изменений цен) может дать лучшее понимание динамики рынков. Стоит понимать, что вся дальнейшая работа - не серебряная пуля для торговли на новостях, а лишь способ понять закономерности в движении цен и индексов. Несмотря на неплохой стабильный результат в 60% на некоторых моделях, я был бы очень осторожен с такими результатами и не делал поспешных выводов на таких данных. Что можно извлечь из моей работы - так это методы, которые потенциально сработают на очень большом хорошо размеченном массиве данных. Кроме того отсюда можно почерпнуть немало интересных выводов и находок, всплывших во время работы, а также учесть минусы и несостоятельность некоторых подходов.

## *Вместо содержания*

Процесс решения задачи выглядит следующим образом: Есть 4 блока - новости(сами тексты), их обработка, модели и выводы Для каждого вида новостей сделана обработка (+- одна и та же), построено много различных моделей и сделаны пометки и выводы.

**Данные 1**

Новости про сбербанк c 2010 по 2020 год, взятые с newsru.com - 1700 новостей.

Очистка текстов довольно стандартная - выкидывание стоп слов русского языка, приведение к нижнему регистру, удаление ссылок, удаление всех символов кроме алфавита (русского и английского), лемматизация, замена ё на е, удаление слов длиной меньше 2, приклеивание частиц не и ни к словам

Пример до и после очистки:

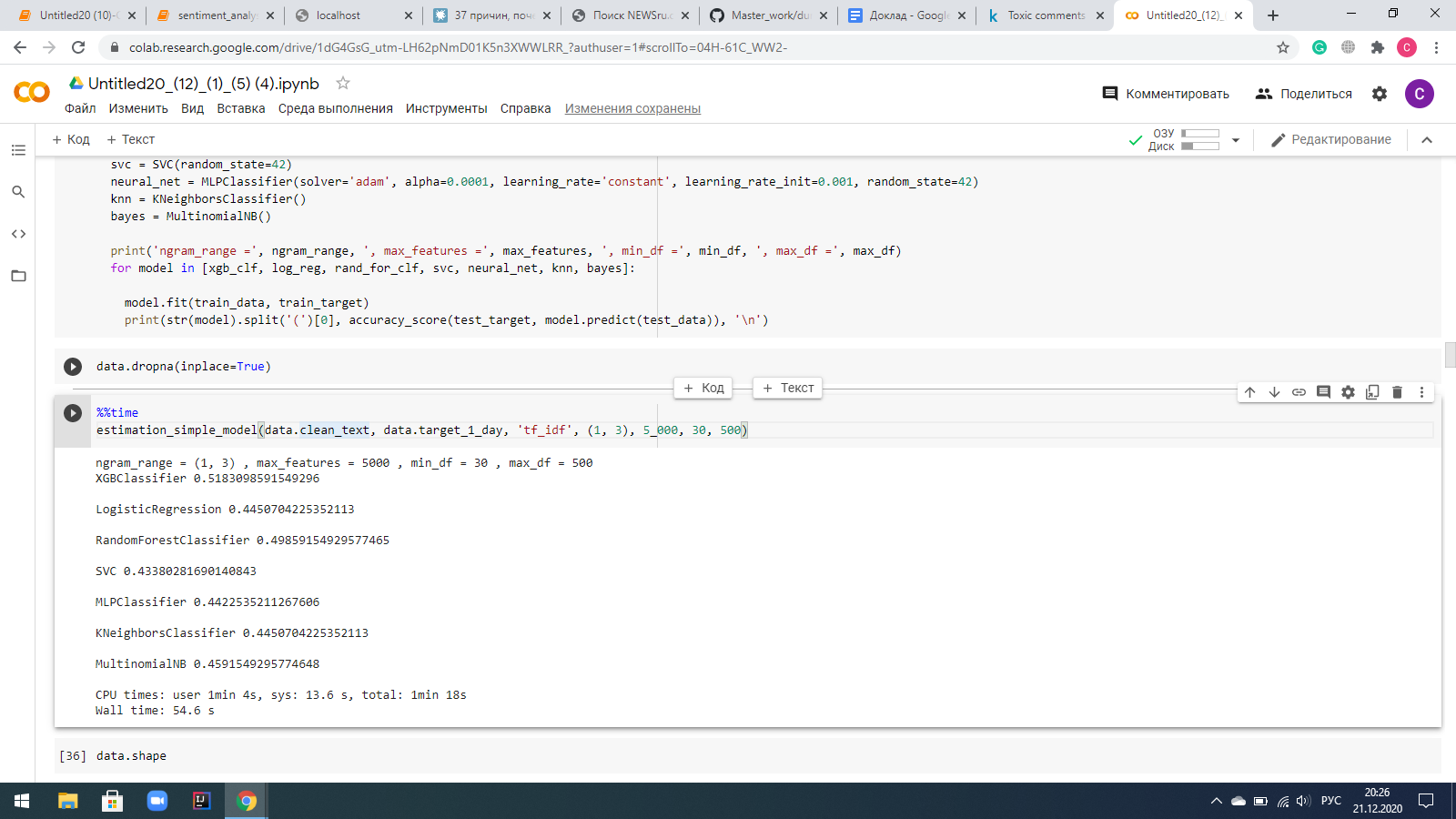
“Второй крупнейший розничный банк России - "ВТБ 24" увеличил с 25 августа ставки по всем вкладам в долларах в среднем на 0,8 процентного пункта и снизил ставки по вновь открываемым депозитам в рублях в среднем на 1 процентный пункт, передает агентство РИА "Новости" со ссылкой на сообщение банка.Как следует из сообщения, максимальная ставка по долларовым вкладам теперь составляет 5,1%, а по рублевым - 7% годовых”

“второи крупныи розничныи банк россия втб увеличить август ставка весь вклад доллар среднее процентныи пункт снизить ставка вновь открывать депозит рубль среднее процентныи пункт передавать агентство риа новость ссылка сообщение банк следовать сообщение максимальныи ставка долларовыи вклад составлять рублевыи годовои”

После удалил частые и редкие слова, целевую переменную взял 1 - если цена на акции сбера завтра будет больше, 0 - меньше, котировки брал с финам.

### Модели

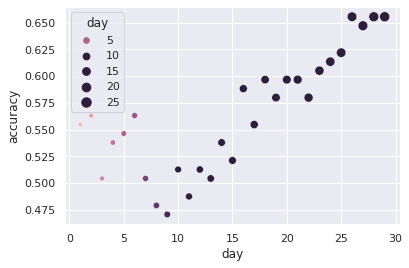
🔘 Самый простой анализ - Tf-idf + простые модели из-под коробки:



Вполне закономерно, что результат плохой, на других датасетах этот подход тоже не показал надежд, поэтому больше его рассматривать не буду. В этом есть логика: цены на акции зависят от новостей явно не таким способом: если есть какие-то слова или выражения, то цена вырастет. Зависимость куда более сложная, но попробовать стоило.

🔘 Другой более интересный подход - vader\_lexicon из NLTK(SentimentIntensityAnalyzer - готовая модель для оценки тональности, но не новости, а текста) дает уже гораздо лучший результат - если брать только новости, в которых модель больше всего уверена, то точность 60% вполне реальная цифра (по крайней мере на этих данных).

Но стоит рассмотреть не очень удачный результат на датасете новостей РБК. На рисунке ниже - качество бинарной классификации в зависимости от того, изменения на какой день я предсказываю - на завтра или на неделю вперед. Судя по графику, модель во-первых не очень устойчивая, о чем говорит разброс на первых днях предсказания, да и рост качества с ростом дней - скорее тренд, нежели какая-то закономерность. Кстати на этом датасете очень много новостей было предсказано как нейтральные, что еще раз говорит о том, что все модели страдают от того, что в нее подаются разные данные.



🔘 Подход, который дает константное качество - бустинг, через который я прогоняю тональность новости, число дней с момента выхода последней новости, ее длину и некоторые другие признаки. Однако его можно развить, но добывать признаки из новостей довольно сложно.

🔘 Подход с применением W2V кажется довольно неплохой идеей - обучить на наших данных(не сработало), или взять пред обученный (лучше, но все равно не то). Дальше, пропуская эмбеддинги через LSTM, я получал не сильно хороший результат. Самое подающее надежды решение - взять не среднее эмбеддингов, а один за другим.Я брал разные W2V модели, у каждой свои проблемы - английская на новостях не подходит, потому что совершенно другое строение наших новостей и зарубежных, наша же модель требует от каждого слова часть речи. Что здесь можно сделать в будущем - раскопать нормально обученный на русских новостях W2V и прогнать данные через нормальную сеть(пока писал отчет, нашел ошибку в коде нейронки)

Более - менее общие замечания по нейронкам

- одного слоя LSTM мало

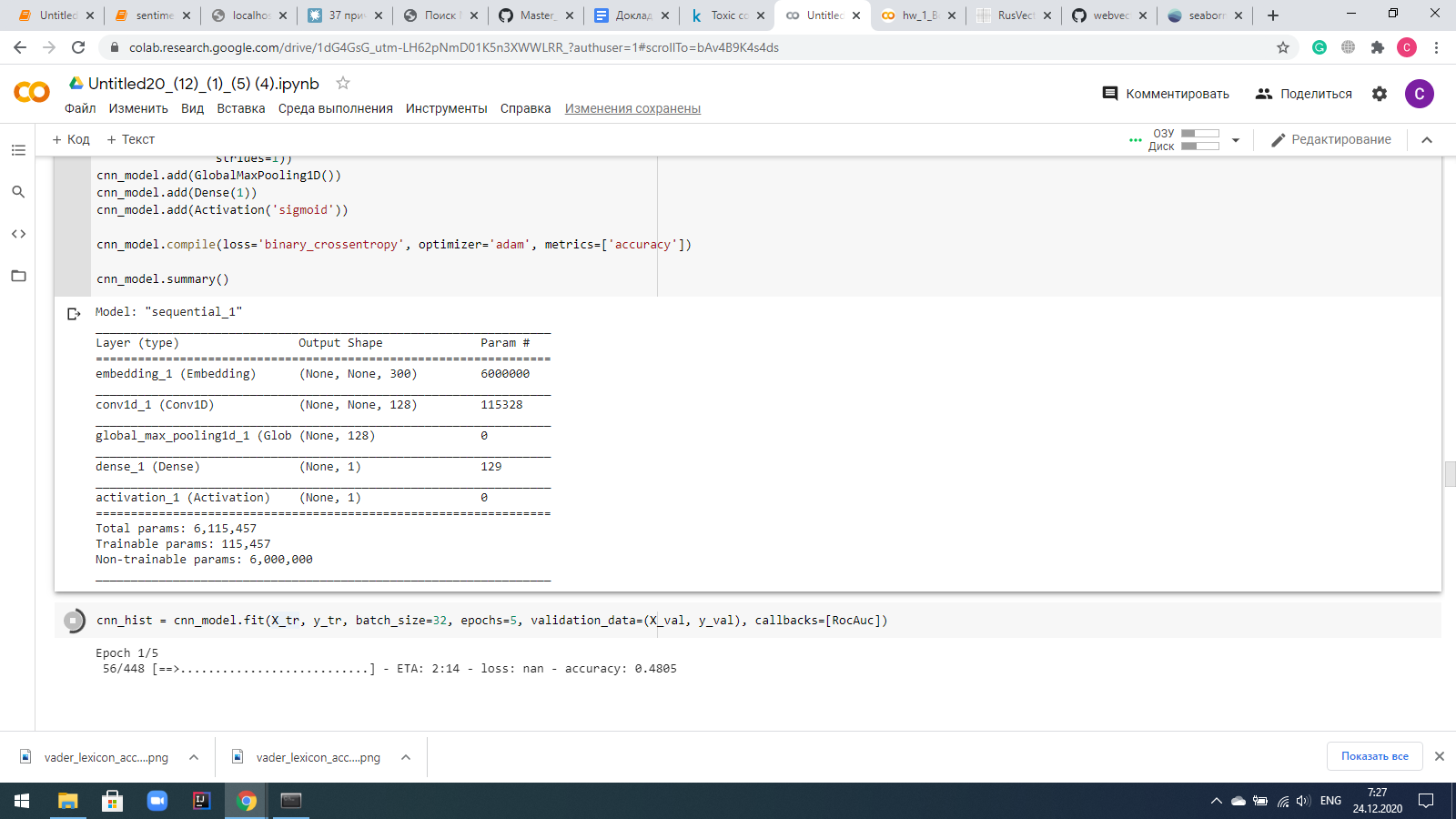
- Dropout обязательно

- линейные слои тоже неплохи, но все же хуже рекурентных

- в нейронки можно запихнуть предсказания тональности(из блока выше)

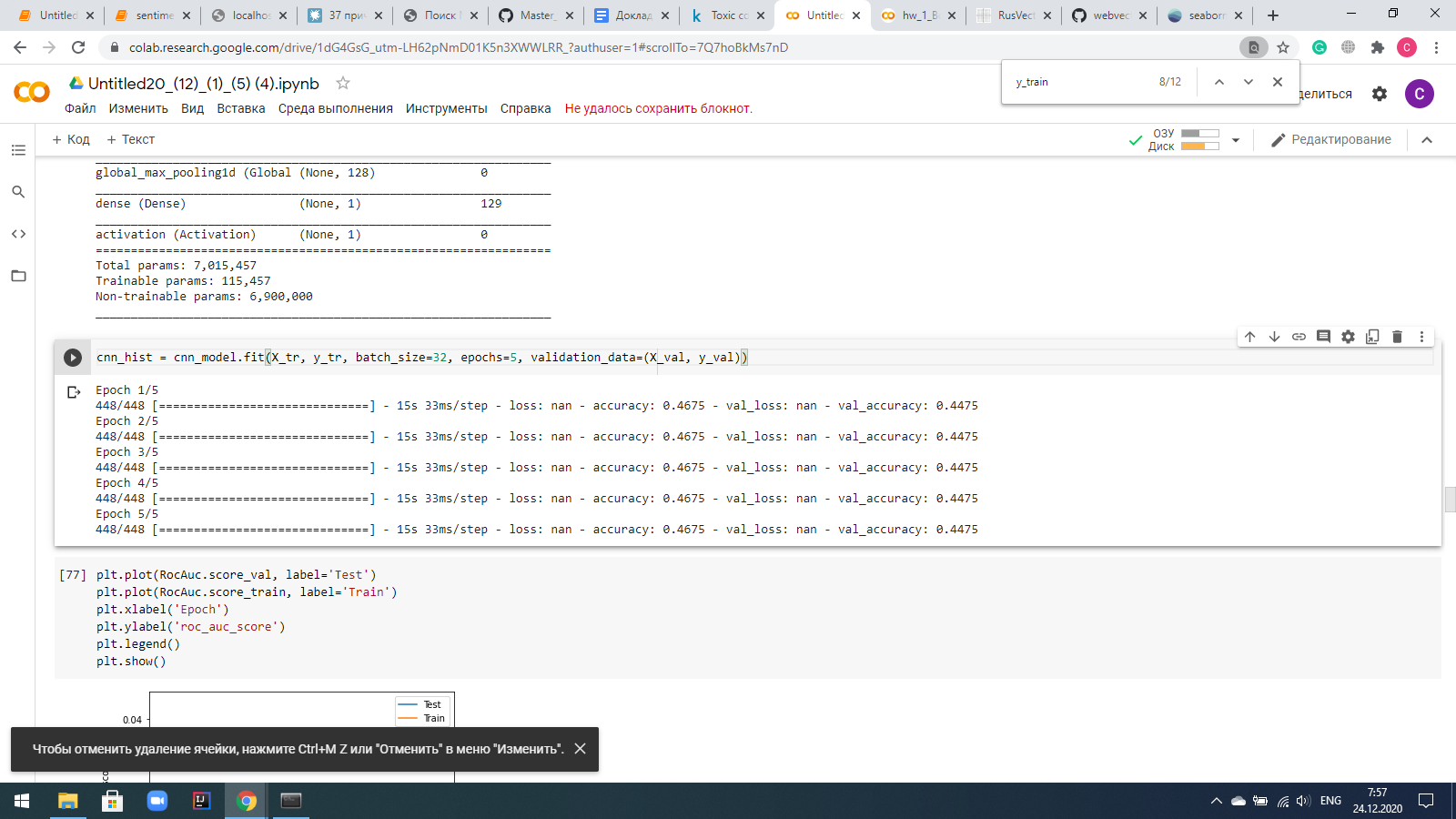
- суммаризация у меня довольно плохая, можно попробовать GPT3 для этой задачи, ожидаю прироста качества

Этот подход хоть не дал мне лучший результат, однако я считаю его самым устойчивым.

В пример приведу одномерные CNN , которые обучаются на корпусе tayga\_upos\_skipgram\_300\_2\_2019отсюда https://rusvectores.org/ru/models/

Текст - новости РБК, слова - с частотой (-100 самых частых) : 5

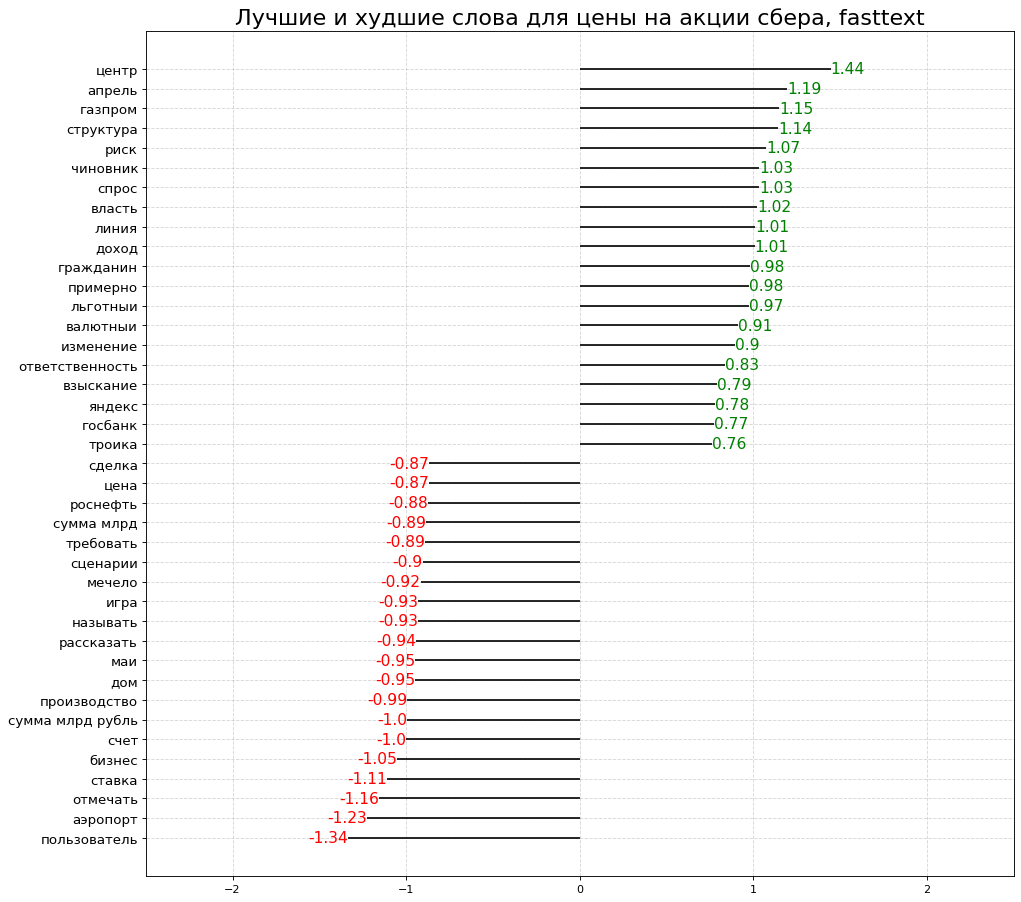
По итогу модель, которая на 5-классовой классификации комментариев давала 98% выдает такое:



Таких случаев с нейронками большинство - просто не учатся. Часто это решается дропаутом или сменой архитектуры (упрощение). К сожалению, с помощью нейронок я добивался лишь 56-57% accuracy.

🔘 fasttext.train\_supervised из коробки дает порой неплохой результат, но его сильно кидает из стороны в сторону, что говорит скорее о неустойчивой модели, которая что-то поймала и пытается выдать это что-то за правду. Что интересного есть в этой модели - это слова, которые она считает важными для принадлежности новости к одному или другому классу. Например логично, что если есть положительная динамика газпрома, то вероятнее сбер тоже растет (эта связь не последовательная, а общая: растут большие компании - растет сбер). После слов спрос и доход в новостях цены растут, согласно этой модели. Не лишено смысла. Несмотря на то, что можно написать и в крайне негативной новости эти слова, но статистически в положительных новостях они встречаются чаще (личное предположение).

Интересно, что слово “пользователь” на первом месте с конца. Бред? А возможно вполне логично. Слово “пользователи” вряд ли будет часто проскакивать в положительных новостях, потому что пользователи если довольны, то нет никакой шумихи. Однако, если они не довольны, то тут же случается информационный бум и в новостях пишут: “пользователи недовольны”. Писать “пользователи довольны” конечно же тоже могут, но первый случай кажется мне более вероятным.



Стоит помнить, что это все - игра с вероятностями, нет строгого результата, нет абсолютно логичных гипотез и на 100% достоверных выводов. Есть некое вероятностное пространство огромной размерности, в котором я хочу предсказать движение цены. Если зафиксировать данные, но поменять обработку и модели - результат будет совсем другим, если поменять данные, но оставить модели - тоже. Учитывая, что ресурсов с данными, равно как и моделей несчетное количество, найти оптимум в этом пространстве невозможно. Но, как по мне, это и не нужно, из той информации, что я получил уже сейчас можно строить вполне реальные гипотезы о динамике рынка. Например, картинка выше может говорить нам о том, что слова “сумма млрд рублей” - скорее всего о сделке, которая топит сбер, “нефть” - про снижение цены на нефть, как следствие - обвал рубля и падение сбера. Такие гипотезы вполне несложно выдвигать, они покрывают большой пласт рынков на определенный момент времени. Таким образом можно предсказывать движение рынков по инерции, тренды. Теперь главный открытый вопрос - как понять, когда закончился тренд? Это можно попробовать сделать с помощью тематического моделирования на разных таймфремах. Допустим, берем предыдущий месяц, ищем темы, сравниваем с темами нынешнего. Если похожи - тренд продолжается. Потом берем таймфреймы поменьше и делаем то же самое. Это план на следующий семестр. + надо бы финберт затестить