Техническое задание

Парсинг товаров интернет магазина Nordstrom

Адрес магазина https://shop.nordstrom.com/

Категории товаров:

- Women - Clothing (~30 000 позиций)
- Women - Shoes (~13 000)
- Men - Clothing (~10 000)
- Men - Shoes (~3500)

Пример товара: https://shop.nordstrom.com/s/thread-supply-double-breasted-peacoat/3738126

Для каждого товара необходимо заполнить поля ("[]" - список, "?" - опционально):

- **url**: URL очищенный, без query params (https://shop.nordstrom.com/s/thread-supply-double-breasted-peacoat/3738126)
- **id**: ID из URL (3738126)
- **name**: Наименование (Double Breasted Peacoat)
- **brand**: Бренд (Thread & Supply) привести к правильному регистру
- **size_addition**: Иногда встречается под брендом (пример: "Regular & Tall", см https://shop.nordstrom.com/s/levis-514-straight-leg-jeans-tumbled-rigid-regular-tall/3394388)
- **price**: Цена ($37.90)
- **original_rice**?: Оригинальная цена ($58.00)
- **discount**?: Скидка (35%)
- **description**: Описание (Tortoiseshell-patterned buttons elevate a double-breasted peacoat detailed with classic button-tab cuffs.)
- **color_variants**[]: Несколько вариантов цветов, для каждого:
    - **color_name**: Название цвета (Camel)
    - **color_url**: URL картинки с цветом, очищенный (https://n.nordstrommedia.com/id/7c59d987-9677-418a-a726-ff9d8c466df4.jpeg)
    - **size_variants**[]?: Комбинации размеров, для каждой:
        - **size_name**: Название размера (Medium)
        - **width_name**?: Ширина, встречается у обуви (пример: "M (Medium)", см. https://shop.nordstrom.com/s/steve-madden-gills-platform-slip-on-sneaker-women/4505146)
        - **inventory**: Наличие ("Not available" | "Available" | "Only X left" -> X)
    - **images**[]: Изображения , для каждого:
        - **url**: URL картинки, очищенный (https://n.nordstrommedia.com/id/6aa324aa-e2d7-4583-a06f-436ded444dff.jpeg)
- **size_info**[]: Строки под "Size Info" ("True to size.", "XS=000, S=00-0, M=2-4, L=6-8, XL=10-12.")

- **details_care**[]: Строки под "Details & Care" ("Tortoiseshell-patterned buttons elevate a double-breasted peacoat detailed with classic button-tab cuffs.", "27' length (size Medium)." ...)
- **avg_review_stars**: Среднее кол-во звезд (86.516%)
- **fit_rating**: Fit rating (runs true to size)
- **reviews**[]: Для каждого отзыва (необходимо пройтись по всем страницам пагинации)
    - **stars**: Кол-во звезд (5)
    - **title**: Заголовок (Perfect, Stylish, EXACTLY what I was looking for)
    - **comment**: Текст отзыва (I hesitated to purchase this coat for 2 months, because several reviews stated it was lighter weight. I'm so glad I decided to go ahead and get it. It is perfect for our mild...)
    - **date**: Дата (Jan 28, 2019 -> 2019-01-29)
    - **fit**: Fit (true to size)
    - **author**: Автор (Brookslucinda)
- **bought_together**[]: Frequently Bought Together
    - **id**: ID из URL (4737821)
    - **url**: URL очищенный (https://shop.nordstrom.com/s/4737821)
- **also_viewed**[]: People Also Viewed
    - **id**: ID из URL (4972497)
    - **url**: URL очищенный (https://shop.nordstrom.com/s/4972497)

Как парсить:

- Предпочтительно использовать Python + Scrapy (можно обсудить другие варианты)
- Использовать кэш запросов и сохранять его, предпочтительно LevelDB (не кэшировать ответы с ошибками)
- Сайт может блокировать доступ для IP адресов за пределами США, возможно придется использовать прокси
- Следовать robots.txt и terms of use, не перенагружать сервер
- На kaggle есть датасет, возможно он будет полезен https://www.kaggle.com/PromptCloudHQ/innerwear-data-from-victorias-secret-and-others#shop_nordstrom_com.csv

Результат представить в виде:

- json файл с результатами парсинга (~56000 записей), каждый товар - 1 строка файла (пример:

  [
  {"id": 234324, "name": "abc..."},
  {"id": 234325, "name": "abc..."}
  ])
- исходный код парсера и паука
- инструкции по запуску (парсер должен заработать и скачать хотя бы несколько товаров без кэша, тестироваться будет на macOS)
- кэш html страниц (повторный запуск в идеале должен получить все результаты используя только кэш)

В отклике, пожалуйста, укажите:

- В кратце ваш опыт работы с парсерами
- Какой стек технологий планируете использовать
- Какие трудности вы видите в парсинге shop.nordstrom.com
- Оценка по сроку и стоимости
- Возможные способы оплаты (предпочтительно Яндекс.Деньги или Paypal)
- Условия оплаты и работы (предоплата, этапы)