**Практическая работа №5**

Язык программирования: Python

**Задание:**

Провести синтаксический анализ отзывов на товар.

**Выполнение:**

1. Загрузка данных. Самостоятельно выбрать интересный товар с большим количеством отзывов (не менее 1000 отзывов, содержащих текст). Если не получится загрузить страницу товара с помощью программных средств – вручную сохраните html код страницы. На некоторых маркетплейсах придется прокрутить загрузить отзывы на страницу прокруткой. Далее необходимо извлечь элементы, содержащие отзывы. Например, на ozon все отзывы находятся в элементе <span class="qw\_29">. Парсинг страницы можно произвести, например, с помощью библиотеки BeautifulSoup.

Другой вариант – взять готовые выгруженные отзывы на 3 товара, но обработку и разбор делать со всеми товарами.

1. Проведите предварительную обработку текстов. Удалите лишнее (смайлы, ссылки, другие конструкции, которые не являются словами), символы (часто пользователи используют символы для написания смайлов, знаки пунктуации должны остаться). Разделите тексты на предложения (заметьте, что некоторые пользователи пишут каждое предложение на новой строке).
2. Провести синтаксический разбор каждого предложения.

Для разбора можно использовать одну из библиотек, представленных в презентации. Для каждого предложения должно получиться дерево. Исключите из деревьев слова, которые относятся к служебным частям речи (кроме некоторых частиц, например, «не»).

Информации о тегах можно изучить здесь: <https://universaldependencies.org/u/dep/>

1. Только после синтаксического разбора можно провести лемматизацию для выделенных слов. Некоторые слова могут встречаться в разных словоформах. В графе должна оказаться именно нормальная форма слова. Для слов на русском языке для лемматизации можно использовать pymorphy3, mystem, natasha.
2. Создать общий граф для всех предложений.

В графе добавить веса: на связи между парами слов (пара слов может встретиться несколько раз) и для вершин (слово может встретиться несколько раз)

Визуализировать граф. Можно пользоваться любыми готовыми решениями. Например, networkx. На графе должна быть видна разница между частыми словами и паросочетнаями и нечастыми.