Использовать можно любую их технологий Pig, Hive или SparkSQL (на лекции не было, но попросили). Данные для задачи (Task 3 из pdf): вместо описанных в задании нужно взять только один файл из http://storage.googleapis.com/books/ngrams/books/datasetsv2.html, нужно взять часть с кодом из двух первых букв фамилии. Например, я должен взять файл http://storage.googleapis.com/books/ngrams/books/googlebooks-eng-all-2gram-20120701-to.gz https://yadi.sk/d/7hAI7w0LkQq6b

Бонус (20 баллов): сделать статистический генератор текстов на основании биграмм (определить вероятность перехода к определенному слову из текущего). Сгенерировать текст из 1000 слов. Второй вариант задания: -) убрать ориентацию на дугах в графах из заданий 1-2 a) если graph-x.tsv не является связным, то нужно посчитать количество компонент связности b) если graph-x.tsv является связным, то нужно найти приближение величины диаметра этого графа. c) если работает больше часа, то просто об этом сообщить! Для поиска приближенного значения диаметра графа предлагается проделать следующую процедуру: 0) выбрать от 10 до 1000 случайных вершин графа 1) запустить поиск в ширину из каждой из выбранных вершин 2) выбрать минимум из максимальных расстояний от выбранных вершин.