1.1

1. Есть ли среди выбранных вами ключевых слов редкие слова?

Смотря что понимать под «редкими словами».

Мои частотные слова из текста: *история, глаза, год, роман, друг, жизнь.*
Роман самое низко частотное слово из списка.
Ещё, в частотных словах текста были имена и фамилии главной героини романа и автора романа: *Скарлетт О'Хара и Маргарет Митчел.*

2. Есть ли среди выбранных вами слов слова, вошедшие в топ 500 по частоте?

Сначала, решила посмотреть не руками, а по средствам кода. Мне выдало, что 5 из 6 моих слов входят в топ-500 по частоте. Я не поверила и решила проверить руками. Если смотреть по списку, указанному на сайте словаря, то действительно получится, что в топ-500 входят:

- *Год*
- *Жизнь*
- *Друг*
- *Глаз*
- *История*

3. К каким частям речи относятся выбранные вами слова, слов какой части речи больше?

Я старалась брать именно существительные, потому что прилагательные, как правило, не такие частотные, а глаголы – их тоже можно было взять, но только если *быть*. Но такие глаголы не очень интересно смотреть☺

4. Какие слова встретились во всех или в большинстве документов? Каковы их грамматические характеристики.

Конечно же, больше всего было частиц, предлогов и союзов. Они прямо обосновались на «Олимпе» частотного словаря текста. Но, одним из самых частотных слов было слово *год*.

1.2.

1. Назовите те слова, у которых мощность обратного индекса (количество документов, в которых слово встречается) равна количеству документов в коллекции

Список частотных слов на этом этапе: *история, глаз, год, роман, друг, жизнь, литература, время, человек*.
Получается, что мощность обратного индекса равна количеству документов в коллекции у слов: *история, глаз, год, друг, жизнь, время, человек*. То есть все слова, кроме *роман* и *литература.*
Скорей всего это из-за того, что моя коллекция – это разбитый на равные части роман «Унесённые ветром» и всё равно его тематика крутится вокруг одного и того же. Поэтому и тематические и частотные слова одинаковые во всех частях.

1.3.

1. Соответствуют ли те слова, которые попали вверх списка, упорядоченного по убыванию tf.idf, Вашей интуиции?
   Не очень поняла вопрос, если честно.
   Скорее да, чем нет, потому что я знаю тематику романа и поэтому могу навскидку предположить, какие слова будут ключевыми. Но вот например слово «друг» в частотных ещё даже в первых заданиях было для меня удивительным (там ведь война, какие друзья☺)

2. Все ли ключевые слова попали в верхнюю часть списка (в первые шесть слов), ранжированного по tf.idf?
   Нет, *роман* как всегда выпал.

3. Какие слова попали вниз ранжированного списка? Каковы их характеристики с точки зрения грамматических характеристик, семантики;
   В основном, это глаголы, насколько я могу видеть.

4. Как, по-вашему, должен быть устроен список «стоп»-слов, данные о которых нет смысла включать в таблицу?
   Мне кажется, что необходимо добавить некоторые служебные части речи, распространенные сокращения (как типа *стр=страница*, так и когда пропускаются гласные, как например *пжлст=пожалуйста*), а так же аббревиатуры = сокращения.

5. Какие слова из списка тематически значимых слов, составленного вручную, вошли в список топ 20 слов по tf.idf, а какие не вошли;
   Вошли все: *история, глаз, год, роман, друг, жизнь, литература, время, человек*
   Возможно из-за того, что я опиралась на «знания» из предыдущих заданий.

6. Предложите шаги по улучшению результатов выделения ключевых слов: (а), например, можно использовать нормализацию по максимальной частоте слова в документе; (б) можно попробовать посчитать tf.idf для биграмм. (бонусный вопрос)
   В домашнем задании по семинару, я использовала не лемматизацию, как делает pymorphy2, например, а делала стемминг. Стемминг, по сравнению с лемматизацией, легче, потому что лемматизация опирается на словообразование. То есть она определает часть речи и применяет к слову различные способы нормализации. Стемминг же ищет флективную форму в своей таблице поиска, что значительно упрощает и ускоряет работу алгоритма. Кроме того, стемминг хорошо обрабатывает исключения.

1.4.

1. Отличаются ли диаграммы для самых частотных в языке слов и для слов с высоким tf.idf в Вашем списке, если отличаются, то чем?
   Как и ожидалось, графики для частотных слов более сбалансированные в силу того, что эти слова встречаются в каждом тексте коллекции практически равном (или хотя бы в близком к равным) количестве.