Основные шаги указаны и описаны в ipynb, здесь дополнительные комментарии. В тетрадке также есть комментарии по вопросам из файла инструкции (не совсем понятно, какие нужны, какие нет)

1.1

Есть ли среди выбранных вами ключевых слов редкие слова?

Редкие по частотному словарю: трансляция, изнасилование, видеозапись.

Есть те, которых не было в словаре: <фамилия>, спамер, аудио, стенографирование

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

Например, “год”, “закон”

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

В основном существительные

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

Например, “год”.

Сложно сказать, были исключены стоп-слова из текстов.

1.2

Если TF-IDF равен нулю, то в тексте есть слово (так как если частота ненулевая, то и показатель ненулевой, хотя может быть и очень маленьким)
Поэтому по матрице документ-терм можно отфильтровать столбцы в датафрэйме так, что по двум столбцам будет >0 (присутствующие слова), а в третьем (слова, которого нет), наоборот, 0 (частота 0, числитель 0 -> 0).

5docTFIDF.csv файл с tf-idf для 5 файлов

1.3

Соответствуют ли те слова, которые попали вверх списка, упорядоченного по убыванию tf.idf, Вашей интуиции?

Все ли ключевые слова попали в верхнюю часть списка (в первые шесть слов), ранжированного по tf.idf?

-

Какие слова попали вниз ранжированного списка? Каковы их характеристики с точки зрения грамматических характеристик, семантики;

Там заметно больше глаголов, чем в топе. Причем глаголы общей семантики (идти, давать, иметь, сообщать, начинать).

Как, по-вашему, должен быть устроен список «стоп»-слов, данные о которых нет смысла включать в таблицу?

Его можно извлечь автоматически из обученного tf-idf vectorizer. В целом хороший набор в nltk. Туда стоит добавить сокращения (ул., г., стр.), а также расширить список служебных частей речи.

Какие слова из списка тематически значимых слов, составленного вручную, вошли в список топ 20 слов по tf.idf, а какие не вошли;

Предполагаемые слова: Куваев, суд, законопроект, изнасилование, судебный, заседание, педофил

Все вошли.

Задайте пороговое значение по tf.idf для ключевых слов;

0.09 или 0.1 - для выбранного текста это достаточные значения для узкого и более широкого списка ключевых слов.

Какие слова, на ваш взгляд, имеют высокий tf.df (выше порогового значения), но не являются ключевыми;

Возможно, *эпизод, трансляция.*

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)

1. Можно использовать стемминг, чтобы объединить сходные слова, которые немного отличаются.
2. Биграммы

1.4

Отличаются ли диаграммы для самых частотных в языке слов и для слов с высоким tf.idf в Вашем списке, если отличаются, то чем?

Частота заметно ниже, болеее “рваные” столбцы barplot. Для частых слов примерно схожая высота столбцов - они равномерно распределяются по подколлекциям.

(см. тетрадь)