Выводы и таблицы приведены в файле HW1.ipynb. Здесь указано основное.

KeyWords:

суд; 371.1

погибнуть; 95.4

свидетель; 55.7

актриса; 40.4

авария; 29.6

виновник; 8.5

ДТП; 5.7

Голуб; 0.0

# Задание 1.1

## Есть ли среди выбранных вами ключевых слов редкие слова?

Да, есть. Во-первых, это **Голуб** (не встретилось в частотном словаре, но мне показалось, что это имя может считаться ключевым, так как может возникать во многих новостных текстах). Во-вторых, с маленькой частотностью можно считать слова **виновник**, **ДТП** (те, что ipm меньше 10).

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

Да, есть. Это слово **суд** (371.1 > 183.9).

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

1 глагол (**погибнуть**), остальные 7 существительные (включая имена собственные: **ДТП**, **Голуб**)

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

**Если говорить о выбранных ключевых словах, ссуд** встретилось в большинстве текстах (196 из 350). Это существительное. А если говорить о словах вообще, то во всех текстах есть служебные части речи и частотные слова (например, местоимения) (см. HW1.ipynb)

# Задание 1.2

Таблица терм-документ находится в HW1.ipynb

Если tf.idf не равно нулю, то тогда слово присутствует в тексте. Поэтому я отфильтровала матрицу tf.idf так что, у первых двух слов было значение не ноль, а у третьего 0. Так нашлись тексты, удовлетворяющие условию Word1&Word2&¬Word3

*Усложнённый вариант*

Я убрала стоп слова, которые встречаются в тексте, поэтому ни одно «значимое» слово не встречается во всех текстах. Если стоп-слова не убирать, то получается так, что даже частотные слова не встречаются во всех текстах.

# Задание 1.3

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

Да, соответствуют. 5 первых слов попали в мои ключевые слова.

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

Нет, не все. Я не выделила слово **газануть**.

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

Во-первых, туда попали глаголы (неназывательные часть речи). Во-вторых, слова с высокой частотностью (по частотному словарю русского языка), например, как **жизнь**, **человек**.

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

Туда должны входить служебные части речи, самые частотные слова (например, эти слова у нас оказались внизу списка, ранжированнного по tf.idf), но это нужно проверить, так как возможно эти слова и могут войти в топ ключевых слов. Также слова , которые никак не определяют тематику текстов (частотные слова, например, местоимения)

*Дополнительное задание*

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

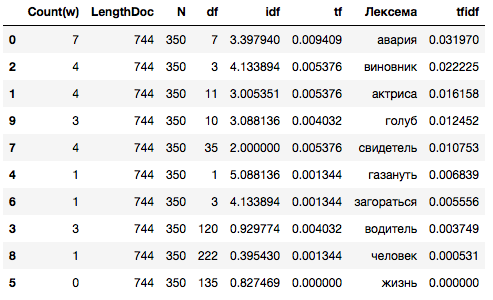
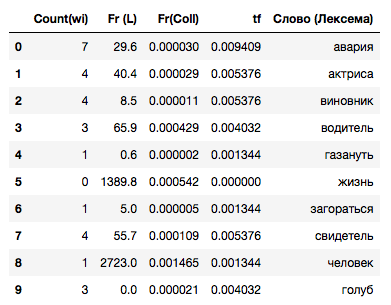
Не вошли **ДТП**, **Голуб**, **суд**. Первые два, возможно, потому что это имена собственные и их частотность мала, или вовсе не встречаются в частотном словаре. Вошли: **свидетель**, **погибать** (я выбрала другую лемму), **актриса**, **авария**, **виновник**, **суд**

## Задайте пороговое значение по tf.idf для ключевых слов

0.007686 (со слова **следствие**)

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

Те, что редки во всей коллекции текста, но ключевыми для этого они тоже не являются (параметр df - маленький: количество документов, в которых встретилось слово)



# Задание 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?

Да, отличаются. Выборка не совсем сбалансирована, поэтому графики частотных слов очень отличаются от слов с высоким tf-idf. Последние встречаются не во всех текстах, но если встречаются, то их частота в этом тексте высокая. Частотные слова появляются в текстах примерно равномерно.

Все диаграммы и другие выводы в файле HW1.ipynb