**УРОВЕНЬ 1**

**файл "5.2 Визуализация данных"**

Графики:

***Диаграммы-линии или графики:*** Для построения диаграмм-линий применяют прямоугольную систему координат. Обычно по оси абсцисс откладывается время (годы, месяцы и т. д.), а по оси ординат – размеры изображаемых явлений или процессов. На осях наносят масштабы.

***Столбчатые и линейные диаграммы (гистограммы):*** Построение столбчатой диаграммы заключается в изображении статистических данных в виде вертикальных прямоугольников или трёхмерных прямоугольных столбиков. Каждый столбик изображает величину уровня данного статистического ряда. Все сравниваемые показатели выражены одной единицей измерения, поэтому удаётся сравнить статистические показатели данного процесса.

***Круговые (секторные) диаграммы:*** Относительная величина каждого значения изображается в виде сектора круга, площадь которого соответствует вкладу этого значения в сумму значений.

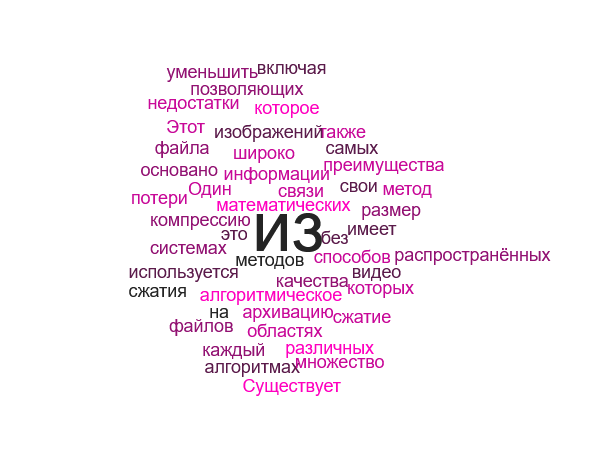
***Радиальные (сетчатые) диаграммы:*** В отличие от линейных диаграмм, в радиальных или сетчатых диаграммах более двух осей. По каждой из них производится отсчёт от начала координат, находящегося в центре. Для каждого типа полученных значений создаётся своя собственная ось, которая исходит из центра диаграммы.

***Таблица:*** организация структуры данных, в которой отдельные элементы помещены в ячейки, каждой из которых сопоставлена пара значений – номер строки и номер колонки.



Облака: Wordcloud:

Wordle:

Word It Out:

tagxedo: ne rabotaet((

Word Cloud by Jason Davies:



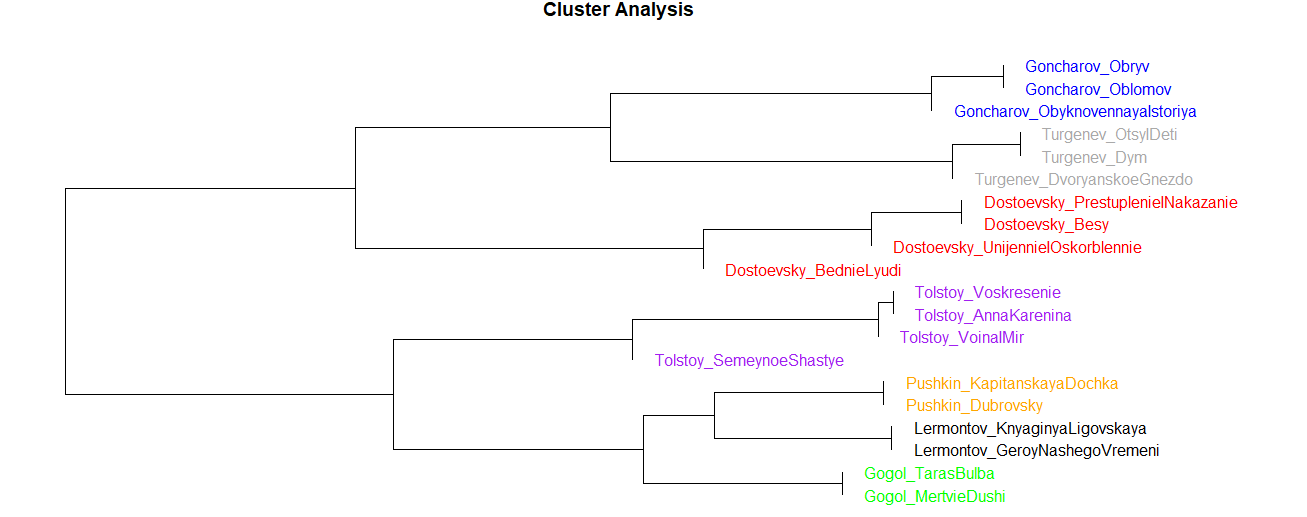
Wordcloud.pro: ne rabotaet((

Tagul: 

Аналоги:

1. библиотека WordCloud для python
2. <https://wordscloud.pythonanywhere.com/>
3. https://wordcloud.online/?ysclid=lve1mo0d71192325670

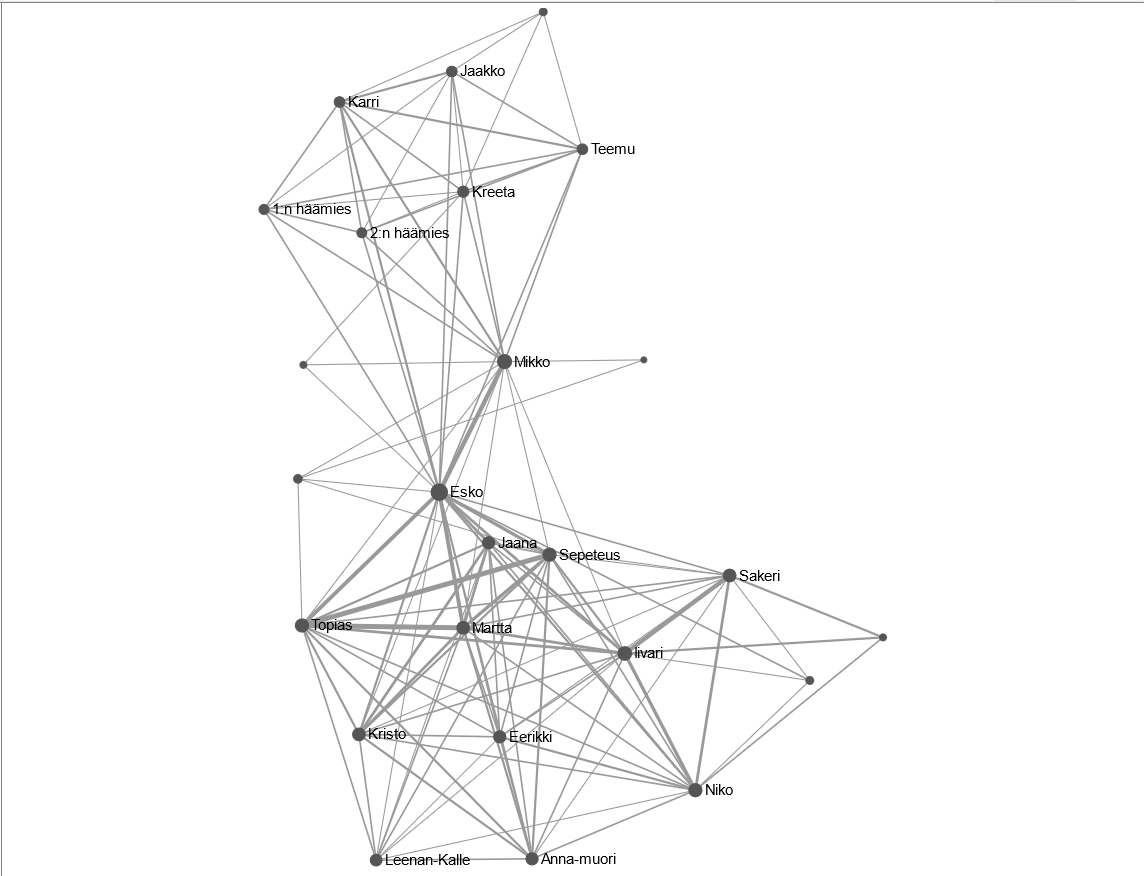
**файл "14. Стилеметрия и тематическое моделирование"**

Для своего стилеметрического исследования я взял текста с гугл диска(<https://drive.google.com/drive/folders/1rc0CzioBpZpud7Re8e_XZHCvN9cUhDW5>) из файлов 1\_pushkin\_lermontov\_gogol.zip и 2\_fourteen\_russian\_novels.zip. 

как видим авторы из первого файла располагаются рядом с собой, в то время как тексты одного автора из второго не всегда совпадают настолько сильно.

**файл “11 Сетевой анализ”**

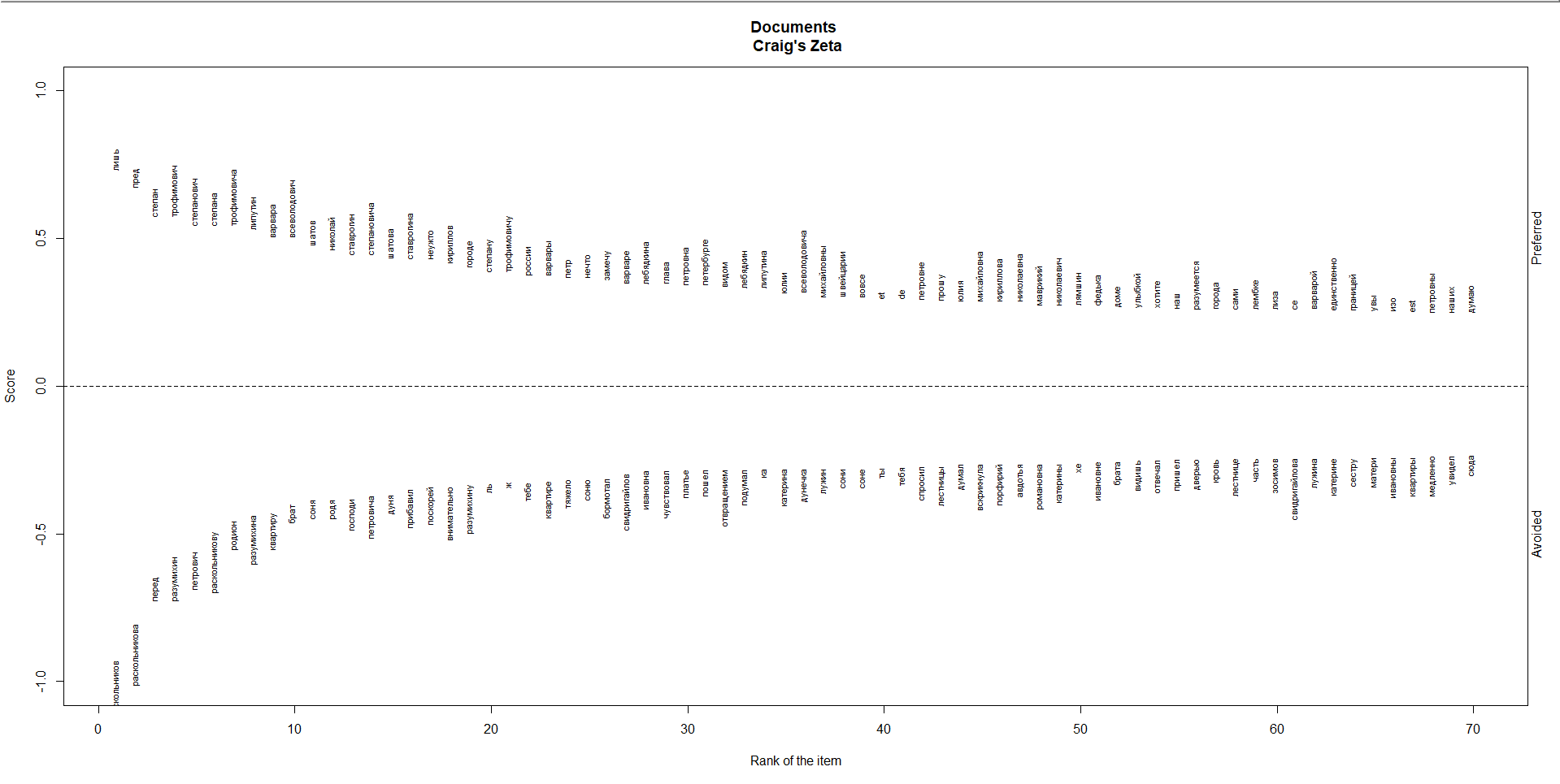
Aleksis Kivi: Nummisuutarit (1864)



**Уровень 2**

**файл "14. Стилеметрия и тематическое моделирование"**

oppose: В качестве двух текстов для сравнения я выбрал “Бесы” и “Преступление и Наказание” Достоевского. Результат можем видеть ниже

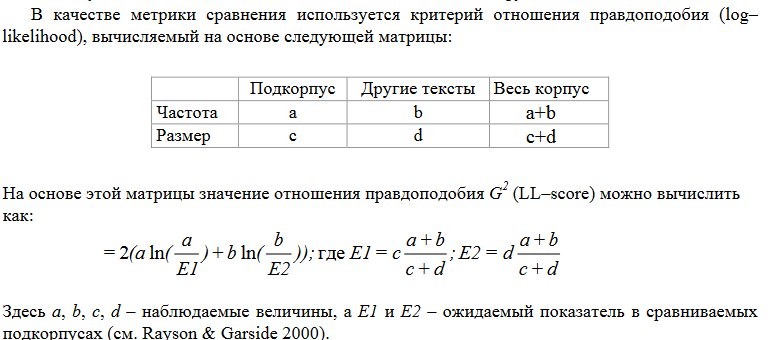


Определения:

1. На основе общего списка лемм, упорядоченного по частоте, леммам присваивается ранг. Самое употребительное слово – и – имеет ранг 1, следующее – в – ранг 2 и т.д., редкие слова имеют ранг 10 000 и больше.
2. относительная частота - во сколько раз чаще слово встречается в подкорпусе по сравнению с остальными текстами.
3. Показатель R (range) отражает количество сегментов корпуса, в которых встретилось слово.
4. Коэффициент Жуйяна отражает равномерность распределения частот в разных сегментах корпуса и вычисляется по следующей формуле:

, где п – количество сегментов, на которые разбит корпус, μ – средняя частота слова по всему

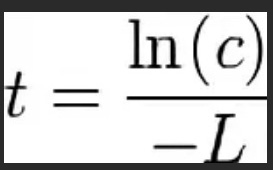
корпусу (т. е. сумма частот в каждом сегменте, поделенная на n), σ – среднее квадратичное отклонение частоты μ на отдельных сегментах.

1. ARF - средняя уменьшенная частота(частота слова взвешивается по расстоянию между отдельными словоупотреблениями)
2. 
3. Критерий Йенсена-Шеннона (или информационный критерий Йенсена-Шеннона) — мера качества статистической модели, основанная на разнице между информацией, которую можно получить из модели, и информацией, которая уже известна. Этот критерий используется для оценки того, насколько хорошо модель может предсказывать данные, основываясь на том, сколько новой информации она предоставляет. Если модель дает мало новой информации, то её качество считается низким. Если же модель даёт много новой информации, то её качество считается высоким.

Граф: сделан и залит на гитпейдж <https://dercomputer.github.io/lingvisticsite/#>

Формулы в глоттохронологии:

Сводеш:

**,** где t - данный период времени от одной стадии языка к другой, c - процент сохранившихся слов в списке к концу этого периода, L - скорость замены для этого списка слов.

Старостин:

, где t - данный период времени от одной стадии языка к другой, c - процент сохранившихся слов в списке к концу этого периода, L - скорость замены для этого списка слов.

**УРОВЕНЬ 3**

**файл "5.2 Визуализация данных"**

25 самых частых слов: man, said, like, don’t, says, get, time, well, would, back, know, make, university, go, first, it’s, new, day, see, asked, people, going, take, could. got.

25 самых частых слов в 1993 году: