Анализ корпуса текстов в научном стиле по радиолокации, газодинамике и наукометрии

Задачи проекта

Задачи проекта

- Загрузить три разных корпуса

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов
  - Распределение косинусного расстояния:

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов
  - Распределение косинусного расстояния:
    - Мешка слов
    - TF-IDF

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов
  - Распределение косинусного расстояния:
    - Мешка слов
    - TF-IDF
  - Возможность определения по косинусному расстоянию:

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов
  - Распределение косинусного расстояния:
    - Мешка слов
    - TF-IDF
  - Возможность определения по косинусному расстоянию:
    - Рубрики журнала

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов
  - Распределение косинусного расстояния:
    - Мешка слов
    - TF-IDF
  - Возможность определения по косинусному расстоянию:
    - Рубрики журнала
    - Перспектив публикации

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов
  - Распределение косинусного расстояния:
    - Мешка слов
    - TF-IDF
  - Возможность определения по косинусному расстоянию:
    - Рубрики журнала
    - Перспектив публикации
    - Рецензентов

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов
  - Распределение косинусного расстояния:
    - Мешка слов
    - TF-IDF
  - Возможность определения по косинусному расстоянию:
    - Рубрики журнала
    - Перспектив публикации
    - Рецензентов
  - Частеречную разметку

Задачи проекта

- Загрузить три разных корпуса
- Проанализировать
  - Частотность слов
  - Распределение косинусного расстояния:
    - Мешка слов
    - TF-IDF
  - Возможность определения по косинусному расстоянию:
    - Рубрики журнала
    - Перспектив публикации
    - Рецензентов
  - Частеречную разметку
  - Коллокации

Корпусы

Характеристика	Корпус		
	Радиолокация	Газодинамика	Наукометрия

Корпусы

Характеристика	Корпус		
	Радиолокация	Газодинамика	Наукометрия
Объем	206	53	8

Корпусы

Характеристика	Корпус		
	Радиолокация	Газодинамика	Наукометрия
Объем	206	53	8
Объем символов	3 704 604	825 349	204 050

Корпусы

Характеристика	Корпус		
	Радиолокация	Газодинамика	Наукометрия
Объем	206	53	8
Объем символов	3 704 604	825 349	204 050
Средняя длина	17 984	15 573	25 506

Корпусы

Характеристика	Корпус		
	Радиолокация	Газодинамика	Наукометрия
Объем	206	53	8
Объем символов	3 704 604	825 349	204 050
Средняя длина	17 984	15 573	25 506
Токены	298 133	70 203	14 829

Корпусы

Характеристика	Корпус		
	Радиолокация	Газодинамика	Наукометрия
Объем	206	53	8
Объем символов	3 704 604	825 349	204 050
Средняя длина	17 984	15 573	25 506
Токены	298 133	70 203	14 829
Средняя длина токенизированного слова (символ)	7,9	8,7	7,7

Корпусы

Характеристика	Корпус		
	Радиолокация	Газодинамика	Наукометрия
Объем	206	53	8
Объем символов	3 704 604	825 349	204 050
Средняя длина	17 984	15 573	25 506
Токены	298 133	70 203	14 829
Средняя длина токенизированного слова (символ)	7,9	8,7	7,7
Средняя длина статьи в токенах	1428	1207	1845

Пример статьи «Радиолокация»

УДК 681.513.6

Построение метода наведения ракеты

с использованием принципа Беллмана

© И. С. Трифонов, 2011

Приведены результаты исследования метода наведения, построенного с использованием опорной модели процесса наведения и принципа Беллмана. Показана возможность построения системы наведения по энергетически выгодным траекториям с увеличением дальности зоны поражения, возможность наведения ракеты на цель при недостаточной информации о цели, возможность наведения на низколетящую цель. Оценены перспективы реализации метода.

Ключевые слова: опорная модель процесса наведения ракеты, принцип Беллмана.

Пример статьи «Газодинамика»

УДК 531.55.011:629.7.076.82

Р. Н. Бухтин

Влияние сферической выемки на траекторию

движения летательного аппарата

Представлены результаты исследования влияния сферической выемки на полет летательного аппарата. С помощью моделирования обтекания выемки построена аналитическая зависимость силы давления на ее поверхности от параметров набегающего потока газа. Рассчитано отклонение летательного аппарата, вызванное наличием сферической выемки, в зависимости от ее положения, начальной скорости полета и угла тангажа.

Ключевые слова: гиперзвуковой летательный аппарат, траектория полета, аэродинамическое сопротивление, моделирование турбулентного течения.

Пример статьи «Наукометрия»

Система массового обслуживания научного журнала

Д. Ю. Большаков https://orcid.org/0000-0001-7694-1454

Концерн воздушно-космической обороны «Алмаз – Антей», г. Москва, Российская Федерация

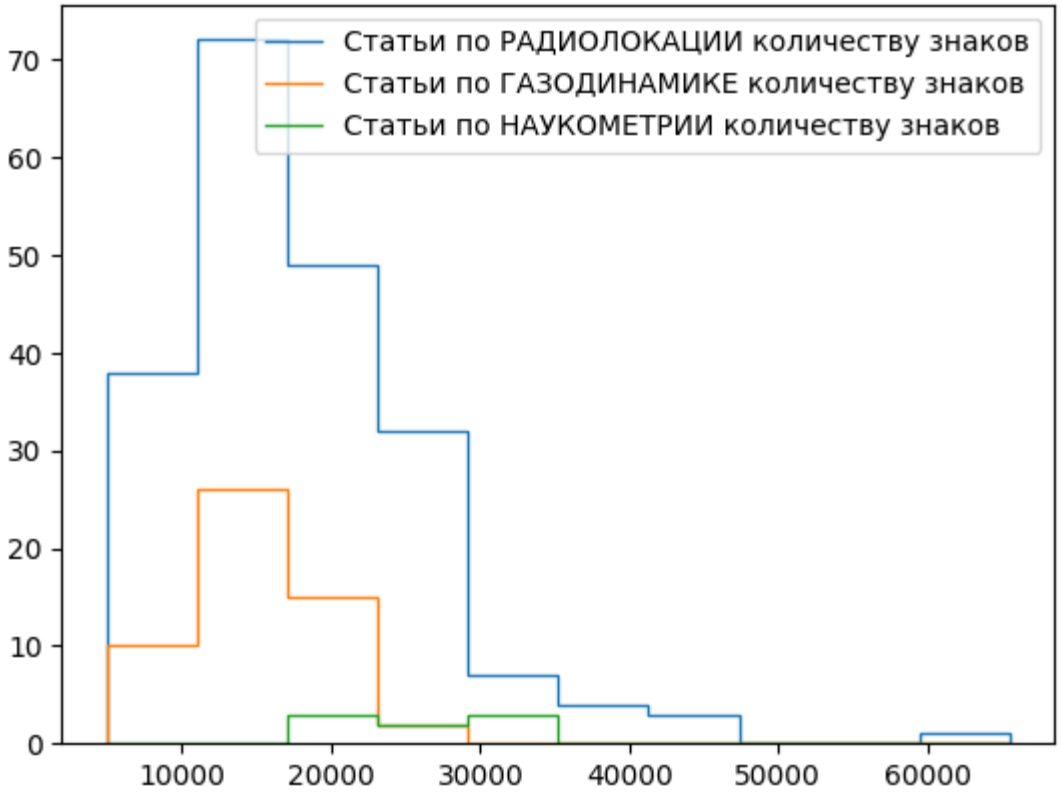
press@almaz-antey.ru

Резюме. Приведены результаты исследования разности календарных дат окончания и начала редакционных процессов в научно-техническом журнале «Вестник Концерна ВКО «Алмаз – Антей», а также количество дат в месяц для начала процессов (поступление статей, передача на литературное редактирование и т. д.) в месяц.

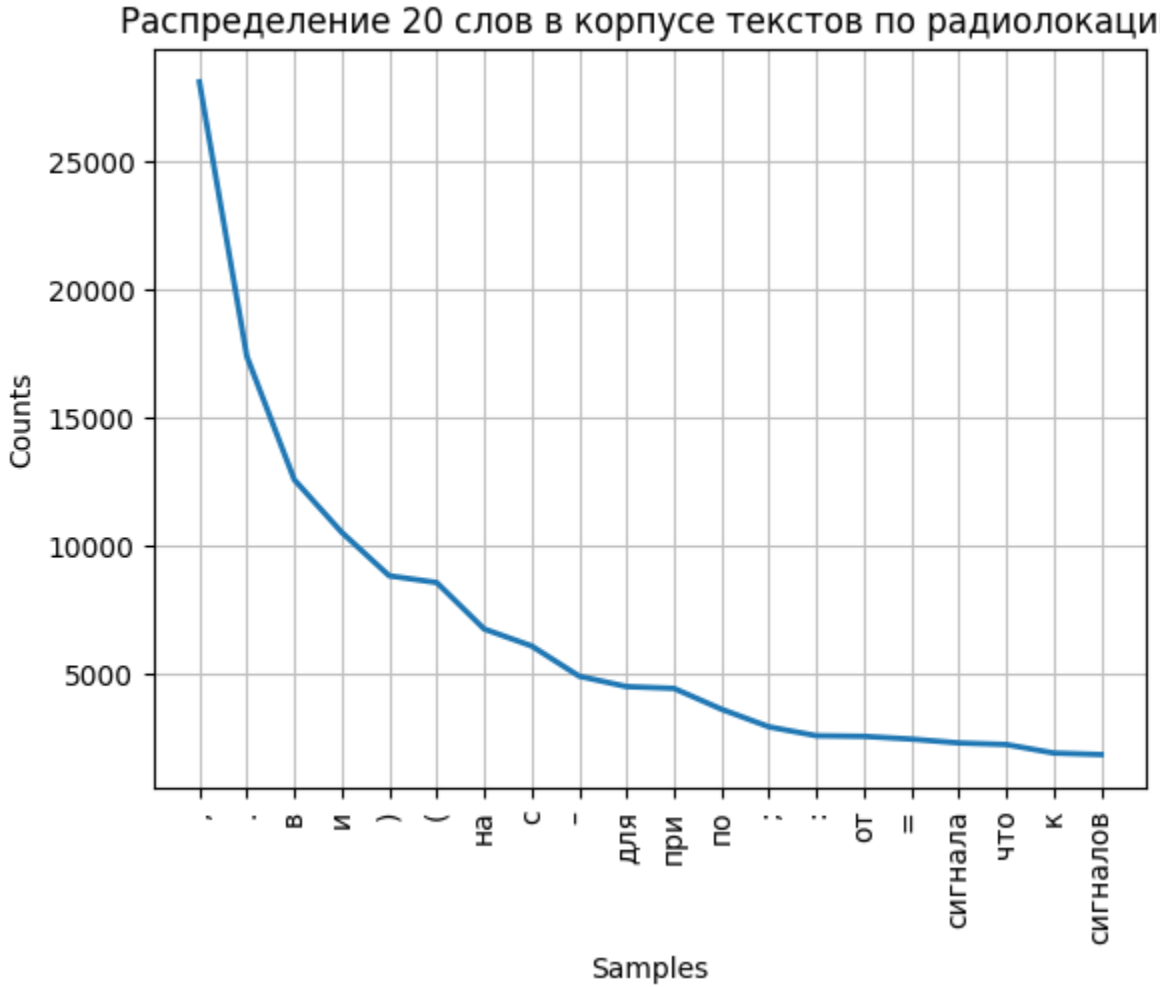
На основании анализа сделан вывод, что поступающая на публикацию статья может рассматриваться как заявка в систему массового обслуживания, которой является научный журнал. И для такой заявки могут быть оценены функции распределения входных потоков, средние и максимальные сроки нахождения в системе, а также функции распределения времени обработки и другие характеристики, которые остаются неизменными (стационарными) во времени для данного научного журнала.

Ключевые слова: научный журнал, система массового обслуживание, обслуживания, временные издержки, редакционные процессы, распределение случайной величины

Гистограмма объема корпусов



Частотный анализ (исходный)



Частотный анализ (исходный)

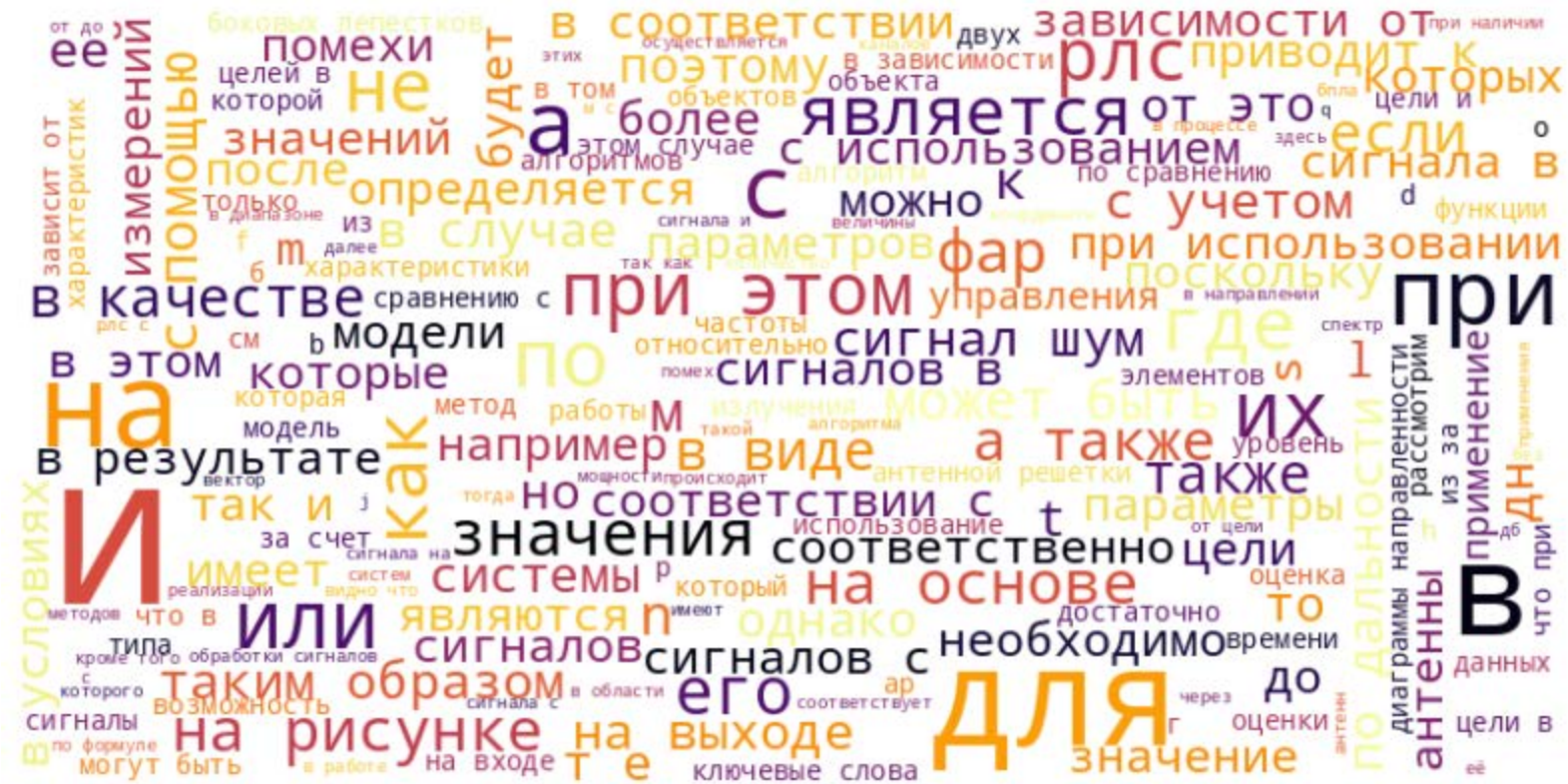
Частотный анализ (исходный)

Частотный анализ (вычищенный)

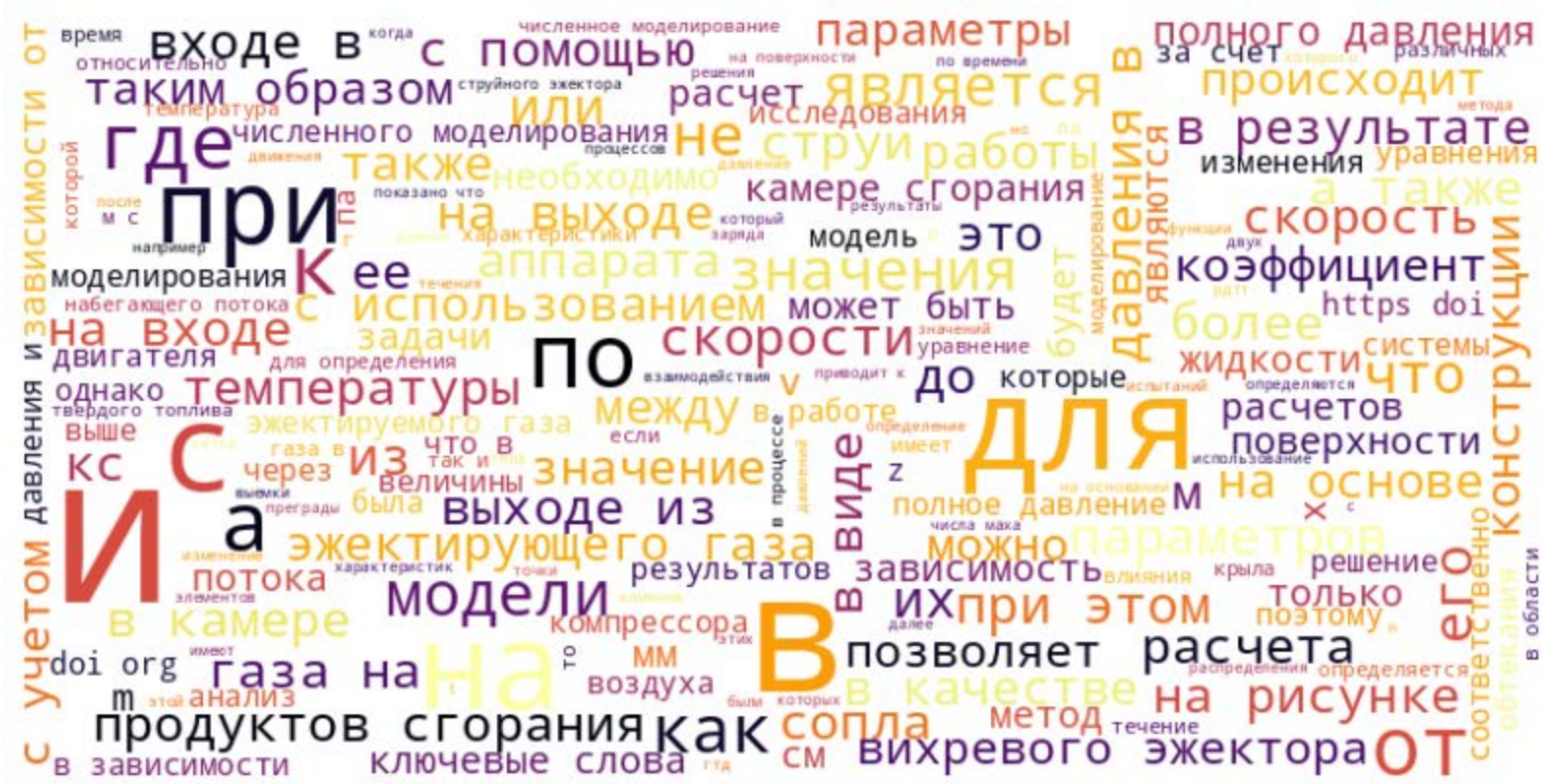
Частотный анализ (вычищенный)

Частотный анализ (вычищенный)

Облако слов (исходное) «радиолокация»

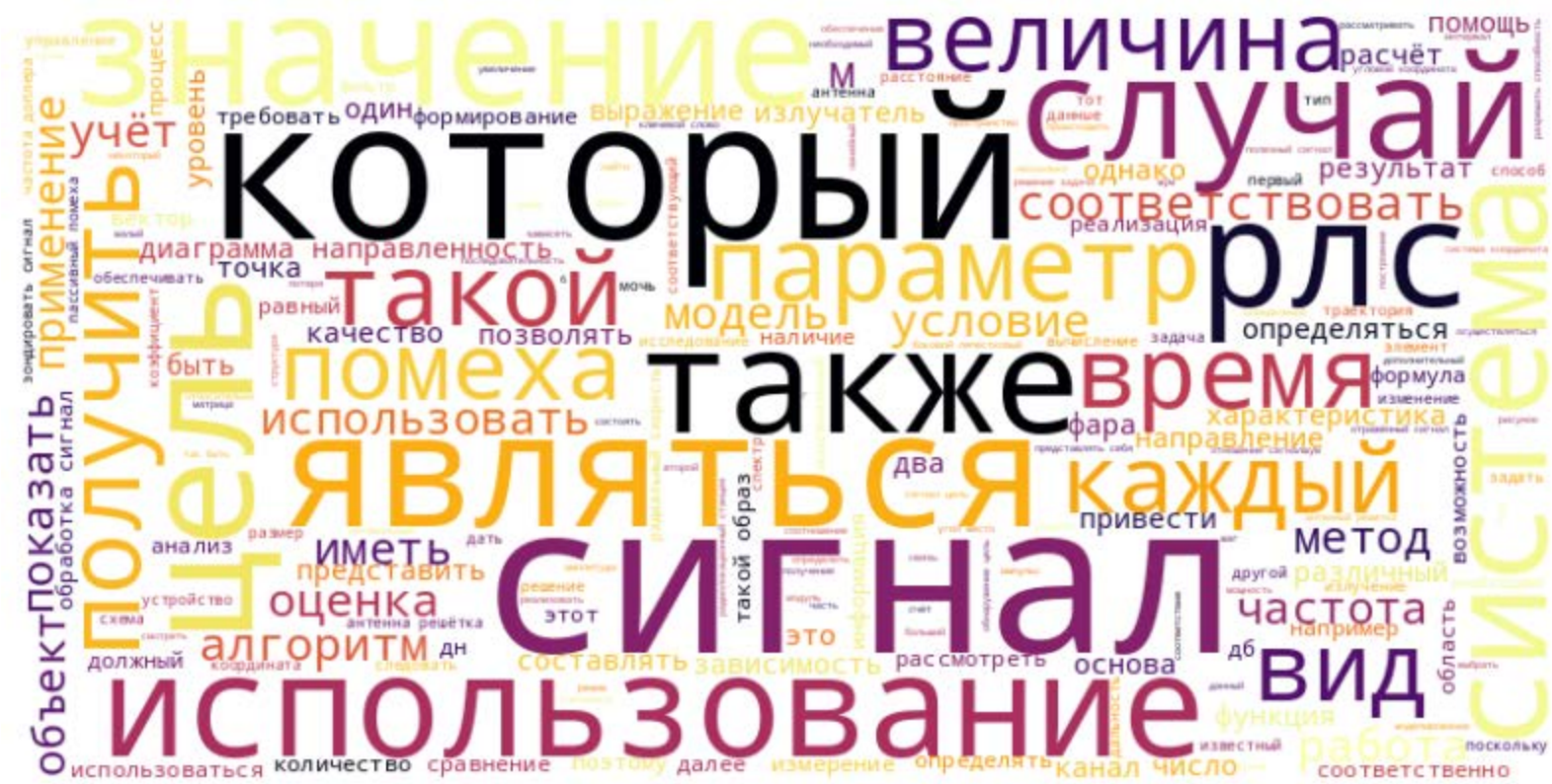


Облако слов (исходное) «газодинамика»

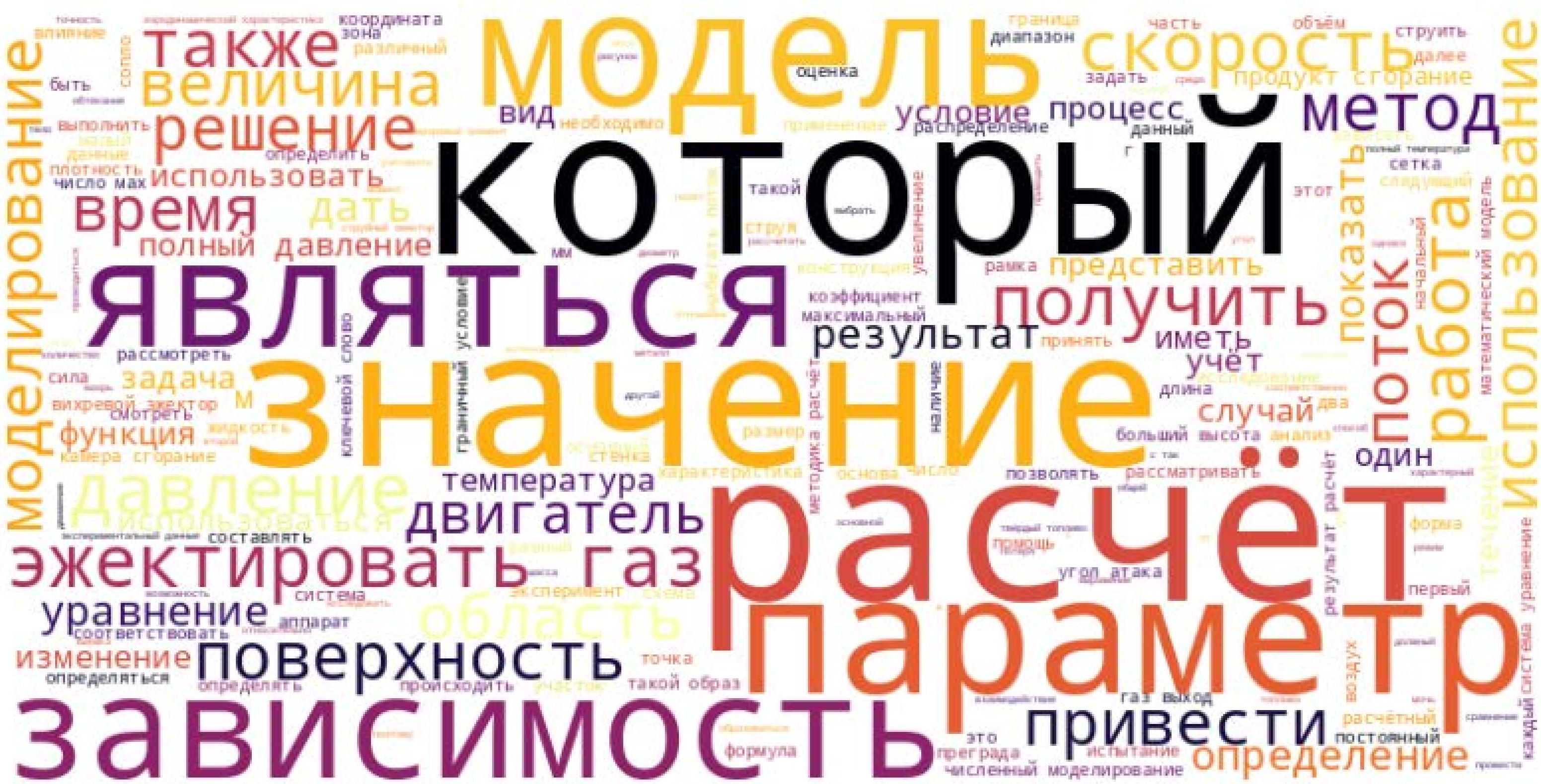


Облако слов (исходное) «наукометрия»

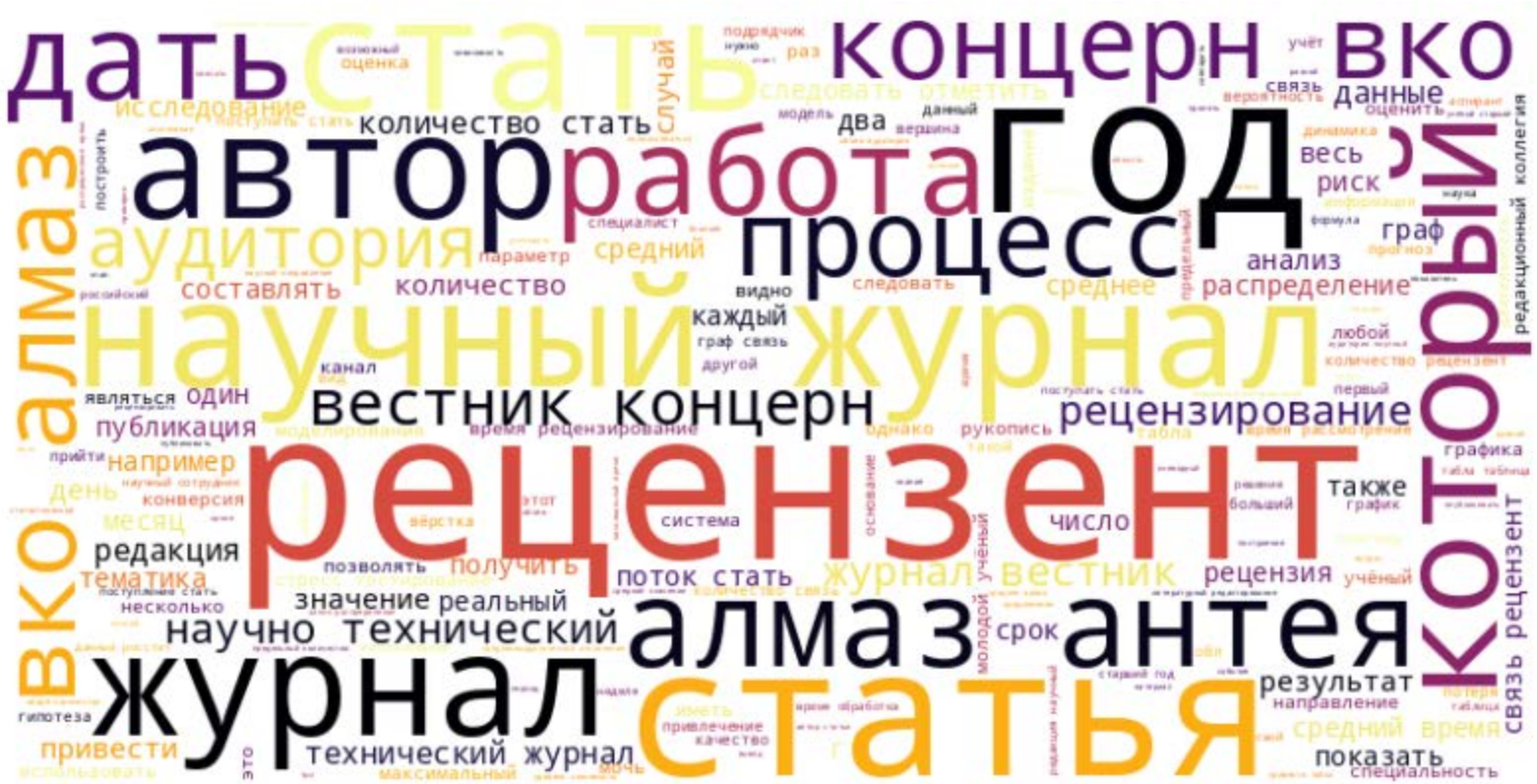
Облако слов (очищенное) «радиолокация»



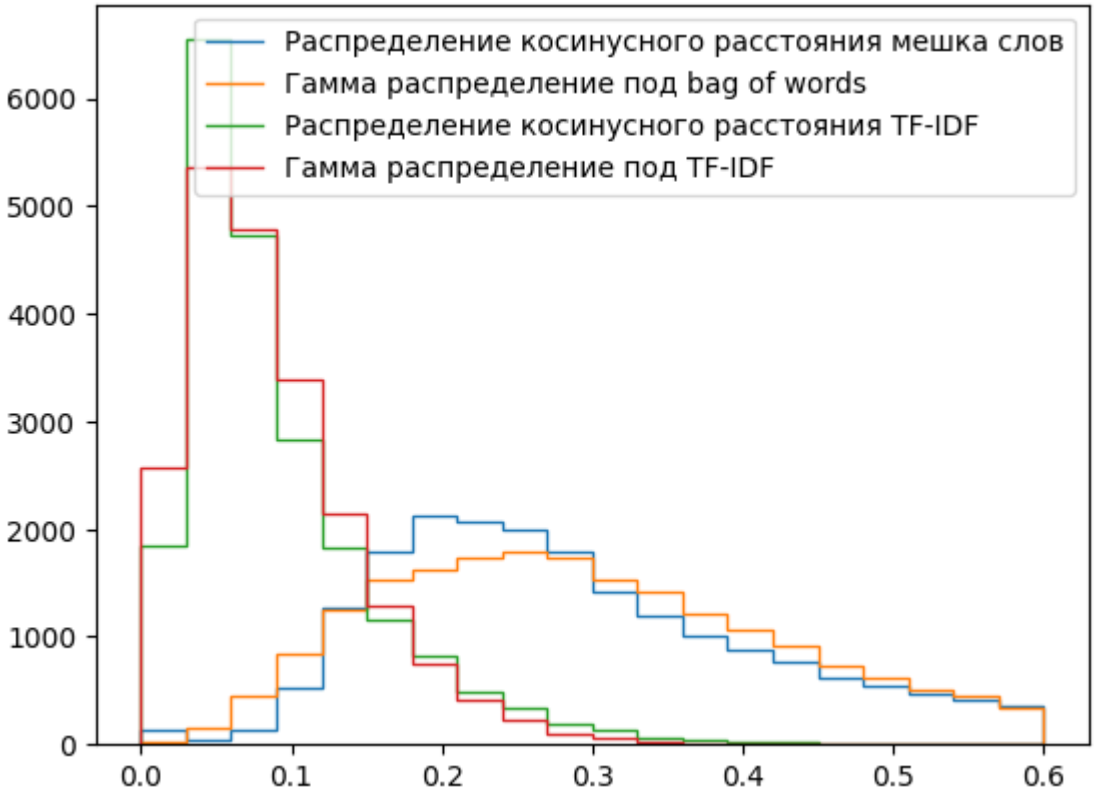
Облако слов (очищенное) «газодинамика»



Облако слов (очищенное) «наукометрия»



Распределение косинусных расстояний (радиолокация)



Количество сочетаний

$$C_n^m = \frac{n!}{(n-m)!\,m!}$$

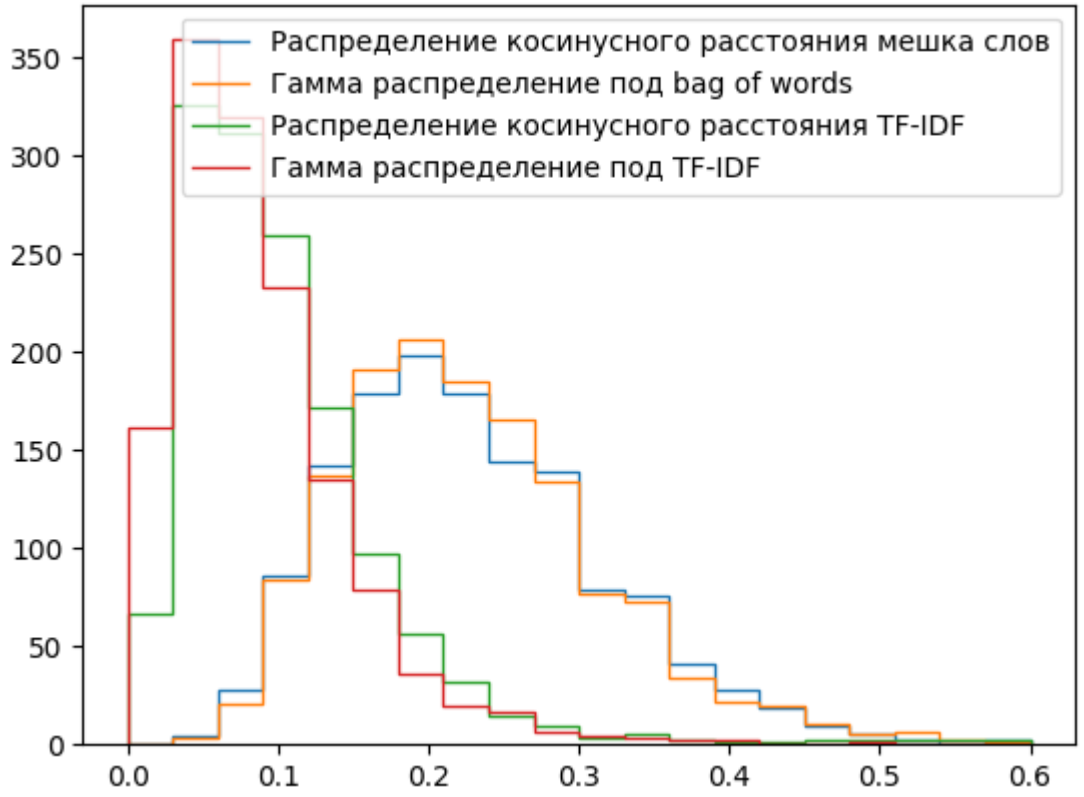
Количество сочетаний

$$C_n^m = \frac{n!}{(n-m)!\,m!}$$

$$C_{207}^2 = \frac{207!}{(207-2)!\,2!} =$$

$$= 207\,\frac{206}{2} = 21\,321$$

Распределение косинусных расстояний (газодинамика)



Количество сочетаний

$$C_n^m = \frac{n!}{(n-m)!\,m!}$$

$$C_{53}^2 = \frac{53!}{(53-2)!\,2!} =$$

$$= 53\frac{52}{2} = 1\,378$$

Распределение косинусного расстояния

$$\Gamma(x) = \begin{cases} x^{k-1}\,\dfrac{e^{-x/\theta}}{\theta^{k}\Gamma(k)}, x \geq 0 \\ 0, x < 0 \end{cases}$$

Г(*x*) – гамма функция

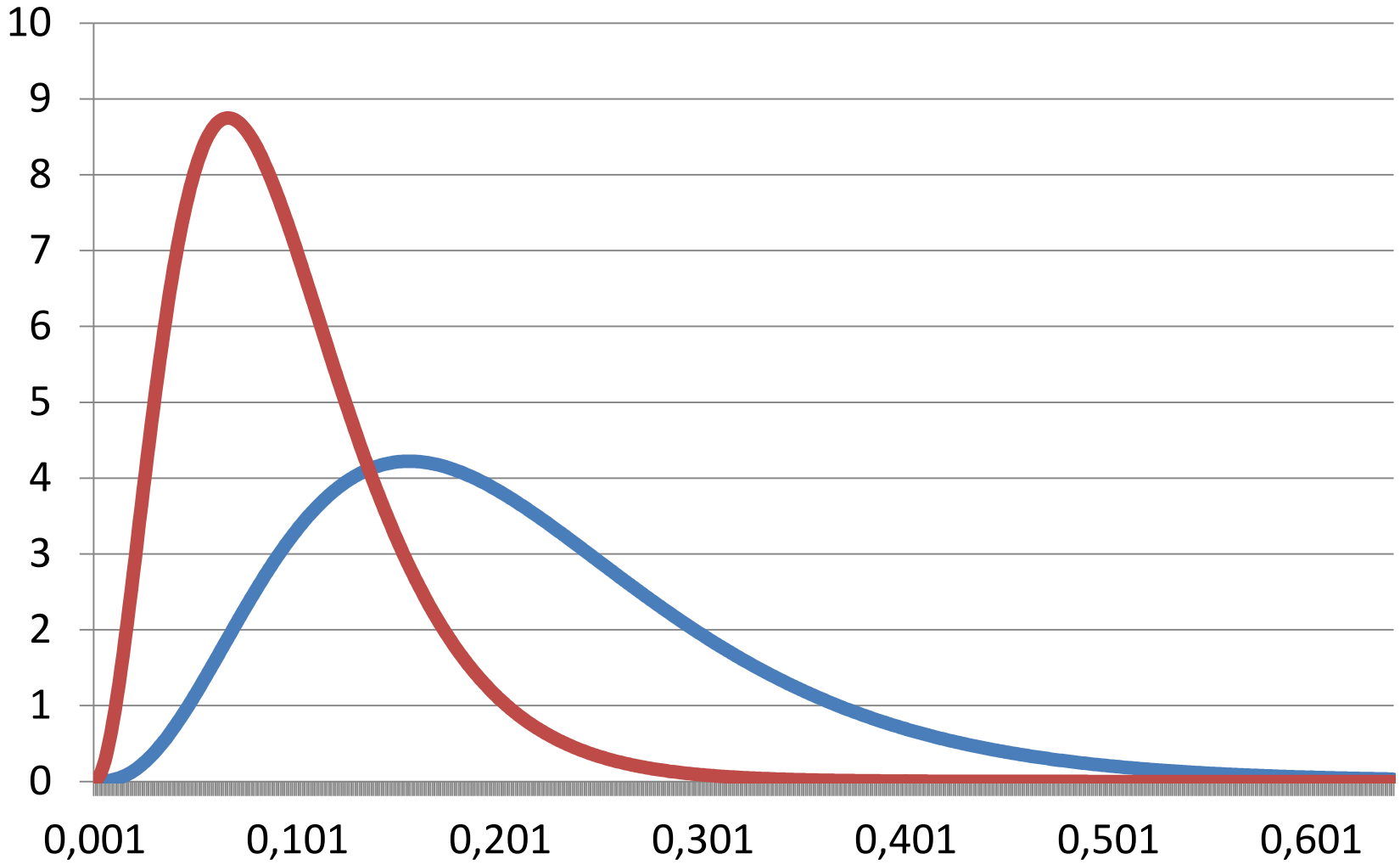
$\theta = Var(x)/E(x)$

$k = E^2(x)/Var(x)$

E(x) – математическое ожидание

Var(x) - дисперсия

Гамма распределение



Косинусное расстояние мешка слов

Корпус	Статьи		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,46 (0,04…0,72)**	0,26 (0,10…0,40)	0,21 (0,14…0,27)

Косинусное расстояние мешка слов

Корпус	Статьи		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,46 (0,04…0,72)**	0,26 (0,10…0,40)	0,21 (0,14…0,27)
Газодинамика	0,24 (0,01…0,41)	**0,49 (0,26…0,66)**	0,21 (0,13…0,30)

Косинусное расстояние мешка слов

Корпус	Статьи		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,46 (0,04…0,72)**	0,26 (0,10…0,40)	0,21 (0,14…0,27)
Газодинамика	0,24 (0,01…0,41)	**0,49 (0,26…0,66)**	0,21 (0,13…0,30)
Наукометрия	0,13 (0,01…0,23)	0,14 (0,05…0,21)	**0,75 (0,69…0,86)**

Косинусное расстояние мешка слов первой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,58**	0,18	0,14

Косинусное расстояние мешка слов первой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,58**	0,18	0,14
Газодинамика	0,29	**0,47**	0,14

Косинусное расстояние мешка слов первой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,58**	0,18	0,14
Газодинамика	0,29	**0,47**	0,14
Наукометрия	0,19	0,10	**0,75**

Косинусное расстояние мешка слов восьмой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,40**	0,32	0,22

Косинусное расстояние мешка слов восьмой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,40**	0,32	0,22
Газодинамика	0,24	**0,60**	0,23

Косинусное расстояние мешка слов восьмой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,40**	0,32	0,22
Газодинамика	0,24	**0,60**	0,23
Наукометрия	0,14	0,18	**0,73**

Косинусное расстояние мешка слов статьи с минимальным значением косинусного расстояния 0,04 (статья 200)

Корпус	Статья
	Радиолокация
Радиолокация	0,048

Косинусное расстояние мешка слов статьи с минимальным значением косинусного расстояния 0,04 (статья 200)

Корпус	Статья
	Радиолокация
Радиолокация	0,048
Газодинамика	0,017

Косинусное расстояние мешка слов статьи с минимальным значением косинусного расстояния 0,04 (статья 200)

Корпус	Статья
	Радиолокация
Радиолокация	0,048
Газодинамика	0,017
Наукометрия	0,018

Косинусное расстояние TF-IDF

Корпус	Статьи		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,30 (0,05…0,56)**	0,1 (0,03…0,19)	0,12 (0,08…0,15)

Косинусное расстояние TF-IDF

Корпус	Статьи		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,30 (0,05…0,56)**	0,1 (0,03…0,19)	0,12 (0,08…0,15)
Газодинамика	0,08 (0,004…0,17)	**0,34 (0,19…0,47)**	0,12 (0,07…0,18)

Косинусное расстояние TF-IDF

Корпус	Статьи		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,30 (0,05…0,56)**	0,1 (0,03…0,19)	0,12 (0,08…0,15)
Газодинамика	0,08 (0,004…0,17)	**0,34 (0,19…0,47)**	0,12 (0,07…0,18)
Наукометрия	0,03 (0,004…0,09)	0,04 (0,01…0,07)	**0,67 (0,61…0,79)**

Косинусное расстояние TF-IDF первой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,34**	0,07	0,14

Косинусное расстояние TF-IDF первой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,34**	0,07	0,14
Газодинамика	0,29	**0,47**	0,14

Косинусное расстояние TF-IDF первой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,34**	0,07	0,14
Газодинамика	0,29	**0,47**	0,14
Наукометрия	0,19	0,02	**0,75**

Косинусное расстояние TF-IDF восьмой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,24**	0,13	0,13

Косинусное расстояние TF-IDF восьмой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,24**	0,13	0,13
Газодинамика	0,06	**0,39**	0,15

Косинусное расстояние TF-IDF восьмой статьи корпуса

Корпус	Статья		
	Радиолокация	Газодинамика	Наукометрия
Радиолокация	**0,24**	0,13	0,13
Газодинамика	0,06	**0,39**	0,15
Наукометрия	0,03	0,05	**0,64**

Косинусное расстояние TF-IDF статьи с минимальным значением косинусного расстояния 0,04 (статья 200)

Корпус	Статья
	Радиолокация
Радиолокация	0,054

Косинусное расстояние TF-IDF статьи с минимальным значением косинусного расстояния 0,04 (статья 200)

Корпус	Статья
	Радиолокация
Радиолокация	0,054
Газодинамика	0,004

Косинусное расстояние TF-IDF статьи с минимальным значением косинусного расстояния 0,04 (статья 200)

Корпус	Статья
	Радиолокация
Радиолокация	0,054
Газодинамика	0,004
Наукометрия	0,004

Перспективы публикации

- Косинусное расстояние распределено гаммой-функцией и по мешку слов и по TF-IDF

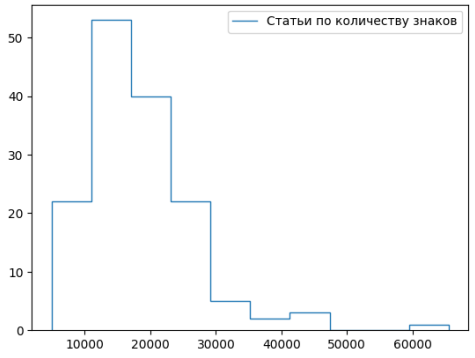
Перспективы публикации

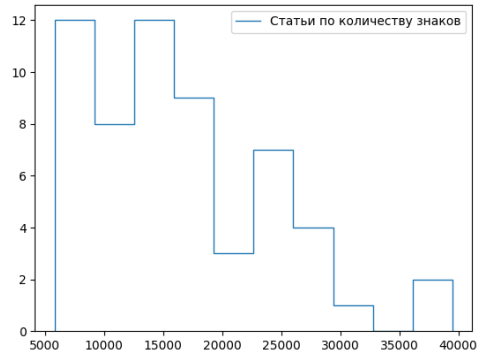
- Косинусное расстояние распределено гаммой-функцией и по мешку слов и по TF-IDF
- Определить статистический критерий по уровню значимости гамма-функции не получилось

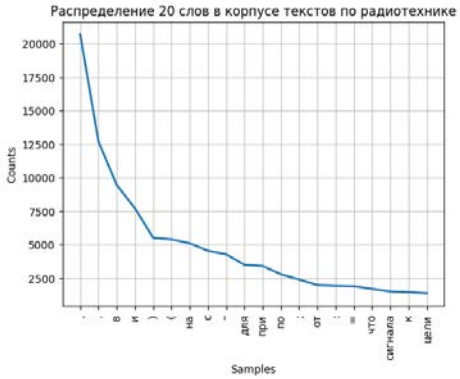
Перспективы публикации

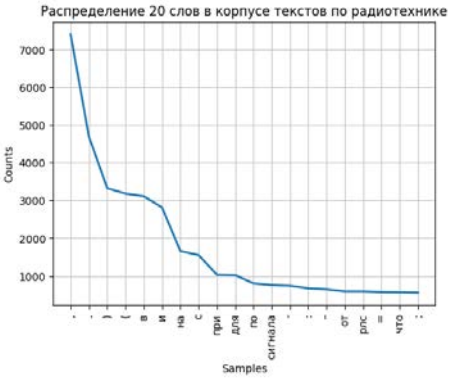
- Косинусное расстояние распределено гаммой-функцией и по мешку слов и по TF-IDF
- Определить статистический критерий по уровню значимости гамма-функции не получилось
- Распределение слов аналогичное у публикуемой и не публикуемой статьи одинаковое

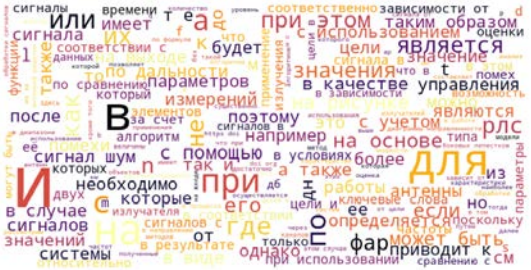
Сравнение

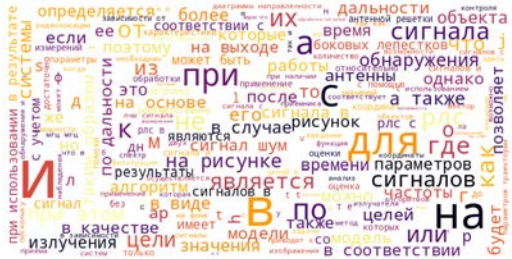

Статьи по количеству знаков


Статьи по количеству знаков


Распределение 20 слов в корпусе текстов по радиотехнике
Counts
Samples

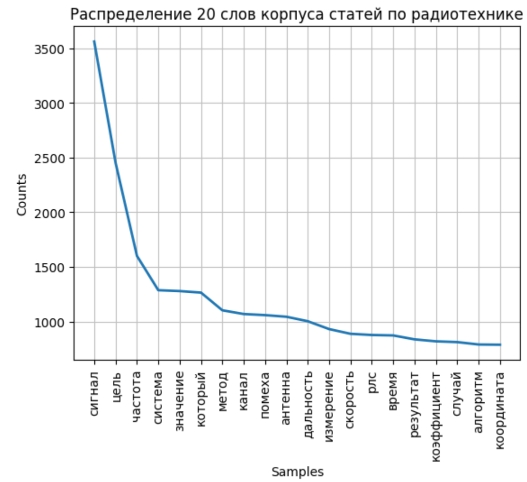

Распределение 20 слов в корпусе текстов по радиотехнике
Counts
Samples

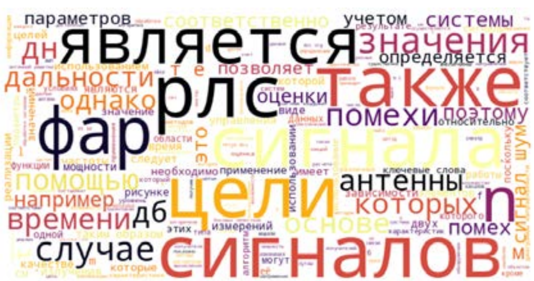


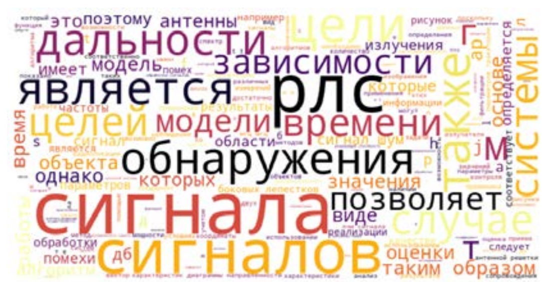


Сравнение

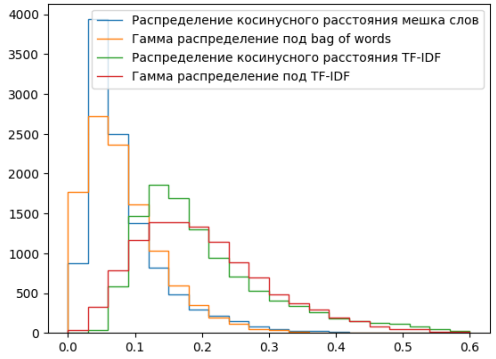

Распределение 20 слов корпуса статей по радиотехнике
Counts
1600
1400
1200
1000
800
600
400
сигнал
.
цель
рлс
частота
система
который
параметр
обработка
значение
обнаружение
алгоритм
фильтр
рисунок
антенна
работа
время
помеха
метод
результат
Samples

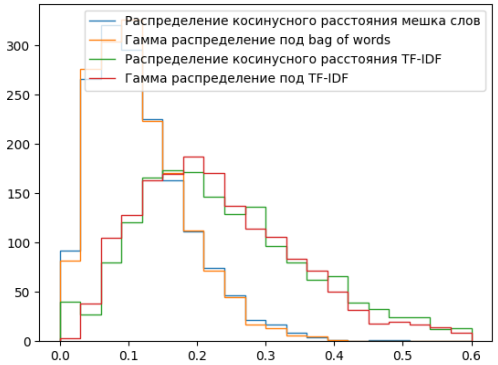

Распределение 20 слов корпуса статей по радиотехнике
Counts
3500
3000
2500
2000
1500
1000
сигнал
цель
частота
система
значение
который
метод
канал
помеха
антенна
дальность
измерение
скорость
рлс
время
результат
коэффициент
случай
алгоритм
координата
Samples

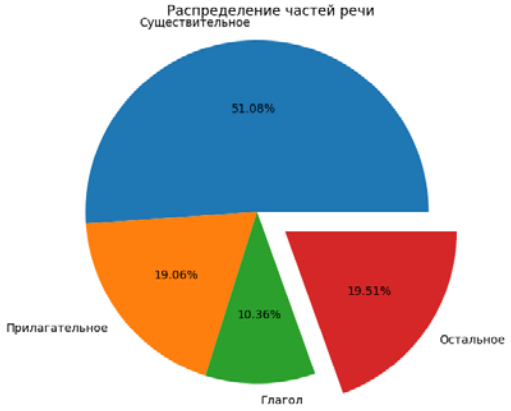

параметров
учетом
системы
является
значения
определяется
дальности
также
рлс
однако
оценки
фар
помехи
поэтому
сигнала
помощью
цели
антенны
например
которых
времени
дб
помех
случае
сигналов


это
поэтому
антенны
цели
дальности
зависимости
является
рлс
также
целей
модели
времени
системы
обнаружения
однако
позволяет
сигнала
случае
сигналов
оценки
таким
образом

Сравнение


Распределение косинусного расстояния мешка слов
Гамма распределение под bag of words
Распределение косинусного расстояния TF-IDF
Гамма распределение под TF-IDF
4000
3500
3000
2500
2000
1500
1000
500
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6


Распределение косинусного расстояния мешка слов
Гамма распределение под bag of words
Распределение косинусного расстояния TF-IDF
Гамма распределение под TF-IDF
300
250
200
150
100
50
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6


Распределение частей речи
Существительное
51.08%
19.06%
Прилагательное
10.36%
Глагол
19.51%
Остальное


Распределение частей речи
Существительное
50.03%
17.63%
Прилагательное
9.67%
Глагол
22.67%
Остальное

Определение рецензента по максимуму косинусного расстояния мешка слов

[0, 1]
[1, 180]
[2, 161]
[3, 87]
[4, 36]
[5, 93]
[6, 2]
[7, 151]

Определение рецензента по максимуму косинусного расстояния мешка слов

[0, 1] **Каплин, Снежинский :**
[1, 180] **Верхоглядов, Каплин, Снежинский :**
[2, 161] **Толоконников, Каплин, Демьянец :**
[3, 87] **Остроженков, Елахин, Ножнин, Порошин :**
[4, 36] **Верхоглядов, Снежинский :**
[5, 93] **Каплин, Корабельников :**
[6, 2] **Гамов, Елахин, Остроженков, Сапельников :**
[7, 151]

Определение рецензента по максимуму косинусного расстояния мешка слов

[0, 1]	**Каплин, Снежинский :**	Верхоглядов Каплин Снежинский
[1, 180]	**Верхоглядов, Каплин, Снежинский :**	Каплин Крыловатых Синяков
[2, 161]	**Толоконников, Каплин, Демьянец :**	Работин Каплин Воронин
[3, 87]	**Остроженков, Елахин, Ножнин, Порошин :**	Дубов Высотин Остроженков
[4, 36]	**Верхоглядов, Снежинский :**	Сажин Ивушкин
[5, 93]	**Каплин, Корабельников :**	Толоконников Каплин Демьянец
[6, 2]	**Гамов, Елахин, Остроженков, Сапельников :**	Гамов Остроженков Елахин
[7, 151]		

Определение рецензента по максимуму косинусного расстояния мешка слов

[0, 1] **Каплин, Снежинский :** Каплин Снежинский

[1, 180] **, Каплин, :** Каплин

[2, 161] **, Каплин, :** Каплин

[3, 87] **Остроженков, , , :** Остроженков

[4, 36] **, :**

[5, 93] **Каплин, :** Каплин

[6, 2] **Гамов, Елахин, Остроженков, :** Гамов Остроженков Елахин

[7, 151]

Определение рецензента по максимуму косинусного расстояния мешка слов

[0, 1] **Каплин, Снежинский :** Каплин Снежинский

[1, 180] **, Каплин, :** Каплин

[2, 161] **, Каплин, :** Каплин

[3, 87] **Остроженков, , , :** Остроженков

[4, 36] **, :**

[5, 93] **Каплин, :** Каплин

[6, 2] **Гамов, Елахин, Остроженков, :** Гамов Остроженков Елахин

[7, 151]

Косинусное расстояние мешка слов:

Радиолокация 55% (72%)

Газодинамика 85% (96%)

Определение рецензента по максимуму косинусного расстояния TF-IDF

[0, 135]
[1, 180]
[2, 161]
[3, 74]
[4, 36]
[5, 93]
[6, 2]
[7, 145]

Определение рецензента по максимуму косинусного расстояния TF-IDF

[0, 135] **Каплин, Снежинский :**
[1, 180] **Верхоглядов, Каплин, Снежинский :**
[2, 161] **Толоконников, Каплин, Демьянец :**
[3, 74] **Остроженков, Елахин, Ножнин, Порошин :**
[4, 36] **Верхоглядов, Снежинский :**
[5, 93] **Каплин, Корабельников :**
[6, 2] **Гамов, Елахин, Остроженков, Сапельников :**
[7, 145]

Определение рецензента по максимуму косинусного расстояния TF-IDF

[0, 135]	**Каплин, Снежинский :**	Сажин Каплин Демьянец
[1, 180]	**Верхоглядов, Каплин, Снежинский :**	Каплин Крыловатых Синяков
[2, 161]	**Толоконников, Каплин, Демьянец :**	Работин Каплин Воронин
[3, 74]	**Остроженков, Елахин, Ножнин, Порошин :**	Толоконников Елахин Гамов
[4, 36]	**Верхоглядов, Снежинский :**	Сажин Ивушкин
[5, 93]	**Каплин, Корабельников :**	Толоконников Каплин Демьянец
[6, 2]	**Гамов, Елахин, Остроженков, Сапельников :**	Синяков Работин Остроженков
[7, 145]		

Определение рецензента по максимуму косинусного расстояния TF-IDF

[0, 135] **Каплин,** : Каплин
[1, 180] , **Каплин,** : Каплин
[2, 161] , **Каплин,** : Каплин
[3, 74]
[4, 36] , **Елахин,** , : Елахин
[5, 93] , :
[6, 2] **Каплин,** : Каплин
[7, 145] , , **Остроженков,** : Остроженков

Определение рецензента по максимуму косинусного расстояния TF-IDF

[0, 135] **Каплин,** : Каплин
[1, 180] , **Каплин,** : Каплин
[2, 161] , **Каплин,** : Каплин
[3, 74] , **Елахин,** , : Елахин
[4, 36] , :
[5, 93] **Каплин,** : Каплин
[6, 2] , , **Остроженков,** : Остроженков
[7, 145]

Косинусное расстояние TF-IDF:
Радиолокация 56% (66%)
Газодинамика 76% (96%)

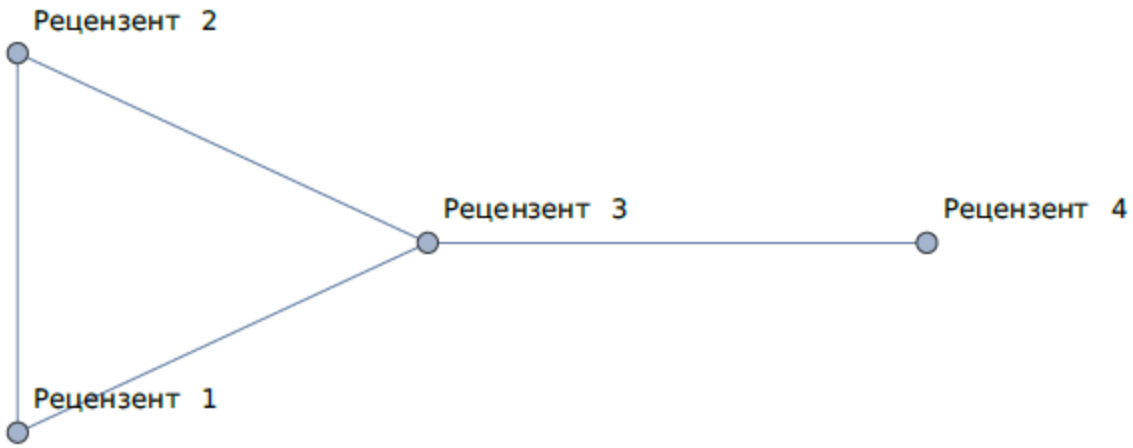
Определение рецензента по максимуму косинусного расстояния TF-IDF

[0, 135] **Каплин,** : Каплин
[1, 180] , **Каплин,** : Каплин
[2, 161] , **Каплин,** : Каплин
[3, 74] , **Елахин,** , : Елахин
[4, 36] , :
[5, 93] **Каплин,** : Каплин
[6, 2] , , **Остроженков,** : Остроженков
[7, 145]

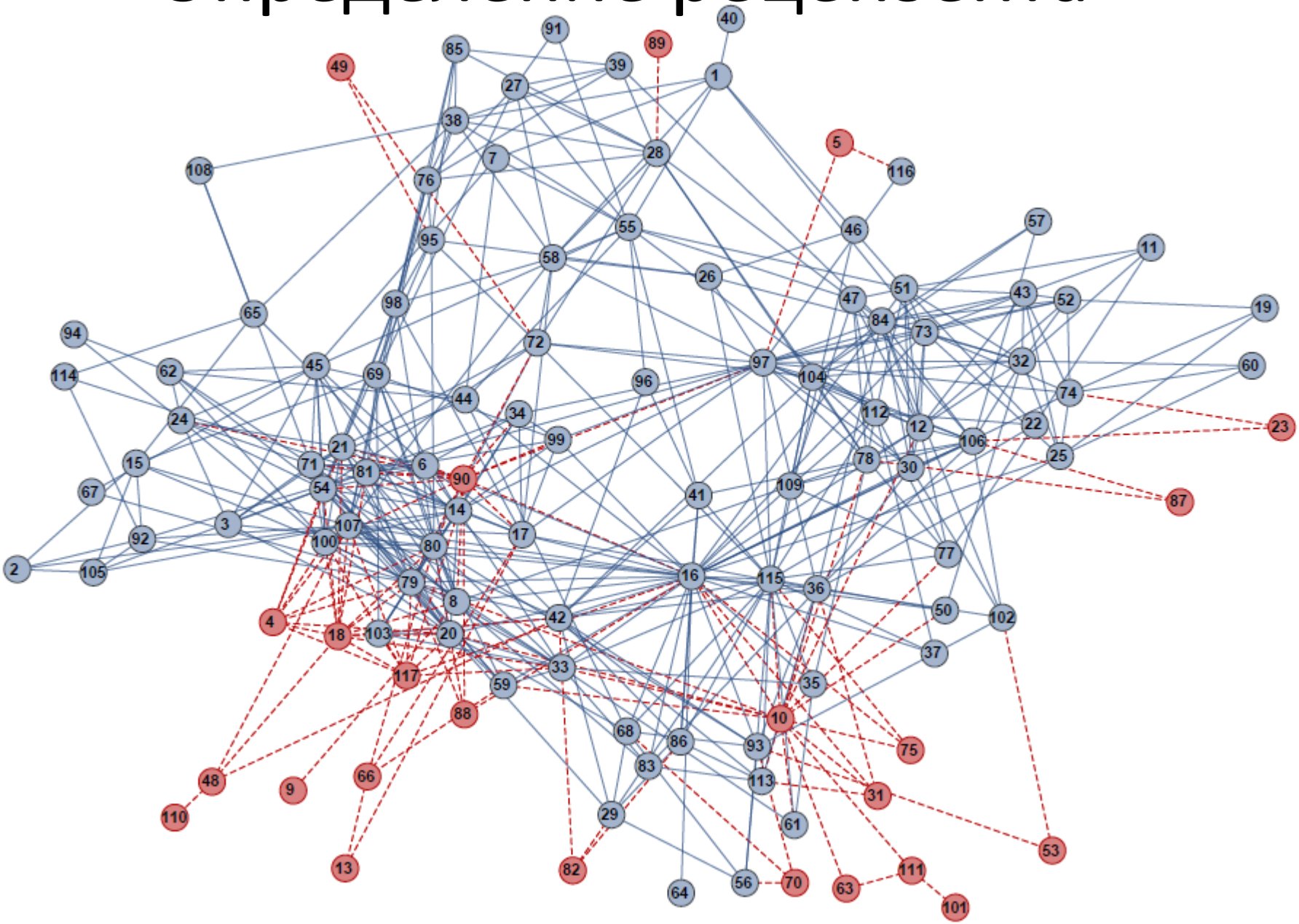
Косинусное расстояние мешка слов:
Радиолокация 55% (72%)
Газодинамика 85% (96%)

Косинусное расстояние TF-IDF:
Радиолокация 56% (66%)
Газодинамика 76% (96%)

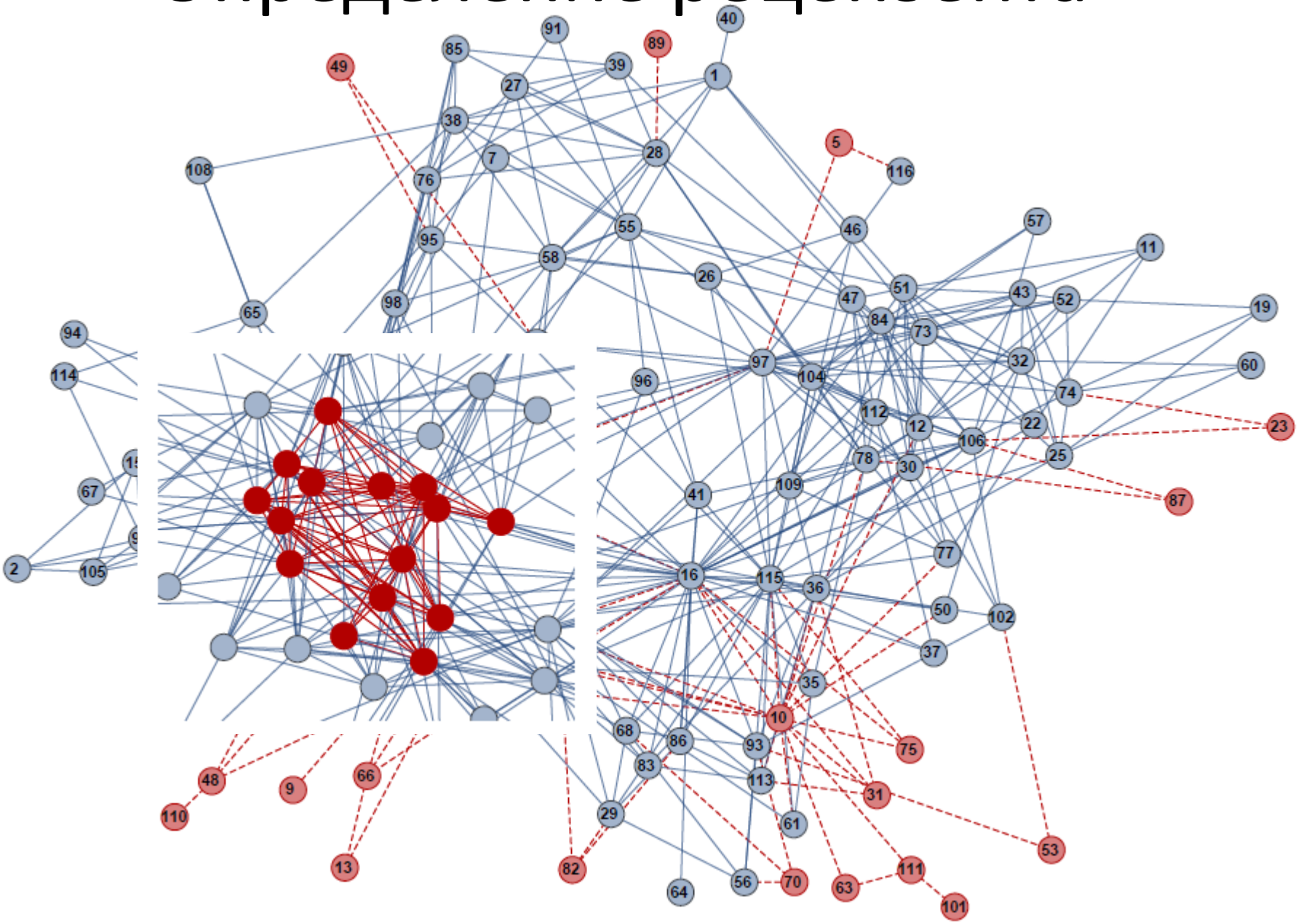
Определение рецензента



Определение рецензента



Определение рецензента



Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF

Bag of words

[0, 1]
[1, 180]
[2, 161]
[3, 87]
[4, 36]
[5, 93]
[6, 2]
[7, 151]

Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF

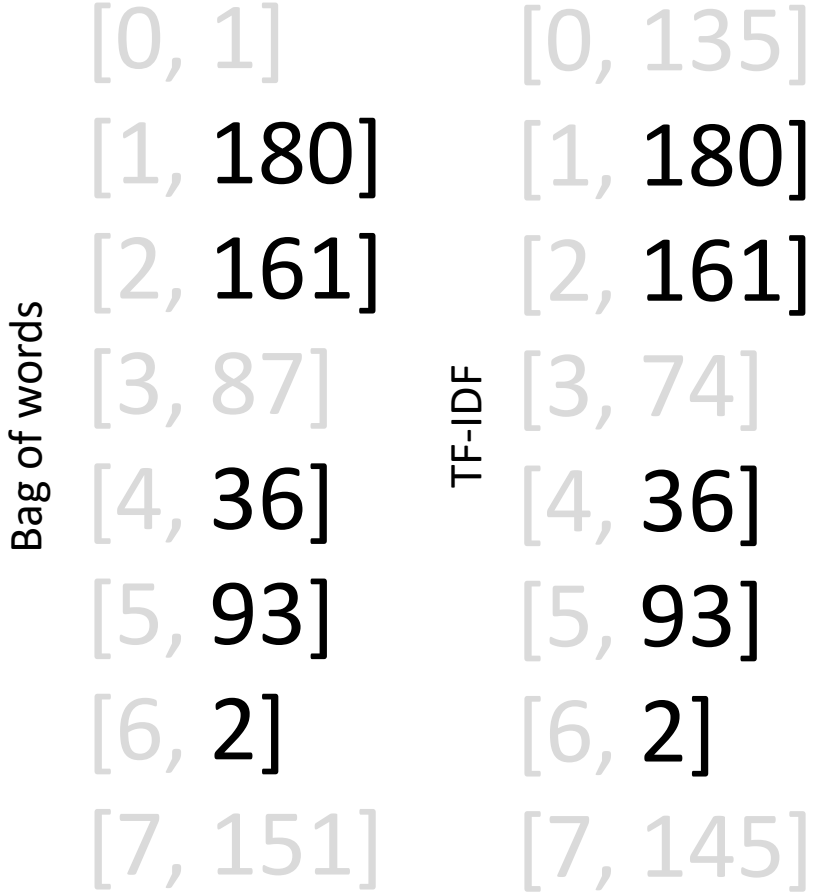
Bag of words

[0, 1]
[1, 180]
[2, 161]
[3, 87]
[4, 36]
[5, 93]
[6, 2]
[7, 151]

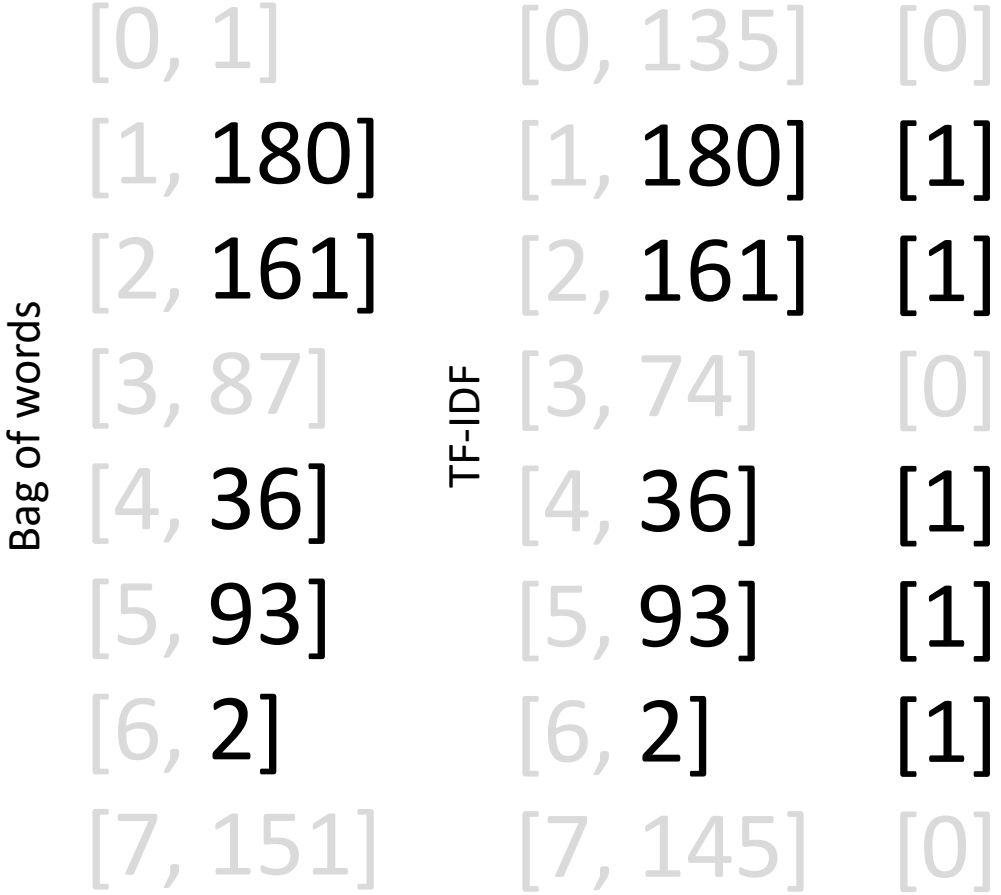
TF-IDF

[0, 135]
[1, 180]
[2, 161]
[3, 74]
[4, 36]
[5, 93]
[6, 2]
[7, 145]

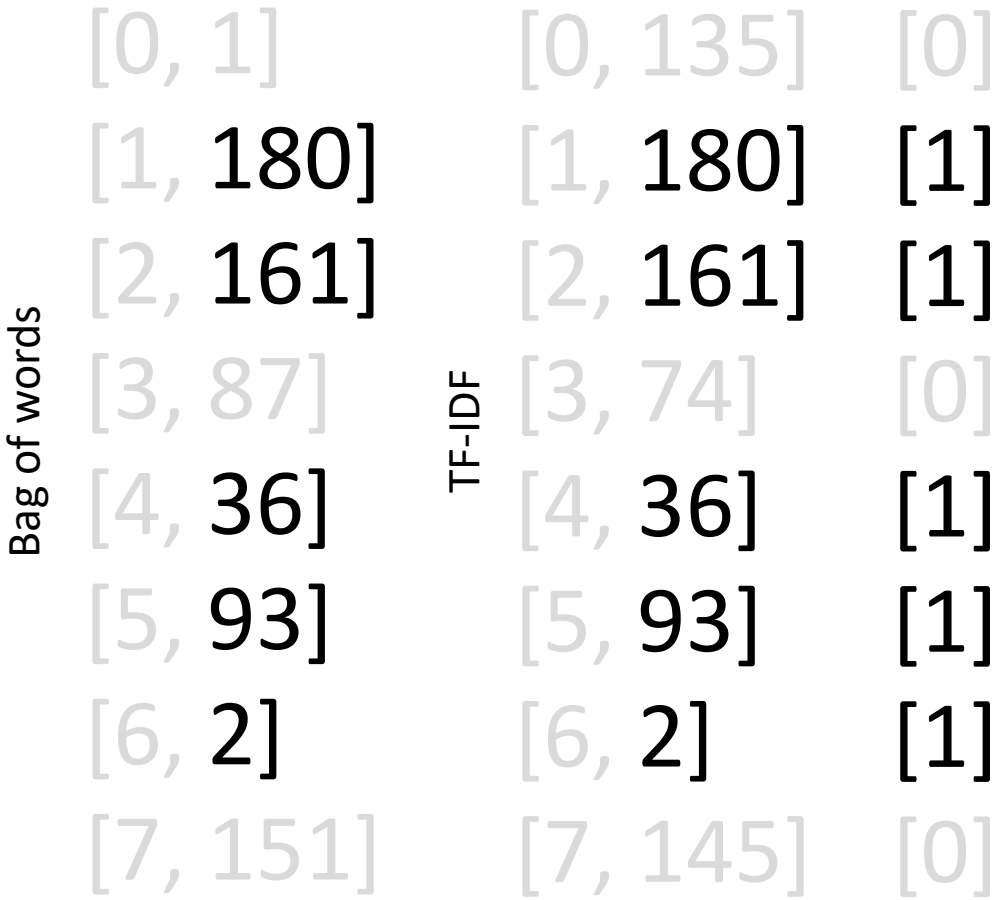
Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF



Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF



Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF



66%

Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF

- Коэффициент корреляции между массивами и номеров статей и расстояний 85%;

Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF

- Коэффициент корреляции между массивами и номеров статей и расстояний 85%;
- Модель мешка слова и TF-IDF даёт одинаковую ошибку в 0,4% случаев;

Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF

- Коэффициент корреляции между массивами и номеров статей и расстояний 85%;
- Модель мешка слова и TF-IDF даёт одинаковую ошибку в 0,4% случаев;
- Время расчёта обоих моделей примерно одинаковое ~30 мин;

Сравнение способов определения рецензента через косинусное расстояние мешка слов и TF-IDF

- Коэффициент корреляции между массивами и номеров статей и расстояний 85%;
- Модель мешка слова и TF-IDF даёт одинаковую ошибку в 0,4% случаев;
- Время расчёта обоих моделей примерно одинаковое ~30 мин;
- Субъективно определение рецензента методом TF-IDF точнее.

Частеречная разметка

Частотное распределение грамматических категорий по жанру научных публикаций (%)

Частотный словарь русского языка / под ред. Л. Н. Засориной М.: Издательство «Русский язык», 1977. 936 с.

Существительное	Глагол	Прилагательное	Наречие	Числительное	Местоимение	Союз	Предлог	Частица	Причастие	Субстантив. причастие	Субстантив. прилагат.	Омонимы типа сущ./глаг.	Остальные
31	13,5	12,5	7,26	1,03	11,6	7,61	11,2	0,67	1,36	0,03	0,52	0,08	1,71

Частеречная разметка

Частотное распределение грамматических категорий по жанру научных публикаций (%)

Частотный словарь русского языка / под ред. Л. Н. Засориной М.: Издательство «Русский язык», 1977. 936 с.

Существительное	Глагол	Прилагательное	Наречие	Числительное	Местоимение	Союз	Предлог	Частица	Причастие	Субстантив. причастие	Субстантив. прилагат.	Омонимы типа сущ./глаг.	Остальные
31	13,5	12,5	7,26	1,03	11,6	7,61	11,2	0,67	1,36	0,03	0,52	0,08	1,71

Существительное	Глагол	Прилагательное	Наречие	Числительное	Местоимение	Союз	Предлог	Частица	Причастие	Субстантив. причастие	Субстантив. прилагат.	Омонимы типа сущ./глаг.	Остальные
46	20	18	0	0	0	0	0	0	0	0	0	0	16

Частеречная разметка

Частотный словарь русского языка / под ред. Л. Н. Засориной М.: Издательство «Русский язык», 1977.

Существительное	Глагол	Прилагательное	Наречие	Числительное	Местоимение	Союз	Предлог	Частица	Причастие	Субстантив. причастие	Субстантив. прилагат.	Омонимы типа сущ./глаг.	Остальные
46	20	18	0	0	0	0	0	0	0	0	0	0	16

Корпус статей по радиолокации

Существительное	Глагол	Прилагательное	Наречие	Числительное	Местоимение	Союз	Предлог	Частица	Причастие	Субстантив. причастие	Субстантив. прилагат.	Омонимы типа сущ./глаг.	Остальные
51	11	19	0	0	0	0	0	0	0	0	0	0	19

Корпус статей по газодинамике

Существительное	Глагол	Прилагательное	Наречие	Числительное	Местоимение	Союз	Предлог	Частица	Причастие	Субстантив. причастие	Субстантив. прилагат.	Омонимы типа сущ./глаг.	Остальные
48	10	14	0	0	0	0	0	0	0	0	0	0	27

Частеречная разметка

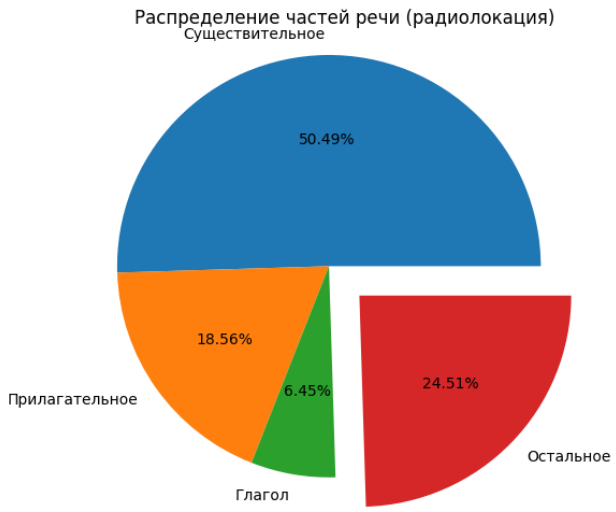
Частотное распределение грамматических категорий (%)

Частотный словарь современного русского языка на материалах Национального корпуса русского языка [Текст] / О. Н. Ляшевская, С. А. Шаров ; Российская акад. наук, Ин-т русского языка им. В. В. Виноградова. - Москва : Азбуковник, 2009. - 1087

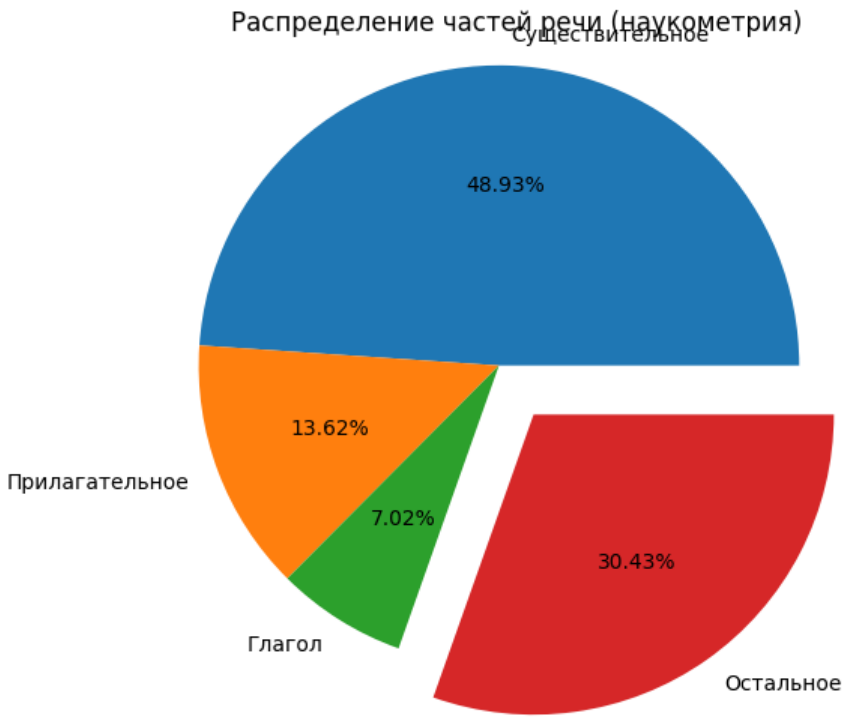
Существительное	Глагол	Прилагательное	Наречие	Числительное	Местоимение	Союз	Предлог	Частица	Причастие	Субстантив. причастие	Субстантив. прилагат.	Омонимы типа сущ./глаг.	Остальные
28,9	16,7	10,4	5	2	11,3	7,60	10,9	3,8	–	-	-	-	1

Существительное	Глагол	Прилагательное	Наречие	Числительное	Местоимение	Союз	Предлог	Частица	Причастие	Субстантив. причастие	Субстантив. прилагат.	Омонимы типа сущ./глаг.	Остальные
48,7	28,1	14,5	0	0	0	0	0	0	0	0	0	0	8

Частеречная разметка



Частеречная разметка



Частеречная разметка

Коллокации (биграммы) likelihood_ratio

- Радиолокация

[('диаграмма', 'направленность'),
('представлять', 'себя'),
('такой', 'образ'),
('ключевой', 'слово'),
('разрешать', 'способность'),
('пассивный', 'помеха'),
('подстилать', 'поверхность'),
('радиальный', 'скорость'),
('отношение', 'сигнал/шум'),
('радиолокационный', 'станция'),
('антенна', 'решётка'),
('несущий', 'частота'),
('антенный', 'решётка'),
('длина', 'волна'),
('период', 'повторение')]

- Газодинамика

[('эжектировать', 'газ'),
('продукт', 'сгорание'),
('число', 'мах'),
('камера', 'сгорание'),
('полный', 'давление'),
('численный', 'моделирование'),
('угол', 'атака'),
('ключевой', 'слово'),
('граничный', 'условие'),
('такой', 'образ'),
('вихревой', 'эжектор'),
('летательный', 'аппарат'),
('набегать', 'поток'),
('представлять', 'себя'),
('математический', 'модель')]

- Наукометрия

[('алмаз', 'антея'),
('вко', 'алмаз'),
('концерн', 'вко'),
('вестник', 'концерн'),
('научный', 'журнал'),
('журнал', 'вестник'),
('научно-технический', 'журнал'),
('следовать', 'отметить'),
('редакционный', 'коллегия'),
('время', 'рецензирование'),
('массовый', 'обслуживание'),
('литературный', 'редактирование'),
('система', 'массовый'),
('количество', 'стать'),
('уровень', 'значимость')]

Коллокации (биграммы) raw_freq

- Радиолокация

[('диаграмма', 'направленность'),
('такой', 'образ'),
('обработка', 'сигнал'),
('антенна', 'решётка'),
('радиальный', 'скорость'),
('частота', 'доплера'),
('зондировать', 'сигнал'),
('пассивный', 'помеха'),
('угловой', 'координата'),
('антенный', 'решётка'),
('отношение', 'сигналшум'),
('боковой', 'лепестковый'),
('полезный', 'сигнал'),
('ключевой', 'слово'),
('представлять', 'себя')]

- Газодинамика

[('эжектировать', 'газ'),
('полный', 'давление'),
('продукт', 'сгорание'),
('численный', 'моделирование'),
('число', 'мах'),
('камера', 'сгорание'),
('вихревой', 'эжектор'),
('результат', 'расчёт'),
('математический', 'модель'),
('угол', 'атака'),
('граничный', 'условие'),
('такой', 'образ'),
('методика', 'расчёт'),
('газ', 'выход'),
('набегать', 'поток')]

- Наукометрия

[('научный', 'журнал'),
('алмаз', 'антея'),
('вко', 'алмаз'),
('концерн', 'вко'),
('вестник', 'концерн'),
('научно', 'технический'),
('журнал', 'вестник'),
('количество', 'стать'),
('технический', 'журнал'),
('средний', 'время'),
('поток', 'стать'),
('связь', 'рецензент'),
('следовать', 'отметить'),
('время', 'рецензирование'),
('редакционный', 'коллегия')]

Коллокации (триграммы) likelihood_ratio

- Радиолокация

[('**диаграмма**', 'направленность', 'антенна'),
('динамический', '**диаграмма**', 'направленность'),
('лепесток', '**диаграмма**', 'направленность'),
('лепестковый', '**диаграмма**', 'направленность'),
('ширина', '**диаграмма**', 'направленность'),
('парциальный', '**диаграмма**', 'направленность'),
('разностный', '**диаграмма**', 'направленность'),
('максимум', '**диаграмма**', 'направленность'),
('**диаграмма**', 'направленность', 'фара'),
('**диаграмма**', 'направленность', 'приёмный'),
('нуль', '**диаграмма**', 'направленность'),
('результирующий', '**диаграмма**', 'направленность'),
('форма', '**диаграмма**', 'направленность'),
('формирование', '**диаграмма**', 'направленность'),
('**диаграмма**', 'направленность', 'основной')]

- Газодинамика

[('**эжектировать**', 'газ', 'выход'),
('**эжектировать**', 'газ', 'вход'),
('**эжектировать**', '**эжектировать**', 'газ'),
('давление', '**эжектировать**', 'газ'),
('энергообмен', '**эжектировать**', 'газ'),
('температура', '**эжектировать**', 'газ'),
('энергия', '**эжектировать**', 'газ'),
('расход', '**эжектировать**', 'газ'),
('**эжектировать**', 'газ', 'завершение'),
('скорость', '**эжектировать**', 'газ'),
('**эжектировать**', 'газ', 'падать'),
('**эжектировать**', 'газ', 'соответственно'),
('**эжектировать**', 'газ', 'кг/с'),
('**эжектировать**', 'газ', 'полный'),
('работа', '**эжектировать**', 'газ')]

- Наукометрия

[('вко', 'алмаз', 'антея'),
('концерн', 'вко', 'алмаз'),
('вестник', 'концерн', 'вко'),
('журнал', 'вестник', 'концерн'),
('научный', 'журнал', 'вестник'),
('научно-технический', 'журнал', 'вестник'),
('оборона', 'алмаз', 'антея'),
('алмаз', 'антея', 'москва'),
('алмаз', 'антея', 'российский'),
('алмаз', 'антея', 'составлять'),
('алмаз', 'антея', 'входить'),
('алмаз', 'антея', '//.-./'),
('алмаз', 'антея', '2015–2019'),
('алмаз', 'антея', '2015–2021'),
('алмаз', 'антея', 'кажущийся')]

Коллокации (триграммы) raw_freq

- Радиолокация

[('уровень', 'боковой', 'лепестковый'),
('фазировать', 'антенный', 'решётка'),
('диаграмма', 'направленность', 'антенна'),
('фазировать', 'антенна', 'решётка'),
('доплеровский', 'сдвиг', 'частота'),
('радиальный', 'скорость', 'цель'),
('вероятность', 'ложный', 'тревога'),
('вероятность', 'правильный', 'обнаружение'),
('разрешение', 'наклонный', 'дальность'),
('амплитудный', 'фазовый', 'распределение'),
('отношение', 'сигнал', 'шум'),
('радиолокационный', 'станция', 'рлс'),
('беспилотный', 'летательный', 'аппарат'),
('вко', 'алмаз', 'антея'),
('концерн', 'вко', 'алмаз')]

- Газодинамика

[('уравнение', 'навий', 'стокс'),
('эжектировать', 'газ', 'вход'),
('газ', 'вход', 'эжектор'),
('давление', 'эжектировать', 'газ'),
('эжектировать', 'эжектировать', 'газ'),
('выход', 'вихревой', 'эжектор'),
('коэффициент', 'лобовой', 'сопротивление'),
('газ', 'выход', 'вихревой'),
('давление', 'камера', 'сгорание'),
('полный', 'давление', 'эжектировать'),
('полный', 'температура', 'эжектировать'),
('утопить', 'часть', 'сопло'),
('газ', 'выход', 'эжектор'),
('давление', 'газ', 'выход'),
('уравнение', 'перенос', 'излучение')]

- Наукометрия

[('вко', 'алмаз', 'антея'),
('концерн', 'вко', 'алмаз'),
('вестник', 'концерн', 'вко'),
('журнал', 'вестник', 'концерн'),
('научно', 'технический', 'журнал'),
('технический', 'журнал', 'вестник'),
('редакция', 'научный', 'журнал'),
('средний', 'время', 'рецензирование'),
('аудитория', 'научный', 'журнал'),
('граф', 'связь', 'рецензент'),
('предельный', 'количество', 'стать'),
('время', 'рассмотрение', 'стать'),
('стать', 'научно', 'технический'),
('учёный', 'старший', 'год'),
('литературный', 'редактирование', 'вёрстка'),
('количество', 'связь', 'рецензент')]

Выводы

- Косинусное расстояние определяет рубрику статьи по максимуму расстояний

Выводы

- Косинусное расстояние определяет рубрику статьи по максимуму расстояний
- Рубрики пересекаются по косинусному расстоянию, но до определенного предела, после которого нет совпадений

Выводы

- Косинусное расстояние определяет рубрику статьи по максимуму расстояний
- Рубрики пересекаются по косинусному расстоянию, но до определенного предела, после которого нет совпадений
- Расстояние TF-IDF и bag-of-words корреляционно связано. Коэффициент корреляции 86%, что говорит о высокой связи

Выводы

- Косинусное расстояние не является мерой принятия решения о публикации

Выводы

- Косинусное расстояние не является мерой принятия решения о публикации
- Распределение косинусного расстояние гамма-функция

Выводы

- Косинусное расстояние не является мерой принятия решения о публикации
- Распределение косинусного расстояние гамма-функция
- Распределение гамма-функции у разных корпусов схожее

Выводы

- Косинусное расстояние не является мерой принятия решения о публикации
- Распределение косинусного расстояние гамма-функция
- Распределение гамма-функции у разных корпусов схожее
- Нельзя сделать вывод о перспективах публикации статьи по косинусному расстоянию, частотности слов, коллокациям