Задание 3

Цель работы: анализ информационных характеристик дискретных источников.

Подготовил: Файтельсон Антон

Домашняя работа

Текст для анализа:

Добро пожаловать во второе издание прогит. Первое издание было опубликовано более четырех лет назад. С тех пор многое изменилось, но многие важные вещи остались неизменны. Хотя большинство ключевых команд и концепций по-прежнему работают, так как команда, разрабатывающая ядро гит, фантастическим образом оставляет всё обратно совместимым, произошло несколько существенных дополнений и изменений в сообществе вокруг гит. Второе издание призвано обозначить эти изменения и обновить книгу для помощи новичкам. Когда я писал первое издание, гит ещё был относительно сложным в использовании и подходил лишь для настоящих хакеров. И хотя в некоторых сообществах он уже начинал набирать обороты, ему было далеко до сегодняшней распространённости. С тех пор его приняло практически всё сообщество свободного программного обеспечения. Гит достиг невероятного прогресса в Виндовс, взрывными темпами получил графический интерфейс для всех платформ, поддержку сред разработки и стал использоваться в бизнесе. Прогит четырехлетней давности ничего подобного не подозревал. Одна из главных целей издания — затронуть в гит сообществе эти рубежи. Сообщество свободного программного обеспечения тоже испытало взрывной рост. Когда я лет пять назад впервые сел писать книгу (первая версия потребовала времени), я как раз начал работать в крохотной компании, разрабатывающей сайт для гит хостинга под названием гитхаб. На момент публикации у сайта было лишь несколько тысяч пользователей и четверо разработчиков. Когда же я пишу это предисловие, гитхаб объявляет о десяти миллионах размещённых проектов, около пяти миллионах аккаунтах разработчиков и более 230 сотрудниках. Его можно любить или ненавидеть, в любом случае Гитхаб сильнейшим образом изменил сообщество свободного программного обеспечения, что было едва мыслимо, когда я только сел писать первое издание. Небольшую часть исходной версии прогит я посвятил гитхаб в качестве примера хостинга, с которым мне никогда не было особо удобно работать. Мне не сильно нравилось писать то, что, по-моему, было ресурсом сообщества, а также упоминать в нём о моей компании. Меня по-прежнему волнует это противоречие, но важность гитхаба в гит сообществе бесспорна. Вместо некоего примера гит хостинга, я решил посвятить этот раздел книги детальному описанию сути гитхаба и его эффективному использованию. Если вы собираетесь узнать, как пользоваться гит, то умение пользоваться гитхабом даст вам возможность поучаствовать в огромном сообществе, ценном вне зависимости от выбранного вами гит хостинга. Другим изменением с момента первой публикации стала разработка и развитие хттп протокола для сетевых гит транзакций. Из соображений упрощения, большинство примеров из книги были переделаны из ссх на хттп. Было изумительно смотреть, как за несколько прошедших лет гит вырос из весьма невзрачной системы контроля версий до безусловно лидирующей в коммерческой и некоммерческой сферах. Я счастлив, что прогит так хорошо выполнил свою работу, оказавшись одним из немногих представителей успешной и при этом полностью открытой технической литературы. Я надеюсь, вам понравится это новое издание прогит.

Характеристики текста:

1) Время набора: 9 минут 36 секунд

2) Количество символов: 3062 символа

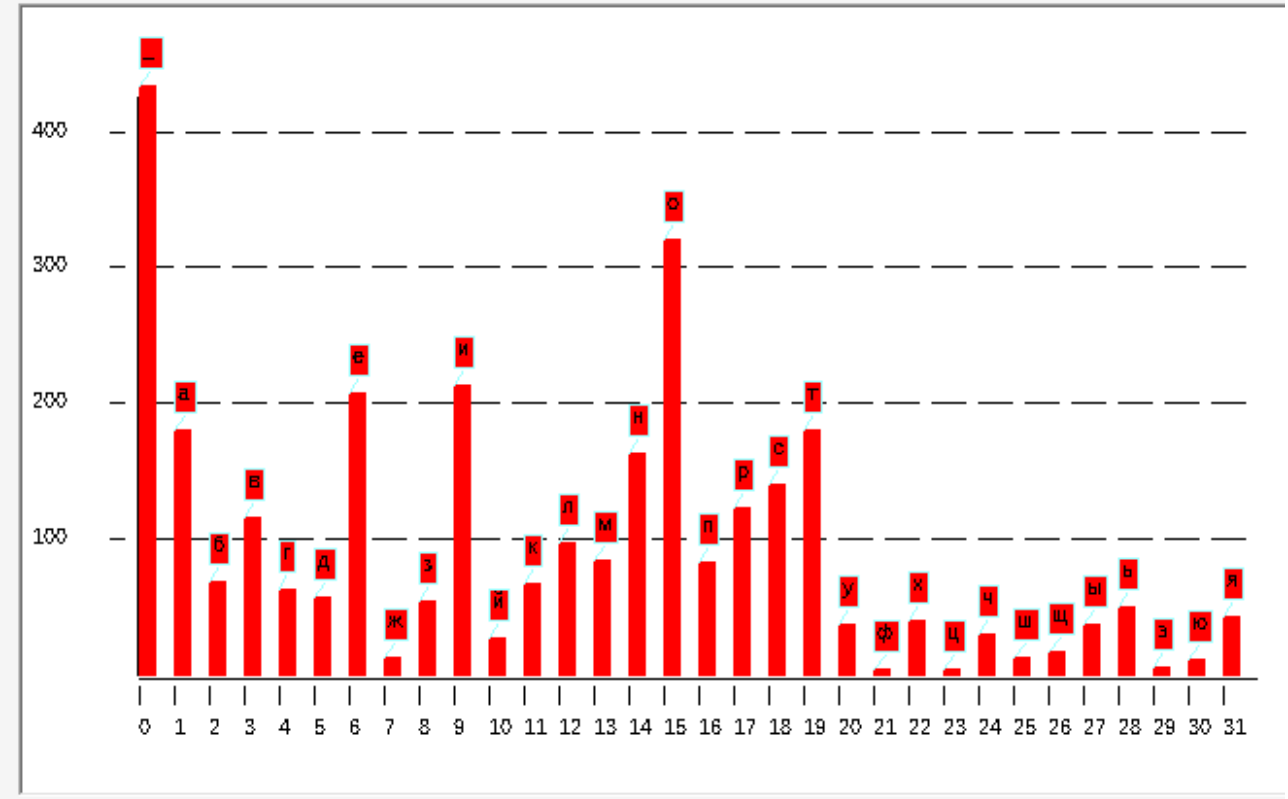
3) Количество информации: 12.03443446068559 бит, полагая энтропию русского языка равной Нрус = 1,37 [бит/симв].

4) Производительность: 7,28288194444 [бит/с], с учетом того, что энтропия Нрус = 1,37 [бит/симв], а скорость выдачи символов: 5,31597222222 [симв/с].

5) Избыточность источника: 0.726, считая, что объём алфавита источника K = 32.

Снятие гистограмм распределения

Текст на русском:

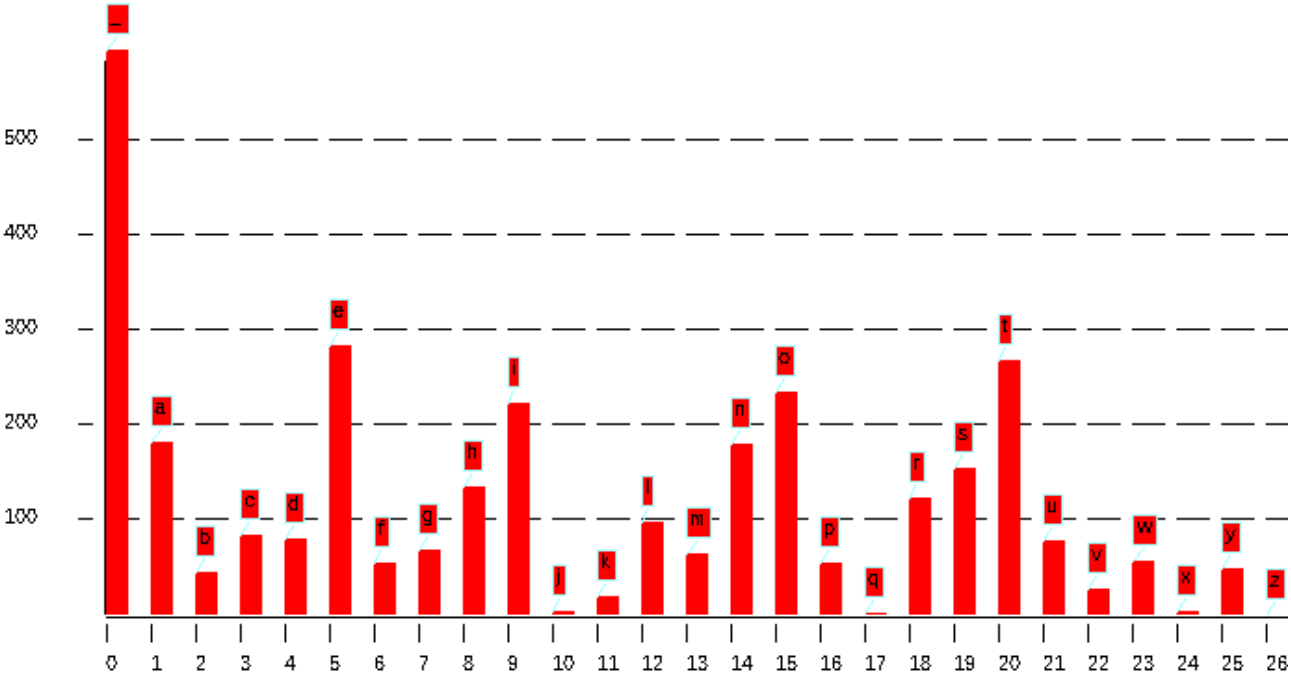


Символ	_	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о
Кол-во	436	183	71	118	65	60	210	15	57	216	29	70	100	87	165	323
Символ	п	р	с	т	у	ф	х	ц	ч	ш	щ	ы	ь	э	ю	я
Кол-во	86	125	142	183	40	7	42	7	32	15	19	40	53	8	13	45

1) Частость пробела: 0,142390594383 (Длина текста: 3062 [символов].)

2) Средняя длина слова: 6.022935779804498 символа.

Текст на английском:

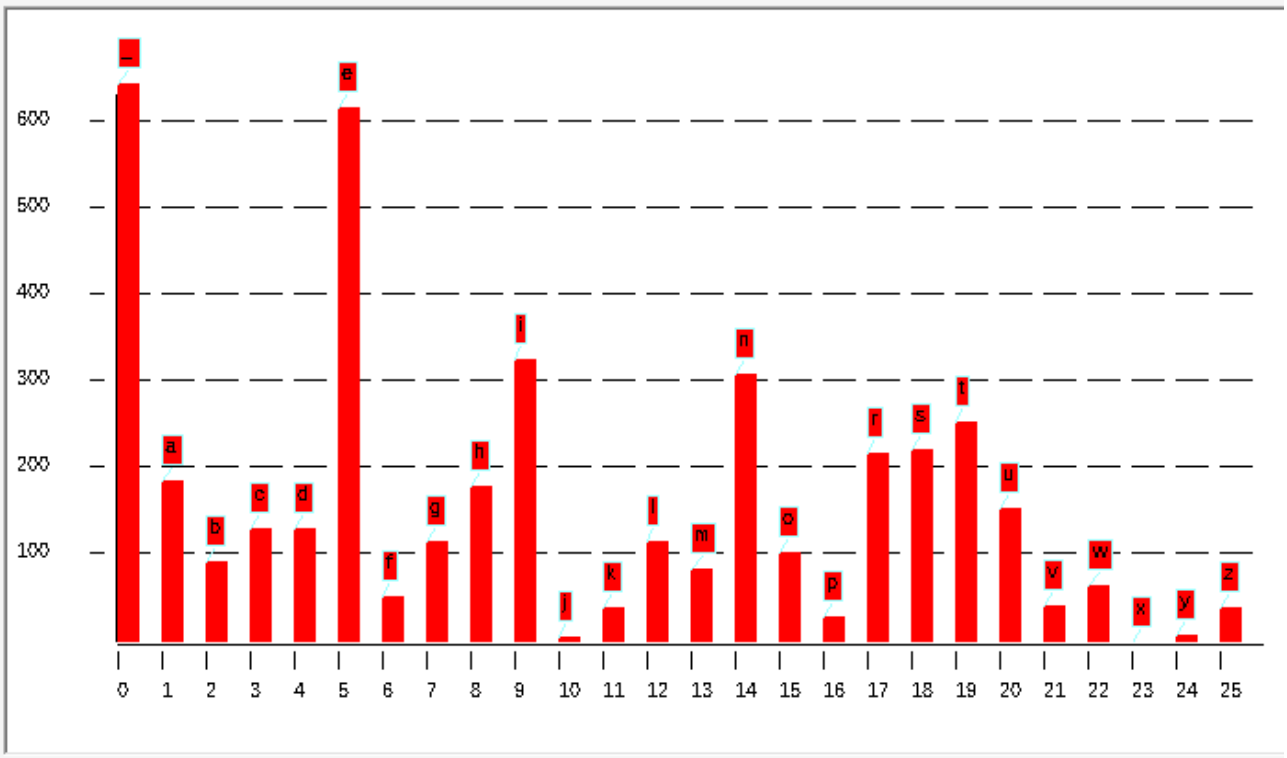


Символ	_	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
Кол-во	597	184	46	85	82	285	56	70	136	224	4	20	99	66	182	237
Символ	p	q	r	s	t	u	v	w	x	y	z					
Кол-во	55	2	124	156	270	79	29	58	5	50	1					

1) Частость пробела: 0,186445971268 (Длина текста: 3202 [символов].)

2) Средняя длина слова: 4.3634840871009555 символа.

Текст на немецком:

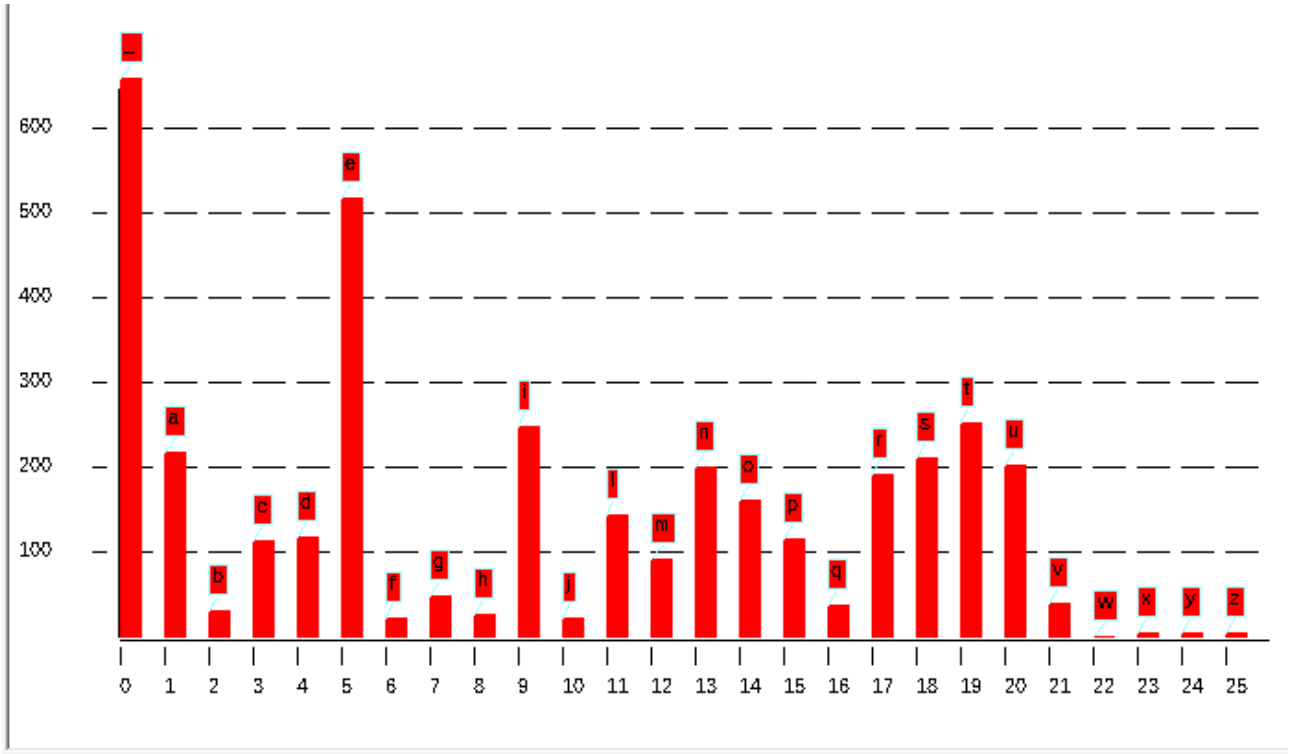


Символ	_	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
Кол-во	646	187	94	133	133	620	54	117	181	328	8	42	117	87	312	105
Символ	p	r	s	t	u	v	w	x	y	z						
Кол-во	31	220	224	256	156	44	67	1	10	42						

1) Частость пробела: 0.15326215895610915 (Длина текста: 4215 [символов].)

2) Средняя длина слова: 5.5247678018575845 символа.

Текст на французском:



Символ	_	a	b	c	d	e	f	g	h	i	j	l	m	n	o	p
Кол-во	662	221	33	117	121	521	25	51	29	251	24	147	95	204	164	118
Символ	q	r	s	t	u	v	w	x	y	z						
Кол-во	40	194	214	255	205	42	3	7	7	7						

1) Частость пробела: 0,142390594383 (Длина текста: 3757 [символов].)

2) Средняя длина слова: 4.675226586102719 символа.

Вывод:

Во всех исследуемых языках пробел — самый часто встречающийся символ. В каждом языке частость пробелов и длина слова разная. Есть свои особенности использования алфавита: в французском и немецком языках — второй по частости символ. В то время как в английском e — хоть и тоже второй по частости символ, но t и o лишь немного по частотности отстают.

Исследование энтропии различных языков

Текст на русском:

Длина n-граммы, n [симв]	Энтропия n-граммы,H(An) [бит/n-грамму]	Удельная энтропия, H(An)/n [бит/симв]	Условная энтропия, H(A|A'n−1)[бит/симв]
1	4,39	4,39	-
2	7,76	3,88	3.37
3	9,76	3,25	2
4	10,60	2,65	0.84
5	10,96	2,19	0.36

Избыточность = 0.22399999999999998

Текст на английском:

Длина n-граммы, n [симв]	Энтропия n-граммы,H(An) [бит/n-грамму]	Удельная энтропия, H(An)/n [бит/симв]	Условная энтропия, H(A|A'n−1)[бит/симв]
1	4,10	4,10	-
2	7,29	3,65	3.19
3	9,26	3,09	1.97
4	10,22	2,56	0.96
5	10,74	2,15	0.52

Избыточность = 0.23236879982139347

Текст на немецком:

Длина n-граммы, n [симв]	Энтропия n-граммы,H(An) [бит/n-грамму]	Удельная энтропия, H(An)/n [бит/симв]	Условная энтропия, H(A|A'n−1)[бит/симв]
1	4,09	4,09	-
2	7,19	3,60	3.1
3	9,29	3,10	2.1
4	10,44	2,61	1.15
5	11,06	2,21	0.62

Избыточность = 0.24288429571425096

Текст на французском:

Длина n-граммы, n [симв]	Энтропия n-граммы,H(An) [бит/n-грамму]	Удельная энтропия, H(An)/n [бит/симв]	Условная энтропия, H(A|A’n−1)[бит/симв]
1	4,00	4,00	-
2	7,09	3,54	3.09
3	9,12	3,04	2.03
4	10,26	2,56	1.14
5	10,87	2,17	0.61

Избыточность = 0.2555028907856801

Вывод:

Удельная энтропия и условная энтропия уменьшаются при увеличении длины n-граммы.

Данное утверждение было проверено на 4 языках.

Избыточность была подсчитана для каждого языка, выяснилось, что, чем больше удельная энтропия, тем меньше избыточность текста. Также избыточность зависит от максимального значения энтропии.

Генератор случайных текстов

N = 1: ьптв_дт_м_рмюкоиун_твлбнзо_се_ипосрннспотб_екнелзоояс

N = 2:_я_ремия_песитиия_печезамитах_взделотралюбе_дос_чнорытесиствабло_этеготлнни

N = 3: ниллико_печеслитхабольно_пишь_из_глатыся_вою_что_сло_сло_с_в_инга_дероящи

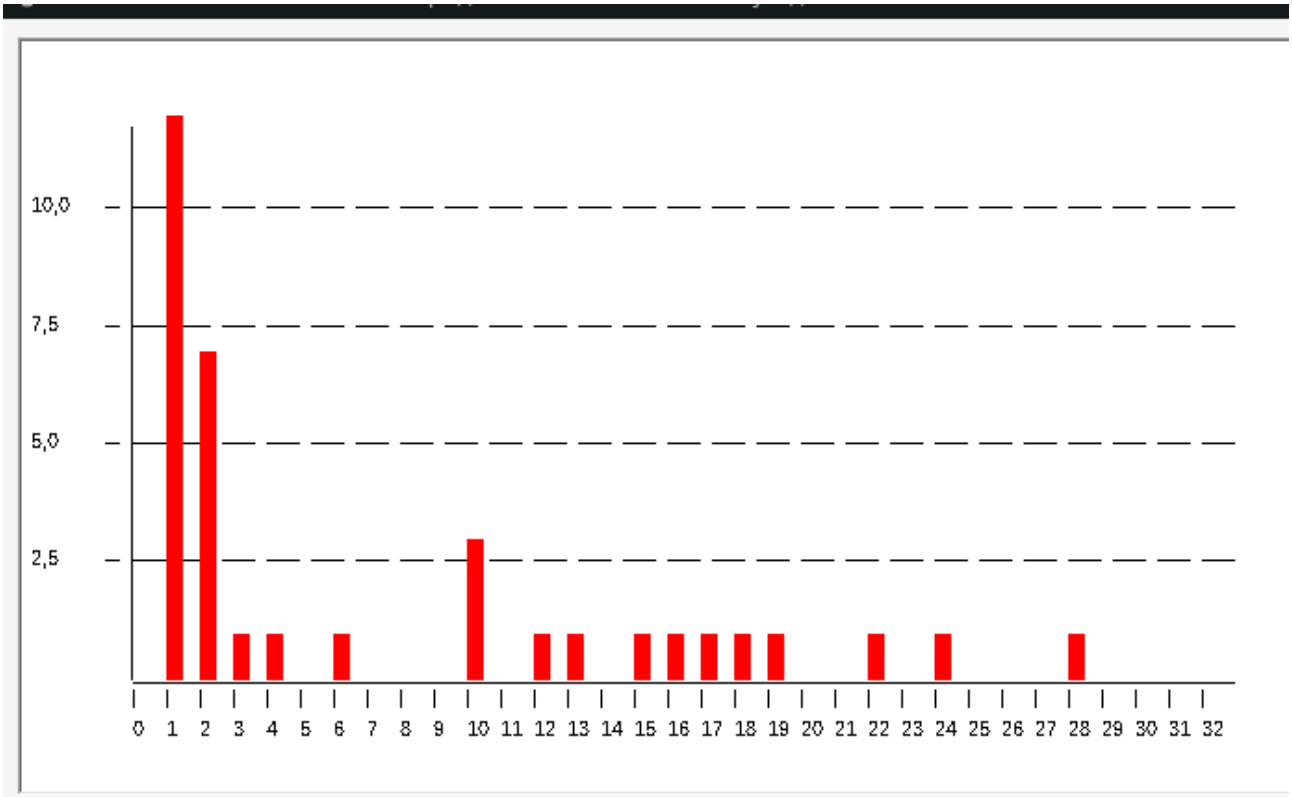
N = 4:но_смотрудним_из_главностально_смотреть_писать_первой_сильно_сообществе_этом

N = 5:_оказавшись_одним_изменил_сообществе_вокруг_гит_первые_сел_писать_гитхаб

Вывод:

При увеличении n-граммы слова становятся более понятными. Генератор случайных текстов начинает выдавать слова из оригинального текста.

Опыт Шеннона



Общий вывод:

Пробел — самый часто встречающийся символ, но частость пробелов и длина слова разная в разных языках. Удельная энтропия и условная энтропия уменьшаются при увеличении длины n-граммы, а избыточность зависит от максимального значения энтропии и удельной энтропии. При увеличении n-граммы слова становятся более понятными в генераторе случайных текстов.

Контрольные вопросы:

1) Дискретный истоник - источник сообщений, который может в каждый момент времени случайным образом принять одно из конечного множества возможных состояний.

Каждому состоянию источника U ставиться в соответствие условное обозначение в виде знака. Совокупность знаков $u_1, u_2,\ldots,u_i,\ldots,u_N$ соответствующих всем N возможным состояниям источника называют его **алфавитом.**

2) Информация (Information)- содержание сообщения или сигнала; сведения, рассматриваемые в процессе их передачи или восприятия, позволяющие расширить знания об интересующем объекте.
Информация - является одной из фундаментальных сущностей окружающего нас мира (акад. Поспелов).
Информация - первоначально - сведения, передаваемые одними людьми другим людям устным, письменным или каким - нибудь другим способом (БСЭ).
Информация - отраженное разнообразие, то есть нарушение однообразия.
Информация - является одним из основных универсальных свойств материи.
Информация - сведения об объектах и явлениях окружающей среды, их параметрах, свойствах и состоянии, которые воспринимают информационные системы (живые организмы, управляющие машины и др.) в процессе жизнедеятельности и работы.

Информационные меры отвечают трем основным направлениям в теории информации: структурному, статистическому и семантическому.
Структурная теория рассматривает дискретное строение массивов информации и их измерение простым подсчетом информационных элементов (квантов) или комбинаторным методом, предполагающим простейшее кодирование массивов информации.
Статистическая теория оперирует понятием энтропии как меры неопределенности, учитывающей вероятность появления, а, следовательно, и информативность тех или иных сообщений.
Семантическая теория учитывает целесообразность, ценность, полезность или существенность информации.

3)

Основной информационной характеристикой дискретного источника является его **энтропия**: среднее количество информации, приходящееся на один символ источника.

$$H(A) = \overline{I(a_i)} = -\sum_{i=0}^{K-1} p(a_i) \log_2 p(a_i).$$

Здесь $I(a_i) = -\log_2 p(a_i)$ — **информация**, содержащаяся в символе a_i, $p(a_i)$ — вероятность его появления, A — множество символов a_i (алфавит источника).

Энтропия измеряется в битах на символ источника: [бит/симв].

4)Сообщения (а также источники, их порождающие), в которых существуют статистические связи (корреляции) между знаками или их сочетаниями, называются сообщениями (источниками) с памятью или марковскими сообщениями (источниками).

Для источника с **памятью** (соседние символы сообщения зависимы) вводится понятие **условной энтропии**:

$$H(A|A') = -\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} p(a_i, a_j) \log_2 p(a_i|a_j).$$

Здесь $p(a_i, a_j)$ — вероятность совместного появления символов a_i и a_j, $p(a_i|a_j)$ — вероятность появления символа a_i при условии, что до него появился символ a_j (условная вероятность символа a_i), A' — множество символов источника на предыдущем шаге (a_j), A — на текущем шаге (a_i).

5)

Максимально возможное значение энтропии (максимально возможное среднее количество информации на символ) достигается при равновероятном выборе символов источником:

$$H_{\max}(A) = -\sum_{i=0}^{K-1} \frac{1}{K} \log_2 \frac{1}{K} = \log_2 K.$$

6)Безызбыточный источник — источник с максимальной энтропией(равновероятный выбор символов)

Избыточность источника

$$\rho_{\text{и}} = 1 - \frac{H(A)}{H_{\max}(A)} = 1 - \frac{n_0}{n}$$

характеризует относительное удлинение сообщения по сравнению с источником без избыточности (с максимальной энтропией). Здесь n — длина сообщения с энтропией $H(A)$, n_0 — минимально возможная длина сообщения с энтропией $H_{\max}(A)$.

7)При максимальном значении энтропии, производительность источника будет максимальной.

Среднее количество информации, выдаваемое источником в единицу времени, определяется его **производительностью**:

$$H'(A) = v_{\text{и}} H(A),$$

где $v_{\text{и}}$ — скорость выдачи символов.

Производительность измеряется в битах в секунду: [бит/с].

8)Энтропия языка — статистическая функция текста на определённом языке либо самого языка, определяющая количество информации на единицу текста. Энтропии разных языков различаются так, как различаются алфавиты(по количеству и по частоте появления разных букв, гистограммы распределния это подтверждают) и различается количество информации, которую несет текст, предложение или отдельный зн ак.

С одной стороны, энтропию языка можно оценить как энтропию группы из n символов A_n, поделённую на количество символов в группе:

$$H_n^+ = \frac{H(A_n)}{n}.$$

С другой стороны, энтропию языка также можно оценить как **условную энтропию** n-го символа a_n при известных $n-1$ предыдущих символах — $H(A|A'_{n-1})$:

$$H_n^- = H(A|A'_{n-1}) = H(A_n) - H(A_{n-1}).$$

9) Удельную и условную энтропию можно определить с помощью опыта Шеннона.

10)

Оригинальный способ определения энтропии языка был предложен в 1951 г. Шенноном [3]. Он заключается в отгадывании n-й буквы текста при известных $n-1$ предыдущих. Мера степени неопределённости данного опыта является оценкой сверху условной энтропии.

Из осмысленного текста наугад выбираются $n-1$ символов и кому-либо предлагается угадать n-й символ. Многократное повторение опыта даёт распределение частот правильного угадывания: частоты (вероятности) $w_1, w_2, \ldots, w_K$ того, что символ будет правильно угадан с $1, 2, \ldots, K$-й попытки (K — объём алфавита). Эти вероятности являются оценкой вероятностей символов алфавита, расположенных в порядке убывания частот [4]. Отсюда следует, что энтропия данного распределения будет являться оценкой (сверху) условной энтропии

$$H(A|A'_{n-1}) \leqslant H(W) = -\sum_{i=1}^{K} w_i \log w_i,$$

которая с увеличением n будет стремиться к энтропии языка.

Результат опыта зависит от «литературного чутья» и добросовестности отгадывающего. Для уменьшения влияния этих факторов, Шеннон предложил задавать вопросы ряду лиц и остановиться на том из них, ответы которого окажутся наиболее удачными.

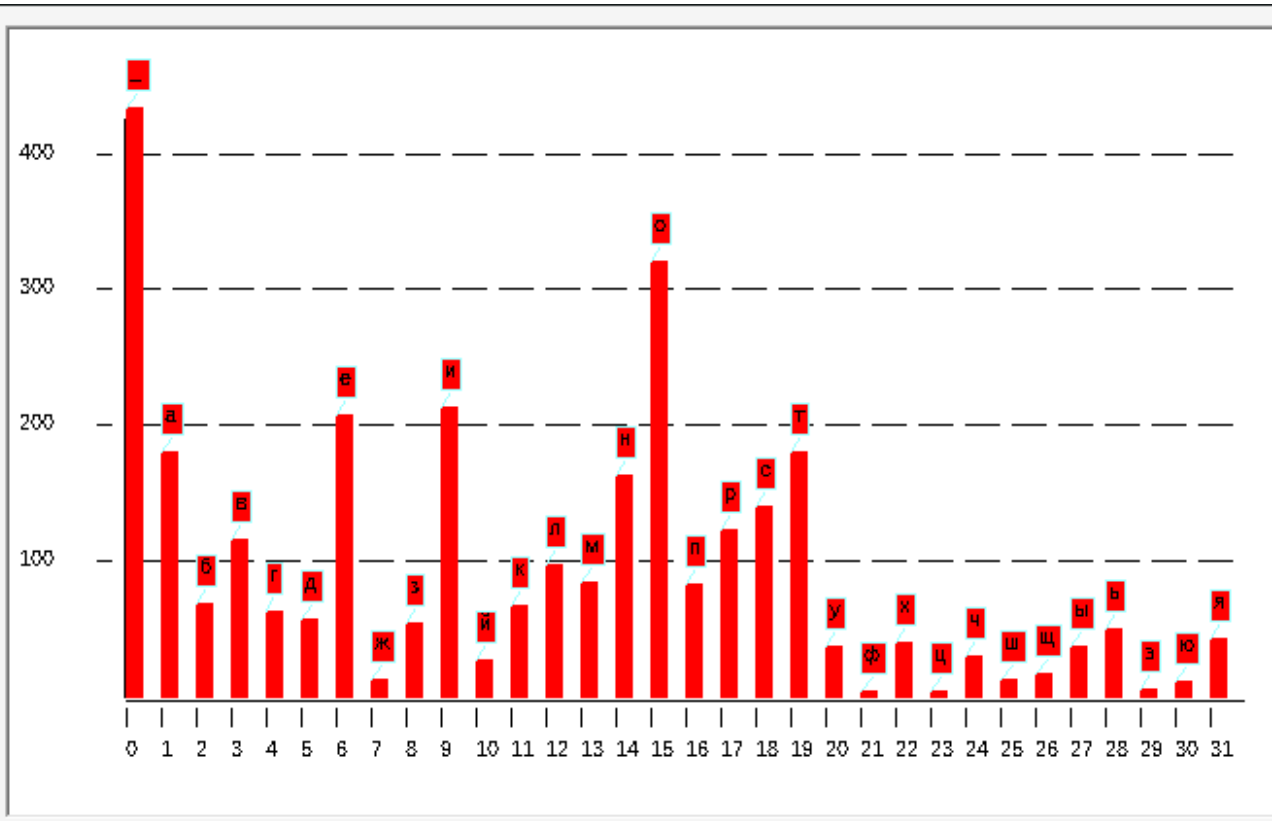
11)

Зная вероятность (частость) появления символа «пробел» ($\sqcup$), можно определить среднюю длину слова:

$$\overline{n}_{\text{сл}} = \lim_{N\to\infty} \frac{N - n_{\sqcup}}{n_{\text{сл}}} = \lim_{N\to\infty} \frac{N - NP(\sqcup)}{NP(\sqcup) + 1} = \frac{1 - P(\sqcup)}{P(\sqcup)}.$$

Здесь N — длина текста (устремлённая к бесконечности), $n_{\sqcup}$ — число пробелов в тексте, $n_{\text{сл}}$ — число слов в тексте, $P(\sqcup)$ — вероятность появления пробела.

12)



Символ	_	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о
Кол-во	436	183	71	118	65	60	210	15	57	216	29	70	100	87	165	323
Символ	п	р	с	т	у	ф	х	ц	ч	ш	щ	ы	ь	э	ю	я
Кол-во	86	125	142	183	40	7	42	7	32	15	19	40	53	8	13	45

Чтобы определить энтропию по данной гистограмме необходимо воспользоваться формулой.

Получив энтропию, нужно будет найти еще максимальную энтропию по формуле Hmax = log2(K), где K — количество букв в алфавите. Тогда можно будет посчитать избыточность по

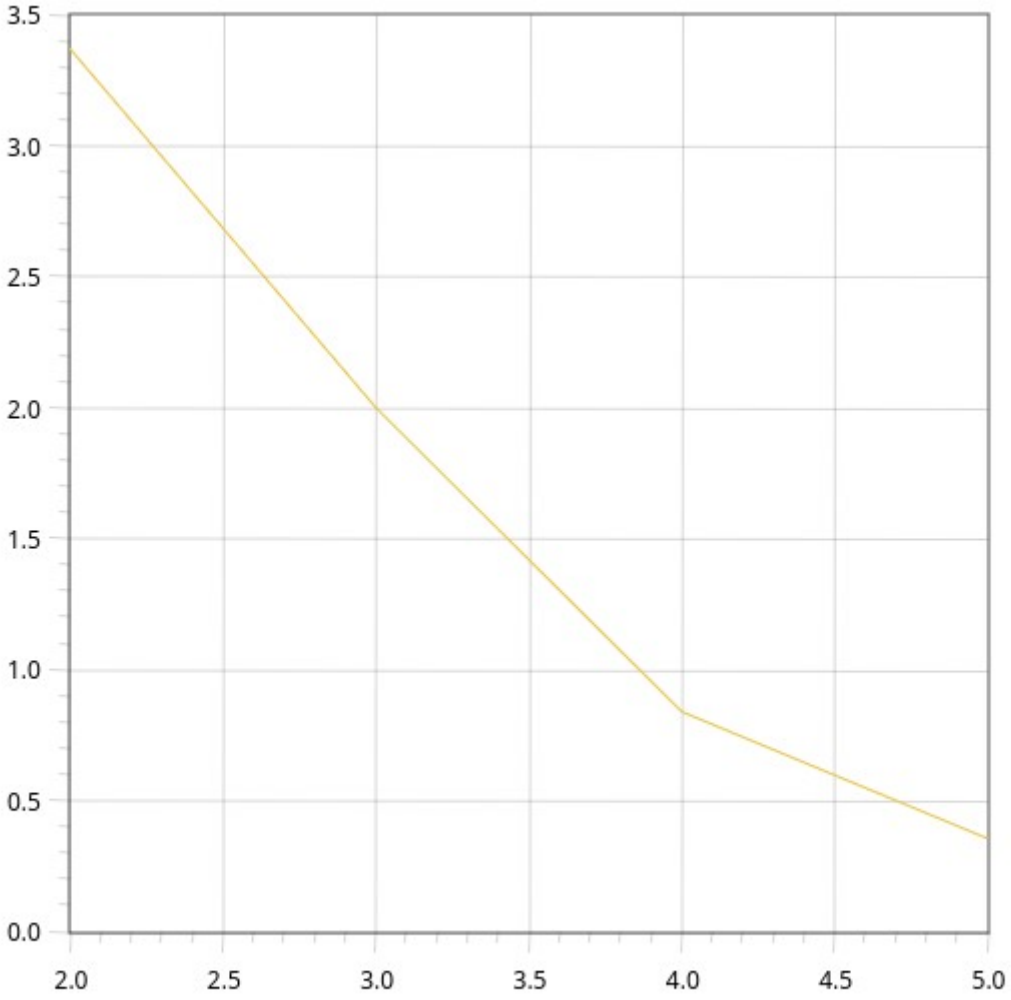
$$H(A) = \overline{I(a_i)} = -\sum_{i=0}^{K-1} p(a_i)\log_2 p(a_i).$$

Здесь $I(a_i) = -\log_2 p(a_i)$ — **информация**, содержащаяся в символе a_i, $p(a_i)$ — вероятность его появления, A — множество символов a_i (алфавит источника).

формуле:

$$\rho_и = 1 - \frac{H(A)}{H_{\max}(A)}$$

13)



В качестве оси y принято значение энтропии, а в качестве оси x принято значение n. Данный график показывает, что при увеличении значения n-граммы, значение условной энтропии уменьшается.