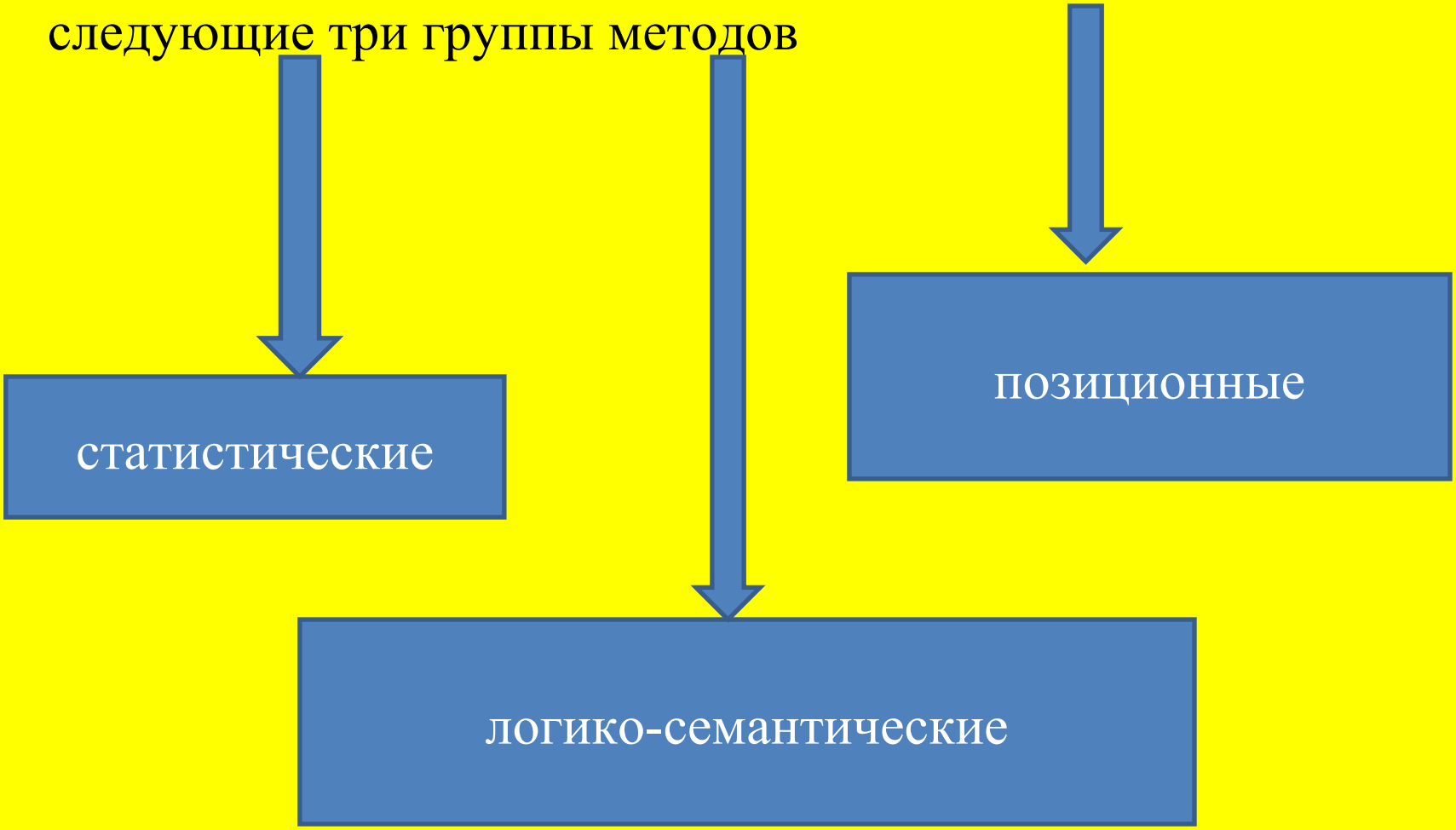
Говоря о двух последних «умениях» компьютера, необходимо помнить, что почти во всех существующих системах автоматического реферирования в качестве **основных смысловых единиц реферата** выступают **ключевые предложения** или **ключевые словосочетания** и **слова исходного текста**.

Первые в их последовательной совокупности (в том порядке, в котором они идут в исходном тексте) образуют текст (квазитекст) реферата.

Второй тип смысловых единиц (ключевые словосочетания и слова) используется компьютером для построения так называемых *табличных рефератов*.

При составлении с помощью компьютера аннотации также используются как ключевые предложения (в том виде, что и при составлении реферата), так и ключевые слова и словосочетания. Последние перечисляются вслед за реляторами вида: «В статье рассматриваются следующие вопросы:...», «Книга посвящена следующим проблемам: ...», «Статья раскрывает следующие понятия: ...» и т.д.

По способам выделения из исходных текстов ключевых словосочетаний и предложений (первые два «умения» компьютера) различают несколько методов автоматического реферирования и аннотирования текстов. Наиболее известны следующие три группы методов



Суть статистической группы методов заключается в том, что

ключевыми словами считаются такие знаменательные слова текста, которые с учетом всех синонимов встречаются в тексте наибольшее число раз

ключевым предложением считается предложение текста, которое

- имеет несколько ключевых слов
- содержит ключевые слова на небольшом расстоянии друг от друга

Принадлежность слова, словосочетания или предложения к числу ключевых определяется **специальными статистическими коэффициентами**.

В ***позиционных методах*** автоматического реферирования и аннотирования ключевым предложением считается предложение, входящее в заголовок, подзаголовок, начало или конец какой-то части текста или всего текста.

Такие предложения, как правило, содержат информацию о **целях, методах, выводах и результатах исследования**, описанного в первичном документе.

Важность тех или иных предложений с указанной точки зрения определяется экспертами путем изучения семантической структуры первичных документов определенного типа.

Логико-семантические методы опираются на исследование структуры и семантики текстов.

Существует несколько вариантов этих методов, но цель их одна — выделить из конкретного текста предложения с **наибольшим функциональным весом**.

Величина эта зависит от многих факторов:

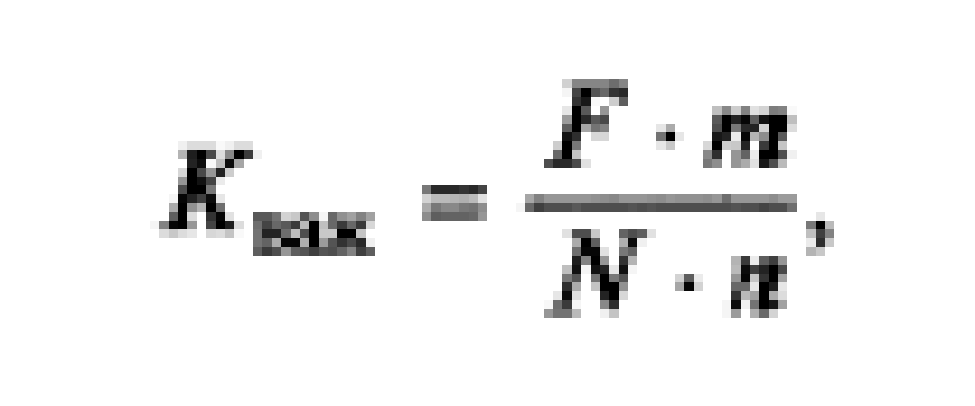
- наличия в исследуемом предложении специальных семантически значимых слов,
- связи этого предложения с другими предложениями текста,
- синтаксического типа самого предложения и т.д.

Формулируя задачу построения системы автоматического аннотирования и реферирования текста, необходимо четко указать

- метод, который используется для выделения ключевых слов предложения
- способ определения ключевых словосочетаний предложения
- критерий выделения ключевых предложений текста
- тип подготавливаемой аннотации: текстовая, в виде релятора с последующими ключевыми словами и словосочетаниями, или табличная
- тип формируемого реферата: текстовый или табличный

Учитывая все сказанное, сформулируем задачу ***автоматического реферирования и аннотирования текста*** следующим образом

На устройстве внешней памяти (например, дискете или винчестере) находится английский научно-технический текст. Начало каждого абзаца в нем обозначено знаком*. Используя для выделения ключевых (опорных) слов текста один из вариантов статистического метода, а именно коэффициент важности слова



В формуле для $K_{важ}$ буквы означают следующее: *F*— частота словоупотреблений в тексте; *m* — число абзацев текста, в которых встретилось слово; *N* — общее число словоупотреблений в тексте; *n* — общее число абзацев в тексте.

Это позволяет получить

аннотацию текста в виде релятора со следующими за ним ключевыми словосочетаниями текста. Ключевым сочетанием считается ключевое имя существительное со стоящим перед ним определением, выраженным именем прилагательным или причастием, не относящимся к числу общеупотребительных

словесный реферат текста в виде последовательной цепочки ключевых предложений. Ключевым предложением текста будем считать предложение, содержащее три и более разных ключевых слова.

Аннотацию и реферат вывести на экран дисплея

Исходным материалом для сформулированной выше задачи автоматического аннотирования и реферирования будет служить текст

•MICROPROCESSORS AND MINICOMPUTERS.

•THE FIRST COMPUTERS FILLED A LARGE ROOM WITH THEIR ELECTRONICS. PHOTOGRAPHS OF EARLY COMPUTERS SHOW MEN AND WOMEN IN BUSINESS SUITS AND LABORATORY COATS STANDING IN THE MIDDLE OF A ROOM SURROUNDED BY A U-SHAPED MACHINE. IN REALITY, PEOPLE OPERATING AND DEVELOPING THE FIRST COMPUTERS DID NOT WEAR SUITS. AIR-CONDITIONING WAS POOR AT THAT TIME AND THE COMPUTER GOT SO HOT THAT THE COMPUTER OPERATORS DRESSED IN T-SHIRTS AND TENNIS SHOES.

•THE DEVELOPMENT OF THE TRANSISTOR IN 1948 MADE IT POSSIBLE TO BUILD ELECTRONIC DEVICES OF A VERY SMALL SIZE. AN INTEGRATED CIRCUIT (IC(CHIP) FOR SHORT) WAS SOON DEVELOPED. AN IC(CHIP) IS A LARGE NUMBER OF TRANSISTORS AND OTHER ELECTRONIC COMPONENTS WIRED TOGETHER. THEY FORM A SPECIAL CIRCUIT AND ARE MICROSCOPICALLY PLACED ON A SMALL PIECE OF SILICON (OR OTHER MATERIAL) WHICH SERVES AS SEMICONDUCTOR.

•INTEGRATED CIRCUITS, ALSO CALLED CHIPS, ARE NOW MANUFACTURED SEPARATELY FROM COMPUTERS. THEY PROVIDE BUILDING BLOCKS TO BUILD A COMPUTER. THE MOST IMPORTANT OF THESE COMPUTER COMPONENTS IS THE CENTRAL PROCESSING UNIT (CPU FOR SHORT) OR MICROPROCESSOR. IT IS THE PART OF THE COMPUTER THAT OBEYS THE INSTRUCTIONS OF A PROGRAM. A MICROPROCESSOR IS A SMALL UNIT CONTAINED ON THE SEMICONDUCTOR CHIP. MICROPROCESSORS ARE USED IN MINICOMPUTERS AND EACH OF THESE INTEGRATED CIRCUITS IS CAPABLE OF PROCESSING 8-BIT OR 16-BIT DATA…

Принципиальный алгоритм решения задачи

Решаемая задача является сложной и состоит из нескольких простых подзадач

В *блоке А* в память компьютера вводится находящийся на устройстве внешней памяти файл с английским научно-техническим текстом.

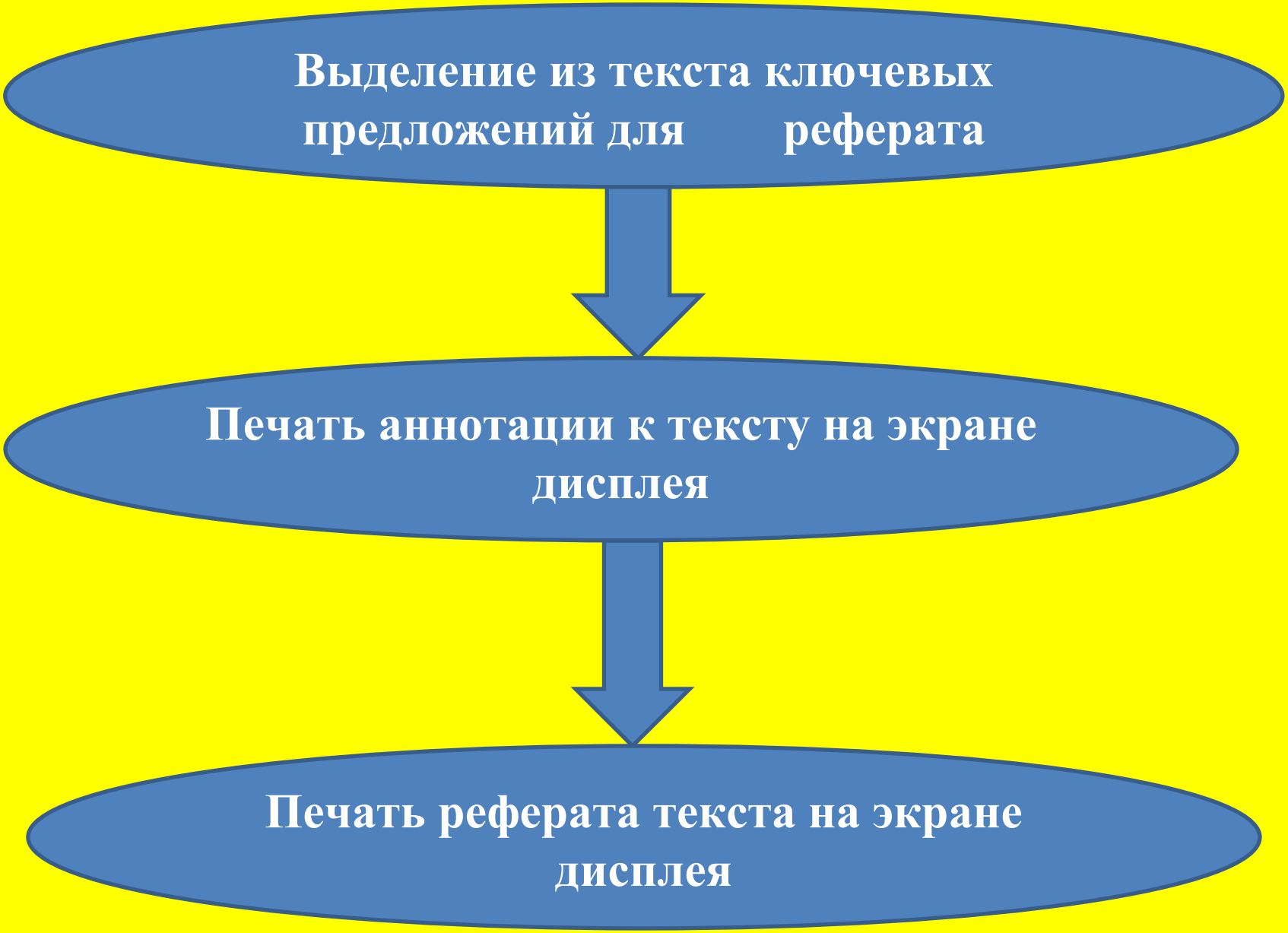
Далее в *блоке В* формируется массив словоформ од-ного абзаца текста, словоформы сортируются по алфавиту, подсчитывается частота употребления каждой словоформы в абзаце и в специальной области компьютерной памяти создается алфавитно-частотный словарь одного абзаца. Подобные действия выполняются со всеми абзацами текста.

В результате выполнения *блоков А и В* алгоритма в специальных областях памяти компьютера будут расположены алфавитно-частотные словари словоформ всех семи абзацев текста.

Блок С объединяет алфавитно-частотные словари абзацев в единый распределительный алфавитно-частотный словарь текста. При этом каждой словоформе приписываются общая частота ее употребления во всем тексте *(F),* число абзацев *(m)* и конкретные номера абзацев (0, 1, 2, 3, 4, 5, 6-й), в которых встретилась эта словоформа.

Принципиальный алгоритм задачи автоматического реферирования и аннотирования текста


Создание словаря потенциальных опорных словоформ текста
Создание словаря главных и второстепенных опорных словоформ текста
Выделение из текста ключевых словосочетаний для аннотации


Выделение из текста ключевых предложений для реферата
Печать аннотации к тексту на экране дисплея
Печать реферата текста на экране дисплея

Распределительный алфавитно-частотный словарь словоформ текста (фрагменты)

Nb п/п	Словоформа	Абсолютная частота употреб-ления, *F*	Общее число абзацев в тексте, *m*	Номера абзацев, в которых встрети-лась словоформа
и	A	25	6	123456
1	AIR-CONDITIONG	1	1	1
2	ALSO	1	1	3
3	AN	2	1	2
4	AND	17	7	0123456
.....				
12	BUILD	2	2	23
18	CALLED	3	3	356

Прежде чем сформулировать решаемую в *блоке D* подзадачу, следует ввести ряд уточнений. Выше, говоря о ключевых словах текста, отмечается три особенности этих слов

это слова-термины

они должны встречаться в тексте несколько раз

должны учитываться все возможные синонимы этого термина в тексте

В исходном тексте могут встречаться пары синонимов. В настоящее время без специальных, очень сложных программ компьютер не может сам их обнаружить. Большую помощь при определении синонимичности слов могут оказать словари-тезаурусы. Поэтому при записи текста на диск эти синонимы указываются. Далее синонимичные словоформы объединяются в одно условное слово с суммированием

- частот их употребления,
- объединением номеров абзацев, в которых они использовались,
- и их общего числа.

Синонимичными для компьютера являются и грамматические формы **одних и тех же слов**. Найти и объединить эти словоформы компьютер может самостоятельно, без специального предредактирования текста человеком. В результате получается единое условное слово. В число потенциальных опорных словоформ текста логично включать только те из словоформ, которые встретились в двух и более абзацах (т.е. $m > 2$).

Наконец, чтобы учесть первую особенность ключевых слов и оставить в числе потенциальных опорных словоформ текста только термины, необходимо из всех словоформ исключить служебные словоформы (предлоги, частицы, артикли, местоимения, наречия, вспомогательные глаголы, числительные и т.д.) и словоформы с общеупотребительным значением. Эти две группы словоформ называют иногда ***отрицательными, запрещенными*** или ***стоп-словами.***

Компьютер последовательно анализирует все словоформы распределительного словаря, начинающиеся с одной и той же буквы.

На первом шаге такого анализа она выделяет у второй из двух сравниваемых словоформ одну последнюю букву и оставшуюся часть второй словоформы сравнивает с первой словоформой.

Если они совпадают, компьютер **суммирует частоты этих словоформ, устанавливает номера абзацев**, в которых встретились эти словоформы, и **определяет общее количество абзацев, в которых они использовались**.

Число выделенных таким образом номеров абзацев и составляет величину *m* для единого условного слова.

Запись двух и более словоформ в виде, когда *F* и *m* указываются лишь для одной формы, а для других словоформ эти величины равны нулю, называется **условным словом**. Номера абзацев в последних словоформах оставлены от результатов предыдущих действий.

Если после отделения у 2-й словоформы одной последней буквы совпадения словоформ не произошло, то у 2-й словоформы выделяются две последние буквы и проводится новое сравнение оставшейся части с 1-й словоформой. В случае их совпадения выполняются аналогичные описанным выше действия.

Объединение данных о синонимичных словоформах проводится с опорой на тот факт, что такие словоформы в тексте по нашему первоначальному условию заключены в скобки и располагаются друг за другом.

Найдя основную словоформу, компьютер объединяет **ее** частоту с частотой **синонима**, уточняет число и конкретные номера абзацев по такому же принципу, как это было показано для объединения грамматических форм одного и того же слова.

Исключаются из распределительного алфавитно-частотного словаря те словоформы, которые встретились лишь в одном абзаце.

Исключение из распределительного словаря любой словоформы служебного или общеупотребительного слова, грамматической формы, синонима, словоформ, встречающихся в одном абзаце, осуществляется в виде **«сжатия» распределительного словаря**, с тем чтобы в нем не осталось ненужных словоформ.

При этом компьютер опирается на заранее заданный список запрещенных слов.

Итогом работы *блока D* является словарь потенциальных опорных словоформ исходного текста.

Основным критерием для создания словаря главных и второстепенных опорных словоформ текста *(блок E)* является коэффициент важности слова

$$K_{\text{важ}} = \frac{F \cdot m}{N \cdot n}.$$

Словарь потенциальных опорных словоформ исходного текста

№ п/п	Словоформа	Абсолютная частота употреб-ления, *F*	Общее число абзацев в тексте, *m*	Номера абзацев, в которых встре-тилась данная словоформа
0	BUILD	2	2	23
I	CHIP	6	3	234
2	CHIPS	10	0	0
3	1C	o	0	0
4	CIRCUIT	5	3	234
5	CIRCUITS	0	0	0
6	COMPONENTS	3	3	234

Среди ключевых словоформ текста может быть установлена следующая иерархия.

Одни из них — ***главные опорные слова*** (ГОС) — являются особенно важными для текста. Они встречаются с наибольшей частотой в большом числе абзацев. Другие опорные слова встречаются с меньшей частотой и в меньшем числе абзацев. Их называют ***второстепенными опорными словами*** (ВОС). Существуют разные методы определения ГОС и ВОС. В качестве критерия их различия используются граничные значения *К1важ* и *К2важ*. Граничные значения этих коэффициентов при таком подходе зависят от числа абзацев текста.

Для текста, содержащего 7 абзацев, эти коэффициенты могут быть вычислены, например, по следующим формулам

$$\frac{9}{N \cdot n} \le K^1_{\text{важ}} < 1;$$

$$\frac{(1/4n+1)^2}{N \cdot n} \le K^2_{\text{важ}} < \frac{9}{N \cdot n},$$

где N – общее число словоупотреблений текста; n – общее число абзацев текста

Для исследуемого текста, имеющего я = 7 абзацев и N=395 словоупотреблений, неравенства преобразуются в следующие

$$\frac{9}{395 \cdot 7} \leq K^1_{\text{важ}} < 1;$$

$$0,0032 \leq K^1_{\text{важ}} < 1;$$

$$\frac{(1/4 \cdot 7 + 1)^2}{395 \cdot 7} \leq K^2_{\text{важ}} < 0,0032;$$

$$0,0027 \leq K^2_{\text{важ}} < 0,0032.$$

Если для каждой словоформы, вошедшей в словарь потенциальных опорных словоформ, вычислить по формуле коэффициент важности $K1_{важ}$ и сравнить его с граничными значениями $K^1_{важ}$ и $K^2_{важ}$ то

главными опорными словоформами текста будем считать те, коэффициент важности которых удовлетворяет неравенству

$$0{,}0032 \le K_{важ} < 1;$$

второстепенными опорными словоформами текста будем считать те, коэффициент важности которых удовлетворяет неравенству

$$0{,}0027 \le K_{важ} < 0{,}0032$$

если же $K_{важ}$ словоформы меньше 0,0027, то эта словоформа относится к числу прочих словоформ текста.

Результатом работы *блока* Е является словарь главных и второстепенных опорных словоформ, представленный в таблице 14. В ней коэффициенты $К_{важ}$ относятся ко всей группе синонимичных словоформ, образующих одно условное слово.

Словарь главных и второстепенных опорных словоформ текста

№ п/п	Тип опорной словоформы	Словоформа	Кваж
0	ГОС	CHIP	0,0065
1	ГОС	CHIPS	—
2	ГОС	1C	—
3	ГОС	CIRCUIT	0,0054
4	ГОС	CIRCUITS	—
5	ГОС	COMPONENTS	0,0033
6	ГОС	COMPUTER	0,0338

Так как аннотация – это кратчайшее изложение содержания исходного, то она должна опираться на наиболее важные в смысловом отношении текстовые единицы, **т.е. на *главные опорные словоформы.*** Они включают несколько имен существительных, одно имя прилагательное, одно причастие и одно сокращение.

В крупных промышленных системах автоматического реферирования и аннотирования в компьютерной памяти находится большой автоматический словарь, в котором каждому слову приписаны различные лексико-грамматические (часть речи, род, число, падеж, время и т.д.), семантические и другие признаки. Поэтому, опираясь на такой словарь, компьютер достаточно легко определит, к какой части речи относится каждая опорная словоформа.

Отнесение определителей к классам имени прилагательного и причастия может быть осуществлено по **автоматическому словарю.**

А выделение среди имен прилагательных и причастий общеупотребительных словоформ может быть сделано лишь путем **сравнения** каждого определения со специальным списком общеупотребительных имен прилагательных и причастий, помещенным в память компьютера.

Подзадачу, которая должна быть решена в блоке *F,* можно сформулировать так:

«Читая последовательно все предло-жения текста, выделить в них ключевые словосочетания (в указанном выше понимании), расположить их по алфавиту и удалить из них одинаковые».

В самом начале работы по автоматическому реферированию и аннотированию текста в память компьютера вводится список предлогов, артиклей, наречий, союзов, числительных, вспомогательных глаголов, местоимений, а также общеупотребительных имен прилагательных и причастий. Обрабатываемый текст уже находится в компьютерной памяти *{блок А)*.

Далее начинается последовательное чтение отдельных предложений. В каждом прочитанном предложении компьютер ищет опорную словоформу — имя существительное.

Если она **найдена**, то компьютер выделяет из предложения словоформу, стоящую перед опорной, и сравнивает ее с введенным ранее в память списком служебных и общеупотребительных словоформ.

Словоформу-определение, **не найденную в таком списке**, компьютер объединяет со стоящим за ней опорным именем существительным и передает полученное словосочетание в специальную область памяти для ключевых словосочетаний текста.

На следующем этапе работы компьютер сортирует этот список по алфавиту. И наконец, на последнем этапе компьютер исключает из рассортированного списка ключевых словосочетаний полностью повторяющиеся словосочетания.

Если в таком списке встречаются словосочетания, различающиеся лишь признаком числа имени существительного *(integrated circuit и integrated circuits),* то в списке остается словосочетание с именем существительным во множественном числе.

В общей формулировке задачи построения системы автомати-ческого реферирования и аннотирования отмечалось, что реферат представляет собой последовательность ключевых предложений текста. Само же ключевое предложение было определено как такое предложение текста, в котором содержится три и более разных опорных слова.

Поэтому подзадачу *блока G* сформулируем следующим образом: «Читая последовательно все предложения текста, выделить и запомнить те из них, в которых содержится три и более разных главных или второстепенных опорных слова данного текста».

Результатом работы *блока G* является ***реферат текста*** — последовательность ключевых предложений, в каждом из которых обнаружено три и более опорных словоформ текста.
В процессе выполнения *блока H* на экран дисплея выводится аннотация текста в виде полученных в *блоке F* ключевых словосочетаний, перед которыми ставится фраза (релятор) «THIS TEXT IS ABOUT:» («В ЭТОМ ТЕКСТЕ ГОВОРИТСЯ:»).

THIS TEXT IS ABOUT: COMPUTER COMPONENTS, ELECTRONIC COMPONENTS, ELECTRONIC COMPUTERS, INTEGRATED CIRCUITS, MEMORY CHIPS, PERSONAL COMPUTER, SEMI-CONDUCTOR CHIP, SPECIAL CIRCUIT.

Выполнение *блока I* позволяет напечатать на экране дисплея все ключевые предложения текста. Они представляют собой его реферат:

AN INTEGRATED CIRCUIT (1C FOR SHORT) WAS SOON DEVELOPED. AN 1C IS A LARGE NUMBER OF TRANSISTORS AND OTHER ELECTRONIC COMPONENTS WIRED TOGETHER. IN-TEGRATED CIRCUITS, ALSO CALLED CHIPS, ARE NOW MANU-FACTURED SEPARATELY FROM COMPUTERS. THE MOST IM-PORTANT OF THESE COMPUTER COMPONENTS IS THE CENTRAL PROCESSING UNIT (CPU FOR SHORT) OR MICROPROCESSOR. MICROPROCESSORS ARE USED IN MINICOMPUTERS AND EACH OF THESE INTEGRATED CIRCUIT IS CAPABLE OF PROCESSING 8-BIT OR 16-BIT DATA. SOME OTHER COMPONENTS OF A COMPUTER ARE COMPUTER MEMORY CHIPS, VOICE SYNTHE-SIZERS THAT PRODUCE SOUNDS SIMILAR TO HUMAN SPEECH AND EVEN INTEGRATED CIRCUITS THAT PERMIT THE COMPUTER TO SEND COMPUTER DATA ON. A TELEPHONE LINE.

С опорой на рассмотренный алгоритм выделяются необходимые переменные и составляется программа автоматического реферирования и аннотирования.