

Лабораторная работа #4

Методы Text Mining (текстовая добыча)

Цель: изучить основы методов Text Mining (текстовой добычи), приобрести навыки работы с методами Text Mining (текстовой добычи) в системе STATISTICA StatSoft, осуществить обработку методами Text Mining индивидуального набора данных и интерпретацию результатов

1 Ход работы

- 1) изучить теоретические сведения
- 2) приобрести навыки работы с методами Text Mining (текстовой добычи) в системе STATISTICA StatSoft, реализуя приведенный ниже пример
- 3) на основе приобретенных практических навыков осуществить все этапы методами Text Mining и интерпретацию результатов согласно варианту индивидуального задания
- 4) оформить отчет и подготовиться к защите лабораторной работы по полученным результатам и контрольным вопросам

2 Содержание отчета и требования к его оформлению

- 1) отчет оформляется в печатном виде
- 2) отчет содержит титульный лист, исходные данные, результаты выполнения этапов обработки данных в виде скриншотов и обязательных комментариев по ходу выполнения работы, выводы - **письменно ответить, чем отличаются полученные результаты от результатов примера, рассчитать параметры классификации**
- 3) к отчету прилагается файл исходных данных *.sta проекта (см. ниже) в электронном виде с целью осуществления выборочного контроля.

3 Варианты исходных данных

- исходные данные – файл Reuters for vars.xlsx, в котором находятся выборочные совокупности, отобранные из генеральной выборки Reuters.sta.

4 Краткие теоретические сведения

Анализ структурированной информации, хранящейся в базах данных, требует предварительной обработки: проектирования БД, ввод информации по определенным правилам, размещение ее в специальных структурах (например, реляционных таблицах) и т. п. Таким образом, непосредственно для анализа этой информации и получения из нее новых знаний необходимо затратить дополнительные усилия. При этом они не всегда связаны с анализом и необязательно приводят к желаемому результату. Из-за этого КПД анализа структурированной информации снижается. Кроме того, не все виды данных можно структурировать без потери полезной информации. Например, текстовые документы практически невозможно преобразовать в табличное представление без потери семантики текста и отношений между сущностями. По этой причине такие документы хранятся в БД без преобразований, как текстовые поля (BLOB-поля). В то же время в тексте скрыто огромное количество информации, но ее неструктурированность не позволяет использовать алгоритмы Data Mining. Решением этой проблемы занимаются методы анализа неструктурированного текста. В западной литературе такой анализ называют Text Mining.

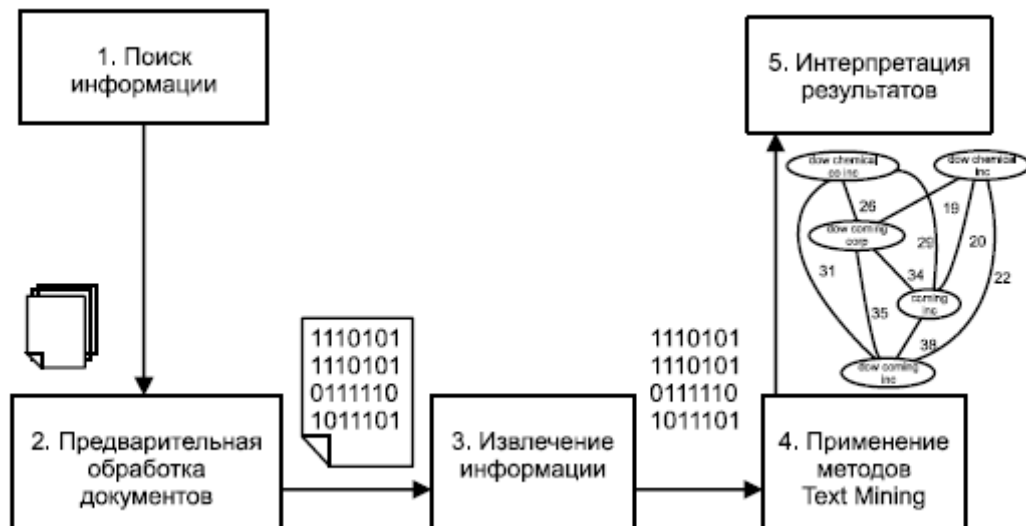
Методы анализа в неструктурированных текстах лежат на стыке нескольких областей: Data Mining, обработка естественных языков, поиск информации, извлечение информации и управление знаниями.

Обнаружение знаний в тексте - это нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных. Как видно, от определения Data Mining, оно отличается только новым понятием "неструктурированные текстовые данные".

Под такими знаниями понимается набор документов, представляющих собой логически объединенный текст без каких-либо ограничений на его структуру. Примерами таких документов являются: Web-страницы, электронная почта, нормативные документы и т. п. В общем случае такие документы могут быть сложными и большими и включать в себя не только текст, но и графическую информацию.

Документы, использующие язык расширяемой разметки XML (eXtensible Markup Language), стандартный язык обобщенной разметки SGML (Standard Generalised Markup Language) и другие подобные соглашения по структуре формирования текста, принято называть полуструктурированными документами. Они также могут быть обработаны методами Text Mining. Процесс анализа текстовых документов можно представить как последовательность нескольких шагов (рис.).

1. Поиск информации. На первом шаге необходимо идентифицировать, какие документы должны быть подвергнуты анализу, и обеспечить их доступность. Как правило, пользователи могут определить набор анализируемых документов самостоятельно - вручную, но при большом количестве документов необходимо использовать варианты автоматизированного отбора по заданным критериям.



2. Предварительная обработка документов. На этом шаге выполняются простейшие, но необходимые преобразования с документами для представления их в виде, с которым работают методы Text Mining. Целью таких преобразований является удаление лишних слов и придание тексту более строгой формы.

3. Извлечение информации. Извлечение информации из выбранных документов предполагает выделение в них ключевых понятий, над которыми в дальнейшем будет выполняться анализ.

4. Применение методов Text Mining. На данном шаге извлекаются шаблоны и отношения, имеющиеся в текстах. Данный шаг является основным в процессе анализа текстов.

5. Интерпретация результатов. Последний шаг в процессе обнаружения знаний предполагает интерпретацию полученных результатов. Как правило, интерпретация заключается или в представлении результатов на естественном языке, или в их визуализации в графическом виде.

Визуализация также может быть использована как средство анализа текста. Для этого извлекаются ключевые понятия, которые и представляются в графическом виде. Такой подход помогает пользователю быстро идентифицировать главные темы и понятия, а также определить их важность.

Предварительная обработка текста

Одной из главных проблем анализа текстов является большое количество слов в документе. Если каждое из этих слов подвергать анализу, то время поиска новых знаний резко возрастет и вряд ли будет удовлетворять требованиям пользователей. В то же время очевидно, что не все слова в тексте несут полезную информацию. Кроме того, в силу гибкости естественных языков формально различные слова (синонимы и т. п.) на самом деле означают одинаковые понятия. Таким образом, удаление неинформативных слов, а также приведение близких по смыслу слов к единой форме значительно сокращают время анализа текстов. Устранение описанных проблем выполняется на этапе предварительной обработки текста.

Обычно используют следующие приемы удаления неинформативных слов и повышения строгости текстов:

- удаление стоп-слов. Стоп-словами называются слова, которые являются вспомогательными и несут мало информации о содержании документа. Обычно заранее составляются списки таких слов, и в процессе предварительной обработки они удаляются из текста. Типичным примером таких слов являются вспомогательные слова и артикли, например: "так как", "кроме того" и т. п.;

- стемминг - морфологический поиск. Он заключается в преобразовании каждого слова к его нормальной форме. Нормальная форма исключает склонение слова, множественную форму, особенности устной речи и т. п. Например, слова "сжатие" и "сжатый" должны быть преобразованы в нормальную форму слова "сжимать". Алгоритмы морфологического разбора учитывают языковые особенности и вследствие этого являются языково-зависимыми алгоритмами;

- N-граммы - это альтернатива морфологическому разбору и удалению стоп-слов. N-грамма - это часть строки, состоящая из N символов. Например, слово "дата" может быть представлено 3-граммой "_да", "дат", "ата", "та_" или 4-граммой "_дат", "дата", "ата_", где символ подчеркивания заменяет предшествующий или замыкающий слово пробел. По сравнению со стеммингом или удалением стоп-слов, N-граммы менее чувствительны к грамматическим и типографическим ошибкам. Кроме того, N-граммы не требуют лингвистического представления слов, что делает данный прием более независимым от языка. Однако N-граммы, позволяя сделать текст более строгим, не решают проблему уменьшения количества неинформативных слов;

- приведение регистра. Этот прием заключается в преобразовании всех символов к верхнему или нижнему регистру. Например, все слова "текст", "Текст", "ТЕКСТ" приводятся к нижнему регистру "текст". Наиболее эффективно совместное применение перечисленных методов.

Задачи Text Mining

В настоящее время в литературе описано много прикладных задач, решаемых с помощью анализа текстовых документов. Это и классические задачи Data Mining: классификация, кластеризация, и характерные только для текстовых документов задачи: автоматическое аннотирование, извлечение ключевых понятий и др.

Классификация (classification) - стандартная задача из области Data Mining. Ее целью является определение для каждого документа одной или нескольких заранее заданных категорий, к которым этот документ относится. Особенностью задачи классификации является предположение, что множество классифицируемых документов не содержит "мусора", т. е. каждый из документов соответствует какой-нибудь заданной категории. Частным случаем задачи классификации является задача определения тематики документа.

Целью *кластеризации (clustering)* документов является автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества. Отметим, что группы формируются только на основе попарной схожести описаний документов, и никакие характеристики этих групп не задаются заранее.

Автоматическое аннотирование (summarization) позволяет сократить текст, сохраняя его смысл. Решение этой задачи обычно регулируется пользователем при помощи определения количества извлекаемых предложений или процентом извлекаемого текста по отношению ко всему тексту. Результат включает в себя наиболее значимые предложения в тексте.

Первичной целью *извлечения ключевых понятий (feature extraction)* является идентификация фактов и отношений в тексте. В большинстве случаев такими понятиями являются имена существительные и нарицательные: имена и фамилии людей, названия организаций и др. Алгоритмы извлечения понятий могут использовать словари, чтобы идентифицировать некоторые термины и лингвистические шаблоны для определения других.

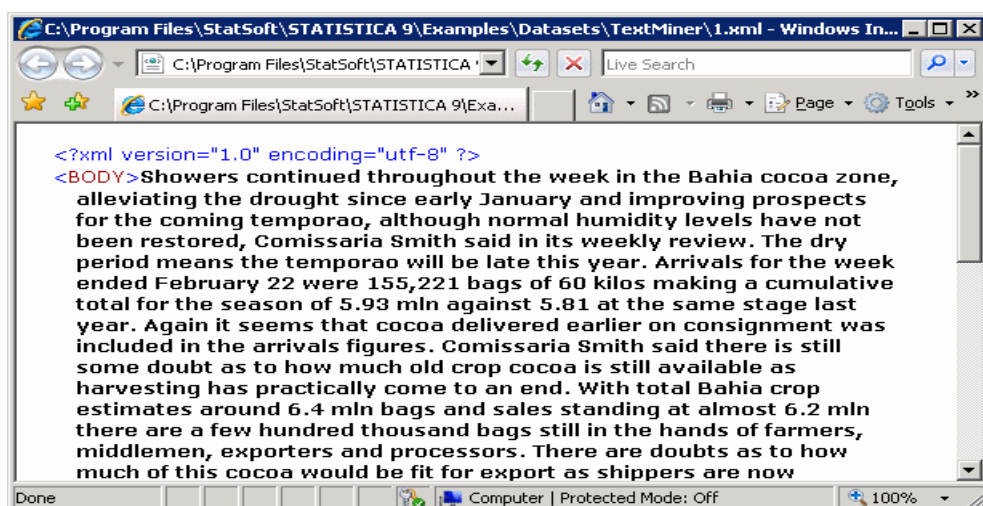
Навигация по тексту (text-base navigation) позволяет пользователям перемещаться по документам относительно тем и значимых терминов. Это выполняется за счет идентификации ключевых понятий и некоторых отношений между ними.

Анализ трендов позволяет идентифицировать тренды в наборах документов на какой-то период времени. Тренд может быть использован, например, для обнаружения изменений интересов компании от одного сегмента рынка к другому.

Поиск ассоциаций также является одной из основных задач Data Mining. Для ее решения в заданном наборе документов идентифицируются ассоциативные отношения между ключевыми понятиями.

5 Пример реализации методов Text Mining в системе STATISTICA StatSoft: автоматическая классификация текста

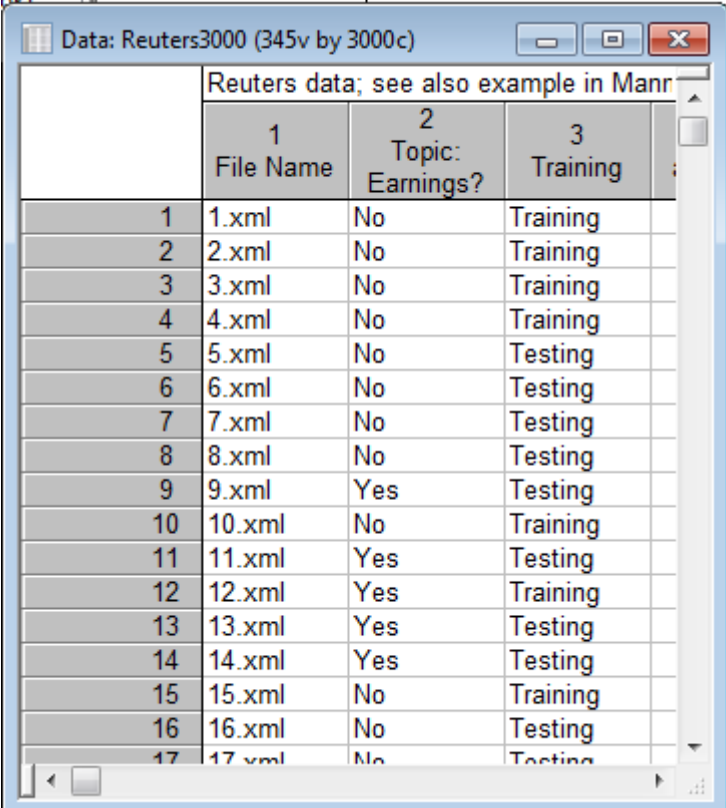
Этот пример основан на «классической» подборке документов Reuters. В частности, 5000 документов были отобраны из базы данных Reuters-21578, которая представляет собой собрание из 21 578 статей из Reuters, появившихся в новостных лентах в 1987 году. Документы были собраны и проиндексированы по категориям сотрудниками Reuters Ltd. в 1987 году. авторские права на эти статьи принадлежат Reuters Ltd. и Carnegie Group, Inc., и эти файлы доступны только для исследовательских и демонстрационных целей. Тело статей было помещено в файлы XML (Extensible Markup Language); ниже показан пример такого файла.



Очевидно, что общая полезность таких методов, которые позволяют автоматически классифицировать большое количество текстов по определенным категориям (например, представляющие интерес или не представляющие интереса; или категории, которые позволяют автоматически направлять документы в соответствующие офисы, отделы и т.д.) может быть весьма значительным. После определения хорошего (точного) метода классификации могут быть сэкономлены значительные ресурсы трудозатрат за счет внедрения автоматизированной системы для выполнения необходимой классификации документов.

Файл данных со ссылками на файлы

Еще раз отметим, что цель этого анализа - получить модель, которая позволит автоматически определять, соответствует ли документ категории «Прибыль». Система интеллектуального анализа текста и поиска документов STATISTICA (STATISTICA Text Mining and Document Retrieval) включает множество опций для поиска документов или ссылок на документы, включая поиск в Интернете или файлах; в данном случае будет использоваться пример файла данных Reuters.sta, который уже содержит необходимую информацию для получения всех документов.



	1 File Name	2 Topic: Earnings?	3 Training
1	1.xml	No	Training
2	2.xml	No	Training
3	3.xml	No	Training
4	4.xml	No	Training
5	5.xml	No	Testing
6	6.xml	No	Testing
7	7.xml	No	Testing
8	8.xml	No	Testing
9	9.xml	Yes	Testing
10	10.xml	No	Training
11	11.xml	Yes	Testing
12	12.xml	Yes	Training
13	13.xml	Yes	Testing
14	14.xml	Yes	Testing
15	15.xml	No	Training
16	16.xml	No	Testing
17	17.xml	No	Testing

Переменная File Name содержит фактические имена файлов, которые необходимо изучить. Вторая переменная «Тема: прибыль?» (Topic: Earnings?) - это то, как эксперты классифицировали каждый документ (как релевантный или не относящийся к прибыли). Кроме того, существует переменная под названием «Обучение» (Training), которая позже будет использоваться во время перекрестной проверки окончательной модели для оценки ее прогностической достоверности и точности.

Определение анализа

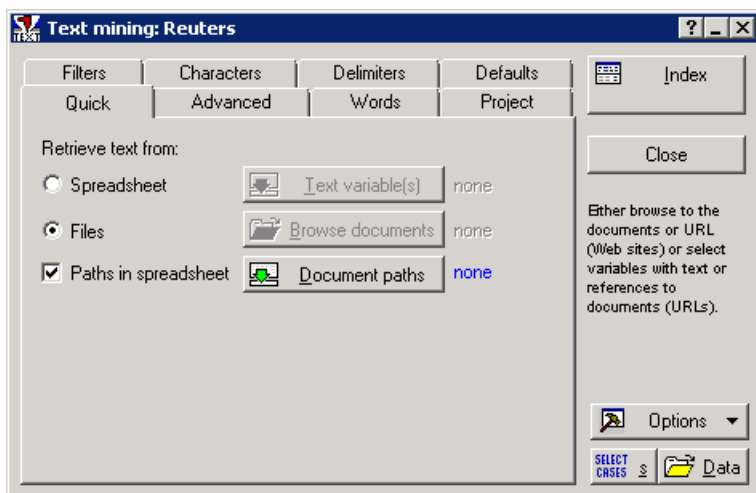
Начните с открытия файла данных примера Reuters3000.sta:

Ribbon bar. Выберите вкладку "Главная" (Home). В группе «Файл» (File) щелкните стрелку «Открыть» (Open) и выберите «Открыть примеры» (Open Examples), чтобы отобразить диалоговое окно «Открыть файл данных STATISTICA» (Open a STATISTICA Data File). Откройте папку Datasets. Файл данных Reuters3000.sta находится в этой папке.

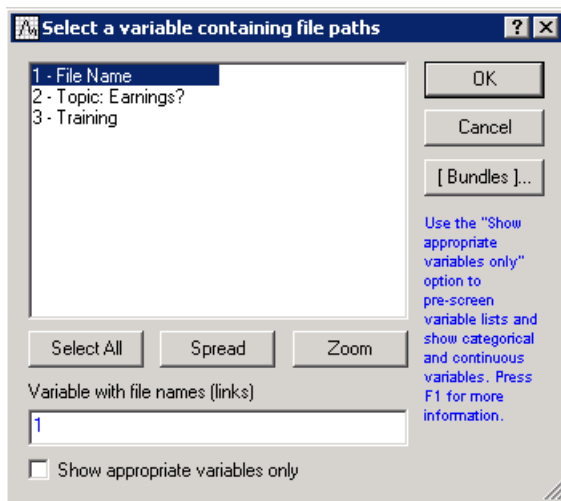
Классическое меню (Classic menus). Выберите «Открыть примеры» (Open Examples) в меню «Файл» (File), чтобы отобразить диалоговое окно «Открыть файл данных STATISTICA» (Open a STATISTICA Data File). Откройте папку Datasets. Файл данных Reuters3000.sta находится в этой папке.

Затем запустите STATISTICA Text Miner: Ribbon bar. Выберите вкладку Data Mining. В группе интеллектуального анализа текста (Text Mining) щелкните интеллектуальный анализ текста (Text Mining), чтобы отобразить панель запуска интеллектуального анализа текста (Text mining Startup Panel).

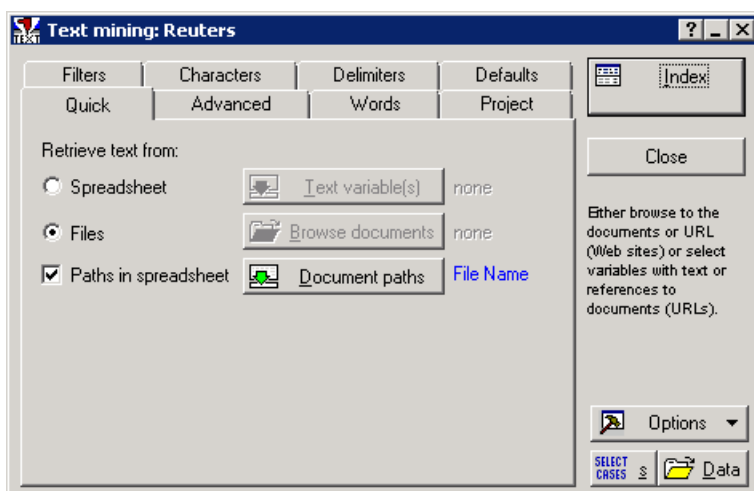
Классическое меню (Classic menus). В меню интеллектуального анализа данных (Data-Mining) выберите интеллектуальный анализ текста и документов (Text & Document Mining), чтобы отобразить панель запуска интеллектуального анализа текста (Text mining Startup Panel). На вкладке Быстрый (Quick) нам нужно указать источник текстовых данных (например, из электронных таблиц, из файлов или из файла в местах, указанных в столбце электронной таблицы): нажмите кнопку выбора файлов (Files) и выберите флажок «Пути в электронной таблице» (Paths in spreadsheet).



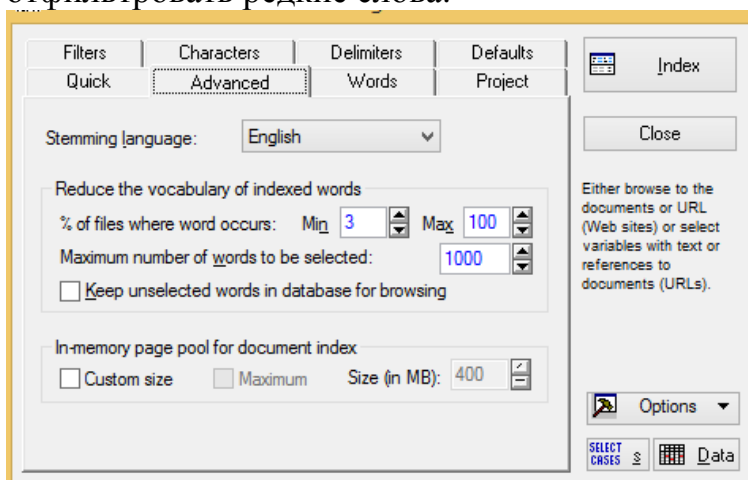
Теперь нажмите кнопку Пути документов (Document paths), чтобы отобразить диалоговое окно выбора переменных, в котором вы выбираете переменную Имя файла (File Name) (которая является переменной, содержащей полные ссылки на файлы [XML] входного документа),



и нажмите кнопку ОК, чтобы вернуться на панель запуска (Startup Panel).



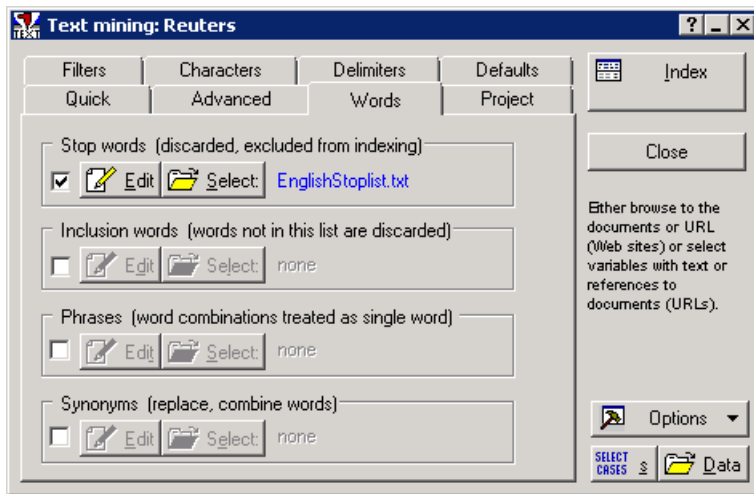
Затем выберите вкладку «Дополнительно» (Advanced), ставите Stemming language - English. Измените параметр % файлов, в которых встречается слово, на 3, чтобы отфильтровать редкие слова.



Теперь выберите вкладку «Слова» (Words) и установите флажок «Стоп-слова (исключено, исключено из индексации)» (Stop words (discarded, excluded from indexing)).

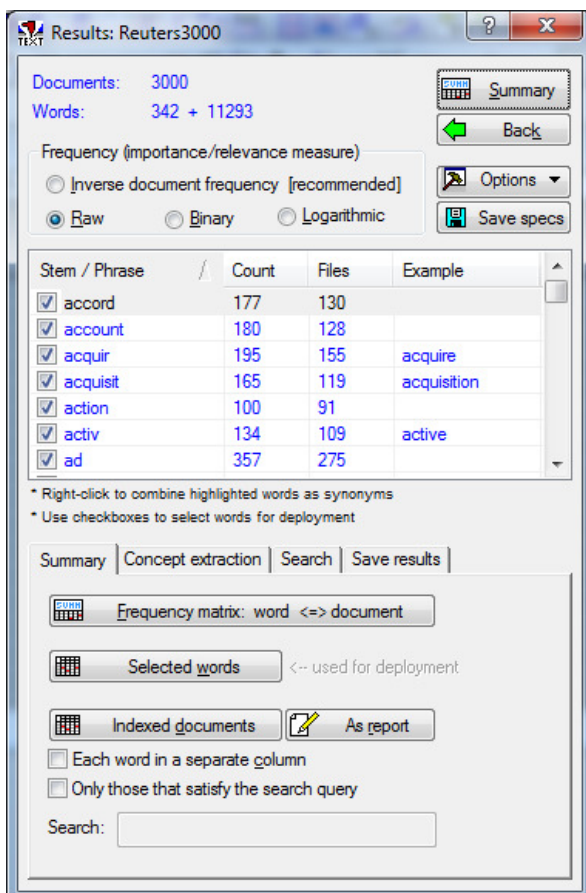
Нажмите соседнюю кнопку Выбрать (Select), чтобы отобразить диалоговое окно Открыть файл стоп-слова (текстовый) (Open stop-word (text) file). Перейдите к файлу EnglishStoplist.txt (который находится в подкаталоге StopLists).

Нажмите кнопку «Открыть» (Open), чтобы загрузить этот файл в качестве стоп-списка по умолчанию, то есть слова и термины, содержащиеся в этом стоп-списке, будут исключены из индексации, которая происходит во время обработки документов.



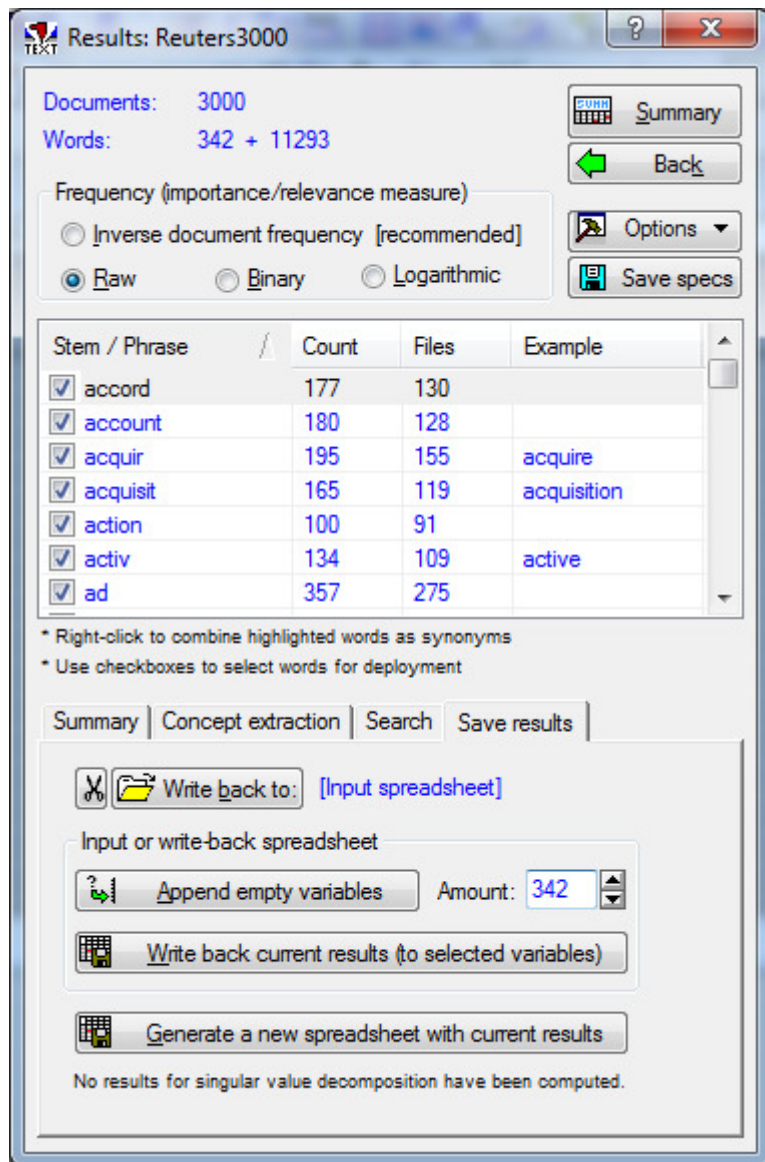
Выполнение анализа данных

Затем нажмите кнопку «Указатель» (Index) на панели запуска (Startup Panel), чтобы начать обработку документов. Через несколько секунд (или минут, в зависимости от скорости аппаратного обеспечения компьютера) отобразится диалоговое окно «Результаты» (Results).



Сохранение извлеченных частот слов во входной файл

Следующим шагом является запись частот извлеченных слов обратно во входной файл, чтобы мы могли использовать эти частоты для дальнейшего анализа. Выберите вкладку Сохранить результаты (Save results). Чтобы записать 342 слова, которые были извлечены обратно во входной файл, нам нужно сначала «освободить место» в файле данных. Для этого введите 342 в поле Amount,



а затем нажмите кнопку «Добавить пустые переменные» (Append empty variables). Если Reuters3000.sta был открыт как файл только для чтения, нам будет предложено сохранить файл в другом каталоге.

Data: Reuters3000* (345v by 3000c)

Reuters data; see also example in Manning and Schutze p. 579

	1 File Name	2 Topic: Earnings?	3 Training	4 NewVar1	5 NewVar2	6 NewVar3	Ne
2985	2985.xml	No	Testing				
2986	2986.xml	No	Testing				
2987	2987.xml	No	Training				
2988	2988.xml	No	Testing				
2989	2989.xml	No	Testing				
2990	2990.xml	No	Training				
2991	2991.xml	No	Training				
2992	2992.xml	No	Testing				
2993	2993.xml	No	Training				
2994	2994.xml	No	Training				
2995	2995.xml	No	Testing				
2996	2996.xml	No	Training				
2997	2997.xml	No	Testing				
2998	2998.xml	No	Testing				
2999	2999.xml	No	Testing				
3000	3000.xml	No	Training				

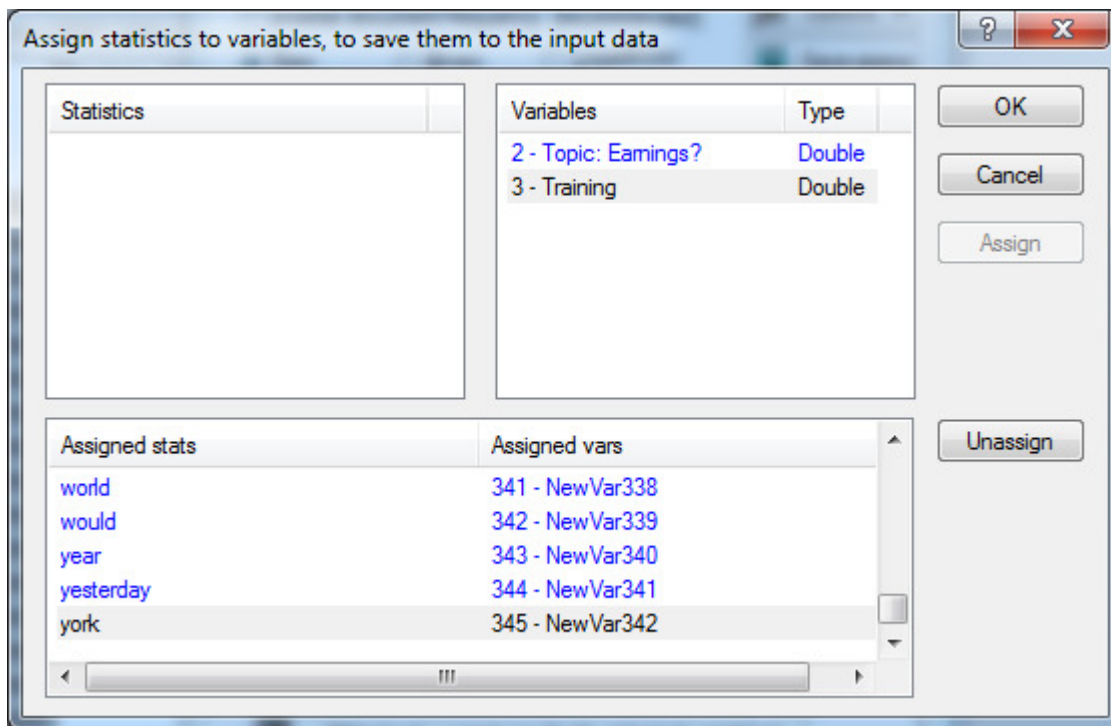
С помощью этой операции к входному файлу будут добавлены 342 пустых новых переменных. Затем нажмите кнопку Записать текущие результаты (в выбранные переменные) (Write back current results (to selected variables)), чтобы отобразить диалог «Назначить статистику переменным, чтобы сохранить их во входных данных» (Assign statistics to variables, to save them to the input data) . Выберите все извлеченные слова (переменные) на левой панели и все вновь созданные переменные на правой панели,

Assign statistics to variables, to save them to the input data

Statistics	Variables	Type	OK
within	339 - NewVar336	Double	Cancel
work	340 - NewVar337	Double	
world	341 - NewVar338	Double	Assign
would	342 - NewVar339	Double	
year	343 - NewVar340	Double	Unassign
yesterday	344 - NewVar341	Double	
york	345 - NewVar342	Double	

Assigned stats	Assigned vars

а затем нажмите «Назначить» (Assign).



Затем нажмите ОК, чтобы завершить эту операцию. Вновь добавленным переменным будут автоматически присвоены соответствующие имена переменных, чтобы отразить соответствующее слово, которое было извлечено, и соответствующие подсчеты частоты будут автоматически записаны в новые переменные.

Data: Reuters3000 (345v by 3000c)

Reuters data; see also example in Manning and Schutze p. 579

	1 File Name	2 Topic: Earnings?	3 Training	4 accord	5 account	6 acquir	ac
1	1.xml	No	Training	0	0	0	
2	2.xml	No	Training	0	0	0	
3	3.xml	No	Training	0	0	0	
4	4.xml	No	Training	0	0	0	
5	5.xml	No	Testing	0	0	0	
6	6.xml	No	Testing	0	0	0	
7	7.xml	No	Testing	0	0	0	
8	8.xml	No	Testing	0	0	0	
9	9.xml	Yes	Testing	0	0	0	
10	10.xml	No	Training	0	0	1	
11	11.xml	Yes	Testing	0	0	0	
12	12.xml	Yes	Training	0	0	0	
13	13.xml	Yes	Testing	0	0	0	
14	14.xml	Yes	Testing	0	0	0	
15	15.xml	No	Training	0	0	1	
16	16.xml	No	Testing	0	0	0	
17	17.xml	No	Testing	0	0	0	

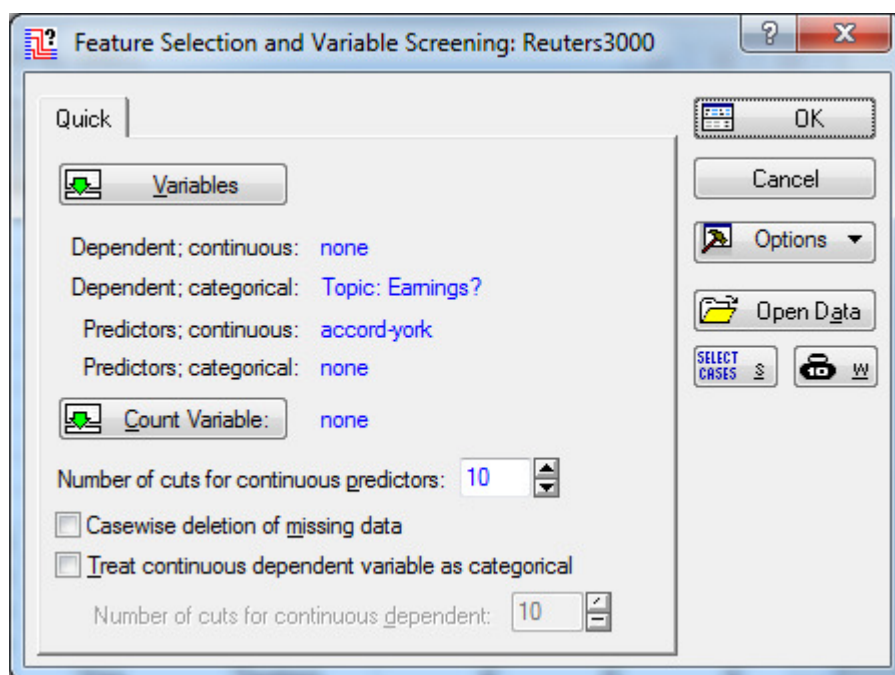
Эти простые шаги завершают конкретную часть этого анализа, посвященную интеллектуальному анализу текста. Остается создать хорошую модель для прогнозирования содержания (доход - да / нет, (Earnings - Yes/No)) новостей, чтобы мы могли автоматически классифицировать их.

Первоначальный выбор функции

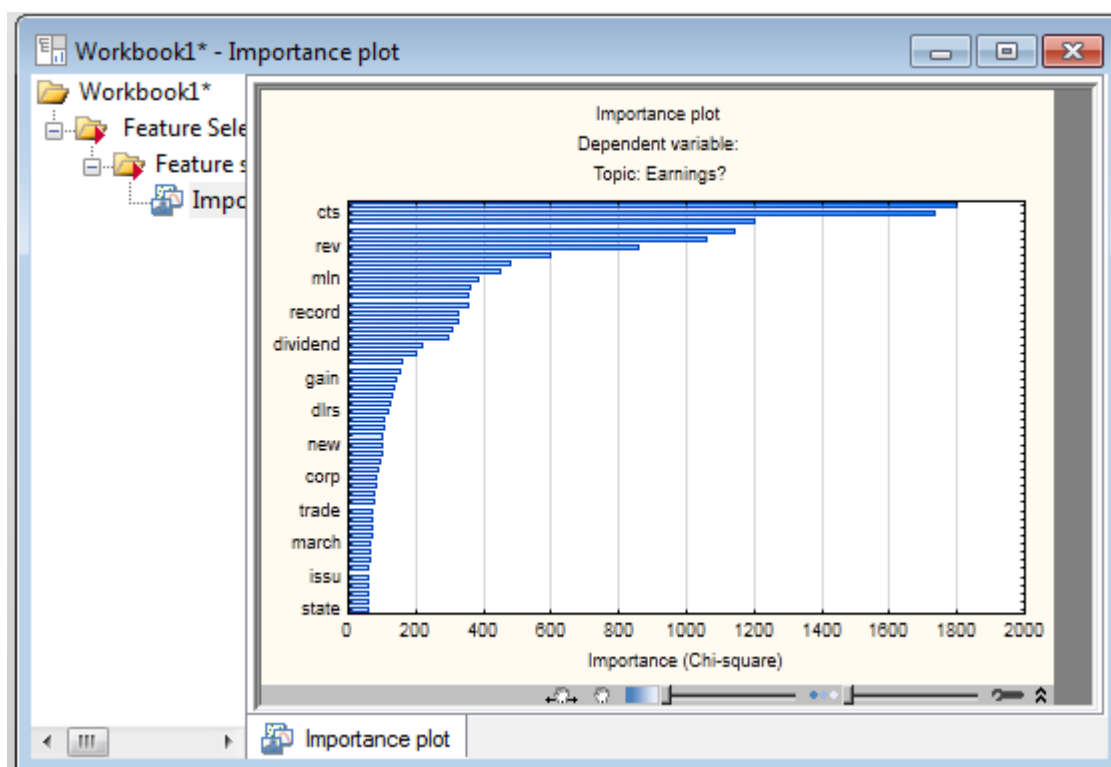
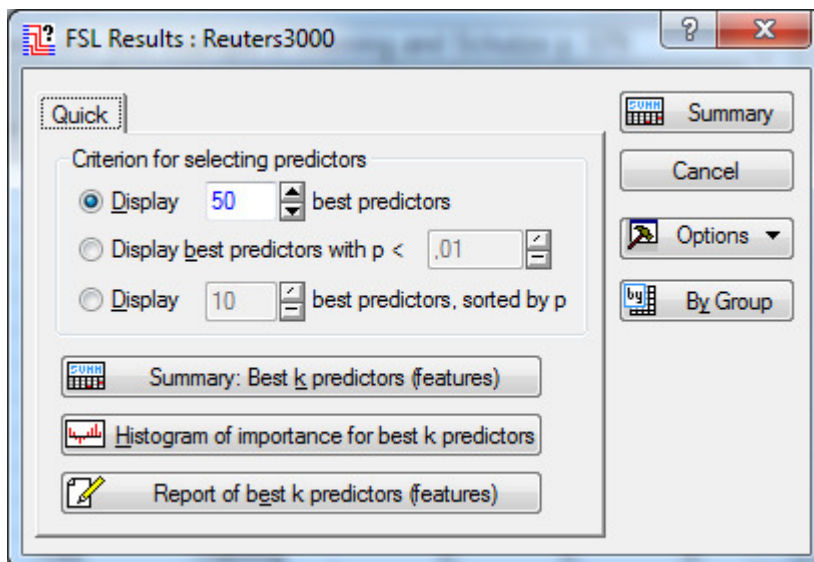
Есть несколько способов, с помощью которых можно продолжить анализ. В качестве первого шага воспользуемся средствами выбора характеристик и скрининга переменных (Feature Selection and Variable Screening), чтобы определить подмножество важных предикторов из 342 слов, которые были извлечены для включения в дальнейшее построение модели.

Следует отметить, что технически в этом нет необходимости, поскольку практически все методы прогнозной классификации, доступные в STATISTICA Data Miner, могут обрабатывать такое количество предикторов. Однако, чтобы проиллюстрировать, насколько быстро могут быть построены модели, давайте сначала воспользуемся методами Feature Selection and Variable Screening.

Выберите Feature Selection and Variable Screening в меню Data-Mining. Затем выберите переменную Тема: Прибыль? (Topic: Earnings?) как категориальную зависимую переменную и все переменные, содержащие количество слов (которое мы записали обратно во входные данные) как непрерывные предикторы.

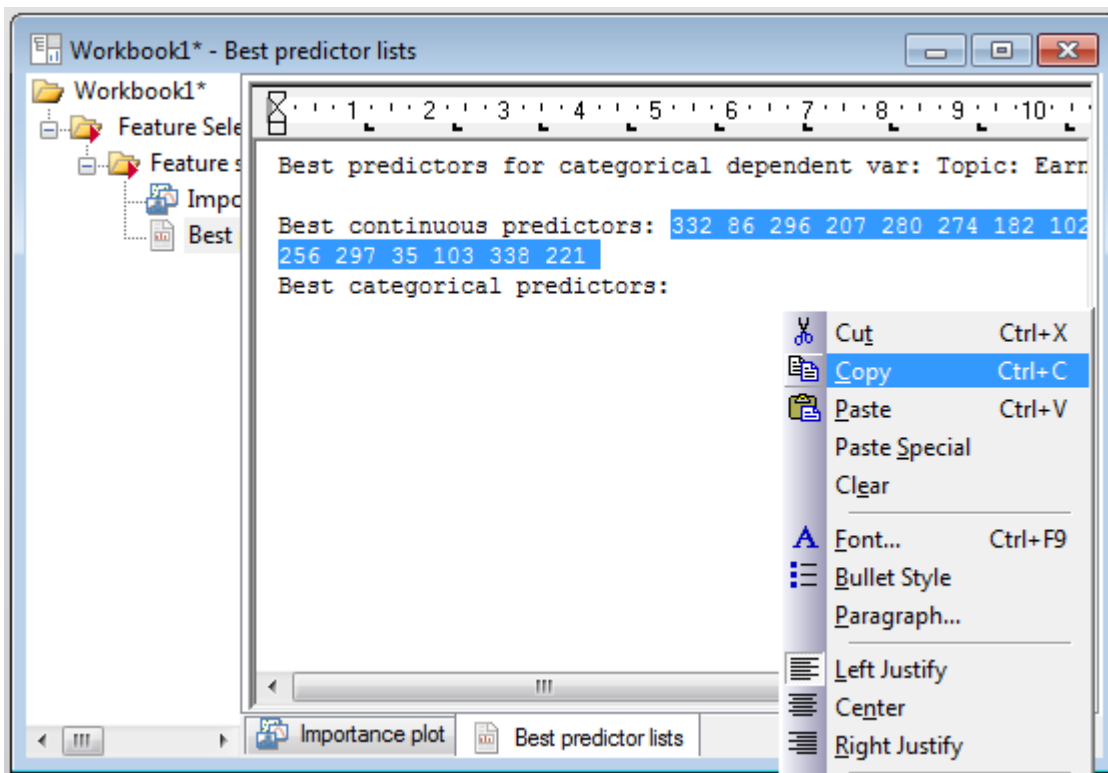


Затем нажмите ОК в диалоговом окне «Выбор функций и переменных» (Feature Selection and Variable Screening), чтобы отобразить диалоговое окно «Результаты FSL» (FSL Results). Укажите, чтобы отобразить 50 лучших предикторов Тема: Прибыль? (Topic: Earnings?) (введите 50 в поле «Показать» (Display)) и создайте график важности предиктора (нажмите кнопку «Гистограмма важности для k лучших предикторов» (Histogram of importance for best k predictors)).



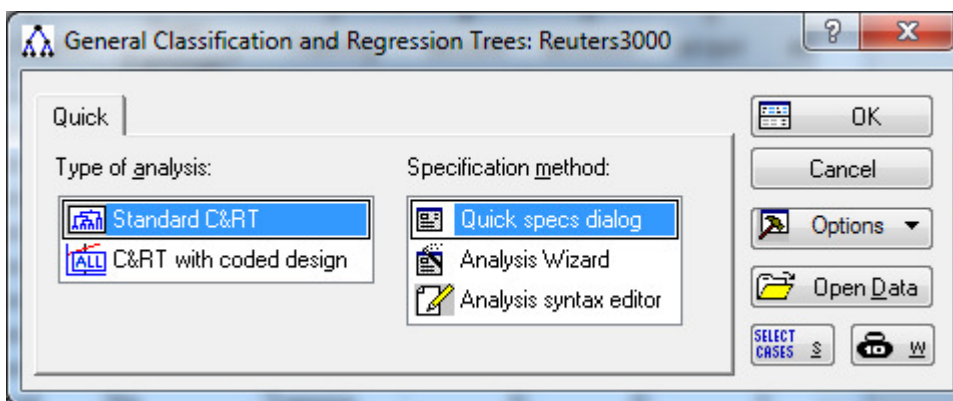
Судя по этому графику, может быть достаточно взять только первые 20 или около того предикторов для окончательного моделирования. Мы будем использовать 20 лучших переменных (слов) в качестве предикторов для дальнейшего построения модели, в частности, чтобы использовать деревья классификации и регрессии (Classification and Regression Trees) для построения окончательной прогнозной модели.

В поле «Показать» (Display) укажите отображение 20 предикторов и нажмите кнопку «Отчет о k лучших предикторах (функций)» (Report of best k predictors (features)), чтобы отобразить список лучших предикторов в отчете. Скопируйте 20 предикторов в буфер обмена, чтобы использовать их в анализе общих деревьев классификации и регрессии (General Classification and Regression Trees GC&RT).

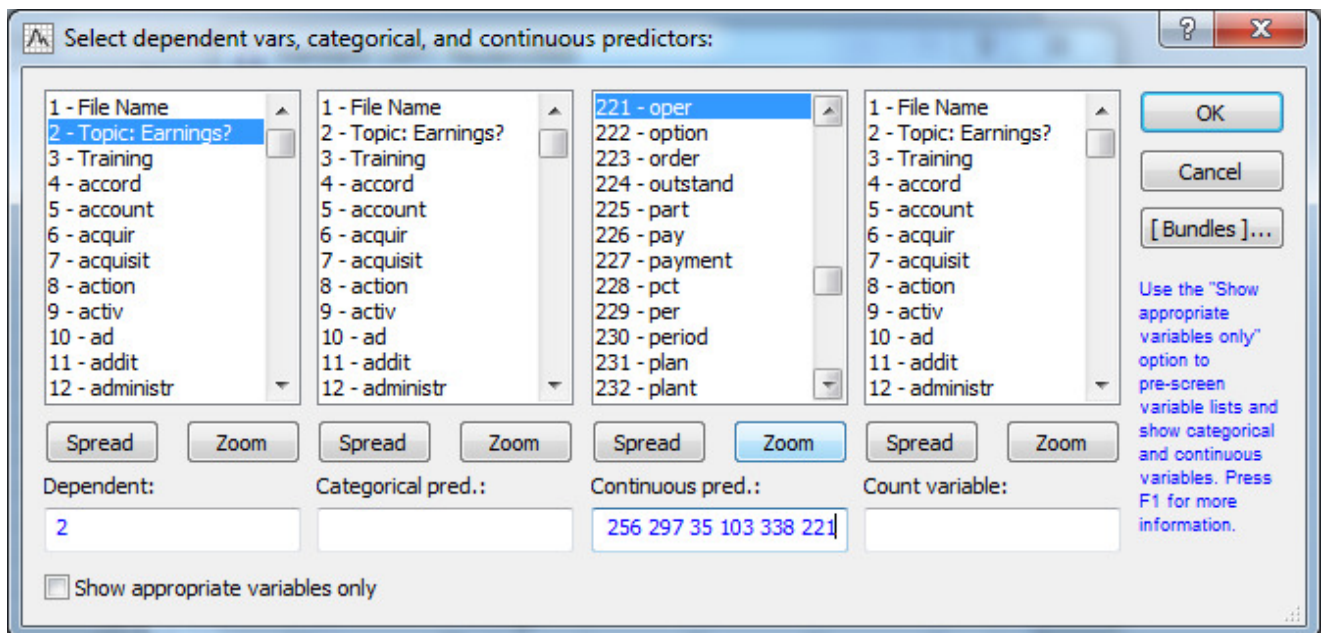


Общая классификация и деревья регрессии

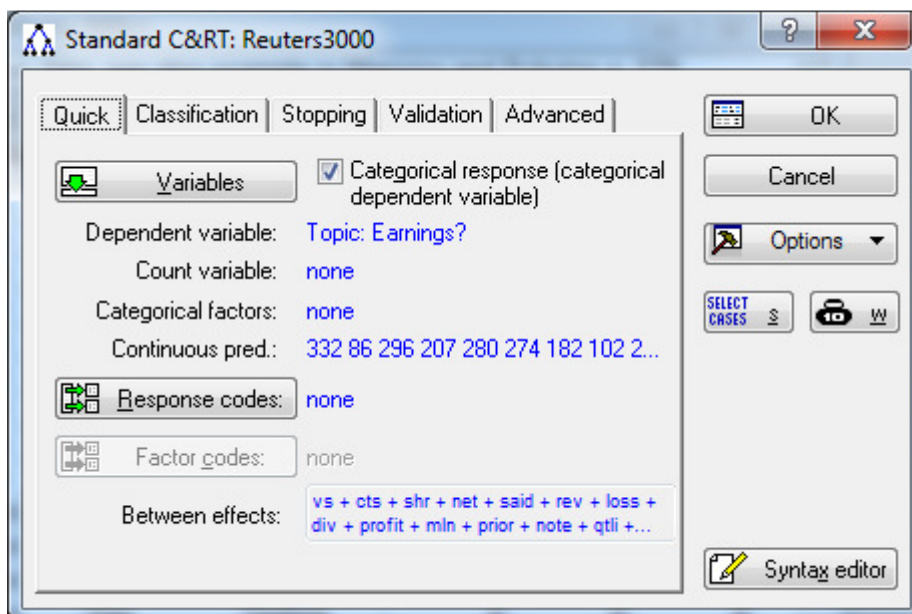
Выберите General Classification / Regression Trees Models в меню Data Mining. Стандартный C&RT выбран по умолчанию.



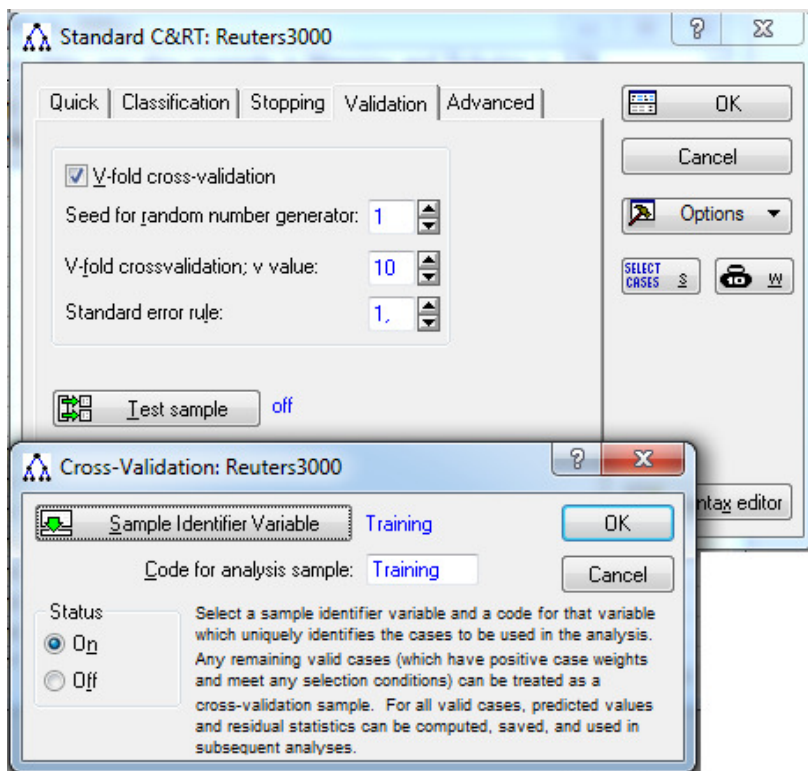
Нажмите кнопку ОК, чтобы отобразить диалоговое окно Standard C&RT, установите флажок Категориальный ответ (категориальная зависимая переменная) (Categorical response (categorical dependent variable)), нажмите кнопку Переменные (Variables) и выберите переменную Тема: Доходы? (Topic: Earnings?) в качестве зависимой (Dependent) переменной и в качестве непрерывных предикторов (Continuous predictors) выберите 20 лучших предикторов (вставьте их в диалоговое окно выбора переменных из буфера обмена), полученных на основе анализа выбора функций и проверки переменных (Feature Selection and Variable Screening).



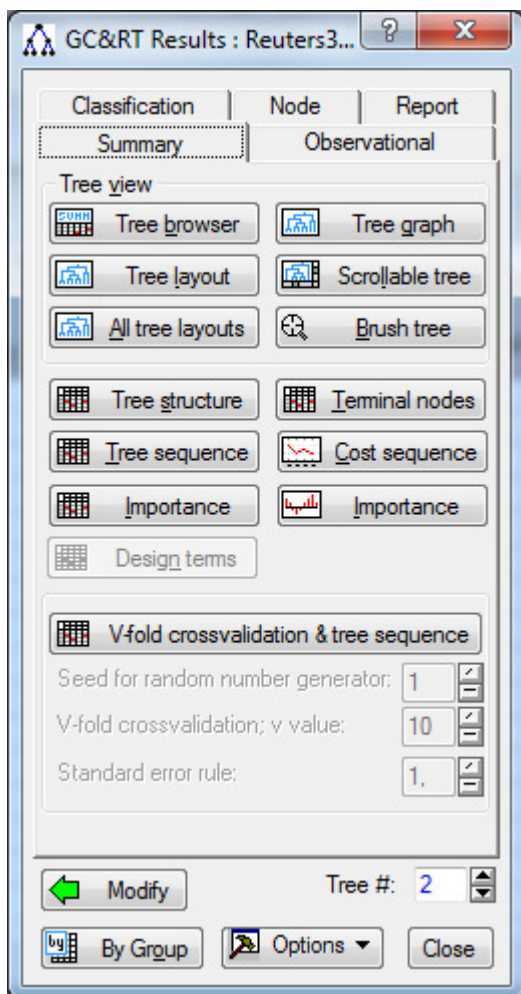
Щелкните ОК.



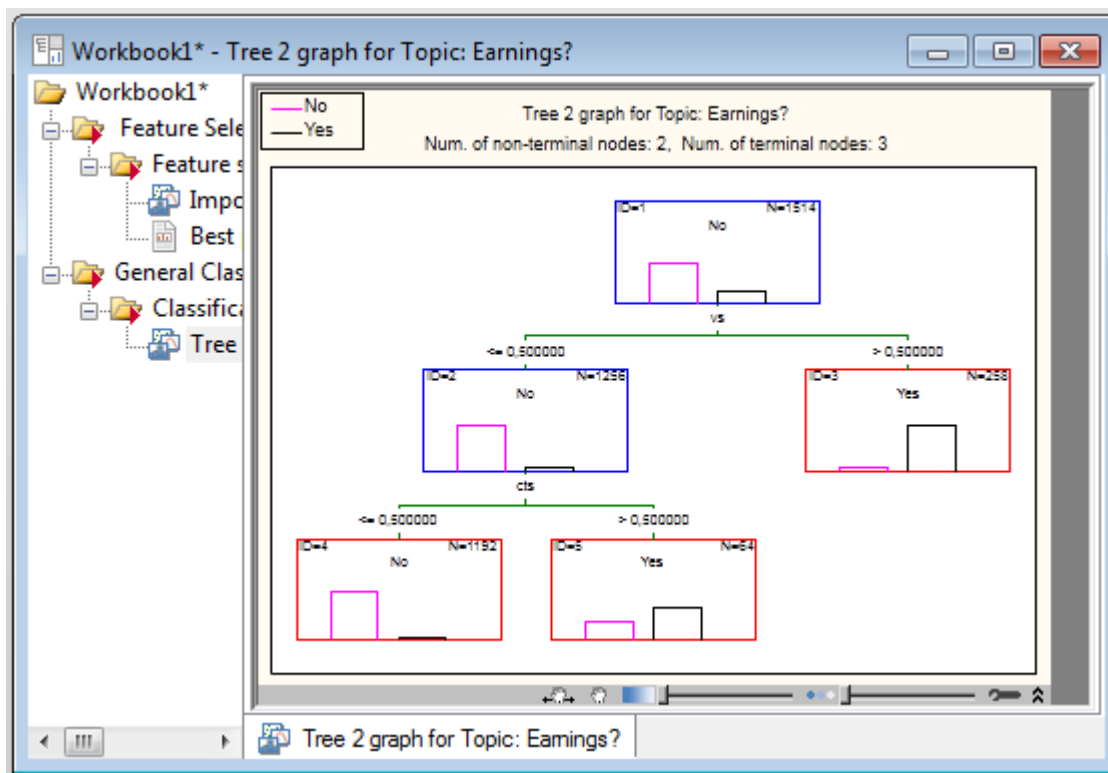
Затем на вкладке Validation установите флажок V-fold crossvalidation (для автоматического выбора надежной модели), а также укажите переменную Training в качестве переменной тестового образца (Test sample) с кодом Training, чтобы определить образец, из которого будем строить модель (будем использовать оставшиеся случаи для проверки предсказательной достоверности модели).



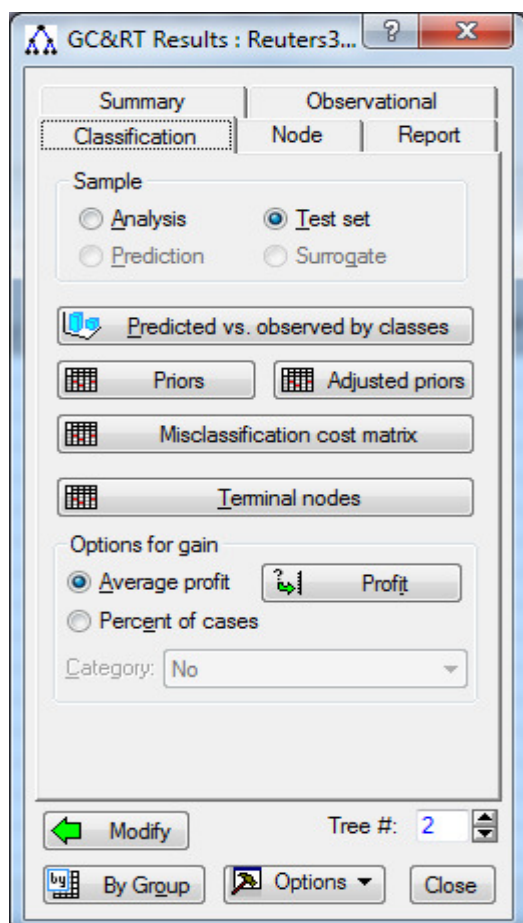
Теперь нажмите ОК в диалоговом окне Standard C&RT, чтобы начать анализ. Через несколько секунд отобразится диалоговое окно результатов GC&RT (GC&RT Results).



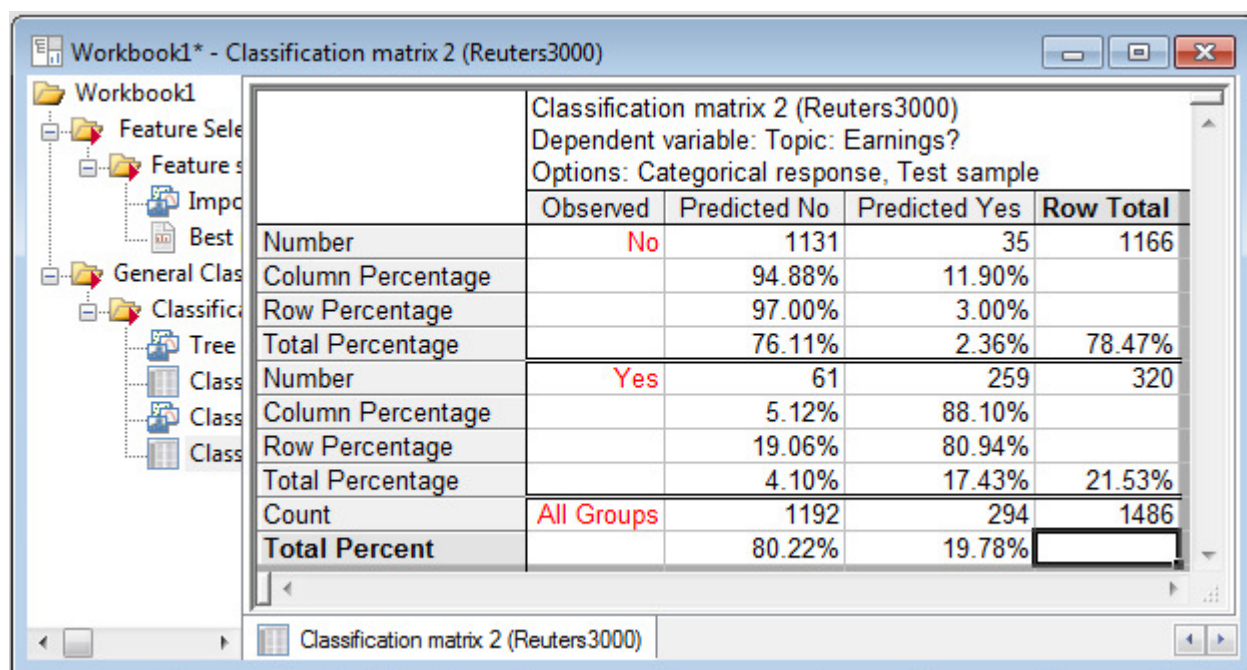
Нажмите кнопку «Древовидный график» (Tree graph) на вкладке «Сводка» (Summary), чтобы просмотреть окончательное дерево.



Выберите вкладку Classification диалогового окна GC&RT Results, выберите опцию Test set, чтобы вычислить прогнозируемую классификацию для (отложенной) тестовой выборки, и нажмите кнопку Predicted vs. observed by classes).



Вычисляется следующая матрица классификации.



Classification matrix 2 (Reuters3000)				
Dependent variable: Topic: Earnings?				
Options: Categorical response, Test sample				
	Observed	Predicted No	Predicted Yes	Row Total
Number	No	1131	35	1166
Column Percentage		94.88%	11.90%	
Row Percentage		97.00%	3.00%	
Total Percentage		76.11%	2.36%	78.47%
Number	Yes	61	259	320
Column Percentage		5.12%	88.10%	
Row Percentage		19.06%	80.94%	
Total Percentage		4.10%	17.43%	21.53%
Count	All Groups	1192	294	1486
Total Percent		80.22%	19.78%	

Интерпретация результатов

Из матрицы классификации можно сделать следующие выводы.

- количество случаев, когда классификация с ответом No (Predicted No) правильная, т.е. соответствует эталонам с ответом No (Observed No) - 1131;
- количество случаев, когда классификация с ответом Yes (Predicted Yes) правильная, т.е. соответствует эталонам с ответом Yes (Observed Yes) - 259;
- общее количество случаев успешной классификации (прогноза) - $1131 + 259 = 1390$;
- количество случаев, когда классификация с ответом No (Predicted No) неправильная, т.е. соответствует эталонам с ответом Yes (Observed Yes) - ошибка 1 рода - 61;
- количество случаев, когда классификация с ответом Yes (Predicted Yes) неправильная, т.е. соответствует эталонам с ответом No (Observed No) - ошибка 2 рода - 35;
- общее количество случаев неуспешной классификации (прогноза) - $61 + 35 = 96$;
- общее количество случаев (элементов выборки, Count) - 1486.

Отсюда можно рассчитать точность прогноза:

$$P = (1390 / 1486) * 100 \% = 93,5 \%$$

Общая ошибка прогноза:

$$P = (96 / 1486) * 100 \% = 6,5 \%$$

Ошибка прогноза 1 рода:

$$P = (61 / 1486) * 100 \% = 4,1 \%$$

Ошибка прогноза 2 рода:

$$P = (35 / 1486) * 100 \% = 2,4 \%$$

Таким образом, методами Data Mining осуществлена весьма точная классификация текстовой информации из информационных массивов большого объема.

Резюме

Таким образом, в этом примере показано, как различные методы в STATISTICA Text и Document вместе со STATISTICA Data Miner могут использоваться для построения высокоточных прогнозных моделей для классификации текста.

6 Контрольные вопросы

1. Понятия, цели и задачи Text Mining
2. Каковы этапы осуществления Text Mining?
3. Каковы цели предварительной обработки текста?
4. Что из себя представляет удаление стоп-слов, стемминг, N-граммы, приведение регистра?
5. Какова цель классификация (classification) как задачи Text Mining?
6. Какова цель кластеризации (clustering) как задачи Text Mining?
7. Что представляет собой автоматическое аннотирование (summarization) в Text Mining?
8. Какова цель анализа трендов позволяет идентифицировать тренды в наборах документов в Text Mining?
9. Какова цель поиска ассоциаций в Data Mining?
10. Пусть в результате реализаций этапов Text Mining получена следующая матрица классификации:

	Observed	Predicted No	Predicted Yes	Row Total
Number	No	1131	35	1166
Column Percentage		94.88%	11.90%	
Row Percentage		97.00%	3.00%	
Total Percentage		76.11%	2.36%	78.47%
Number	Yes	61	259	320
Column Percentage		5.12%	88.10%	
Row Percentage		19.06%	80.94%	
Total Percentage		4.10%	17.43%	21.53%
Count	All Groups	1192	294	1486
Total Percent		80.22%	19.78%	

Требуется рассчитать (или указать в таблице):

10.1 общее количество случаев успешной классификации (прогноза);

10.2 общее количество случаев неуспешной классификации (прогноза);

10.3 общее количество случаев (элементов выборки);

11. Как рассчитать по таблице (см. выше) точность прогноза?

12. Как рассчитать по таблице (см. выше) общую ошибку прогноза?