Министерство образования Республики Беларусь

Учреждение образования

Брестский государственный технический университет

Кафедра ИИТ

Лабораторная работа №4

За 7 семестр

По дисциплине «Методы обработки данных»

Тема: «Методы Text Mining»

Выполнил:

Студент 4 курса, группы ПО-7(1)

Вощук А.Я.

Проверил:

Савицкий Ю.В.

Брест, 2023

**ЛАБОРАТОРНАЯ РАБОТА №4**

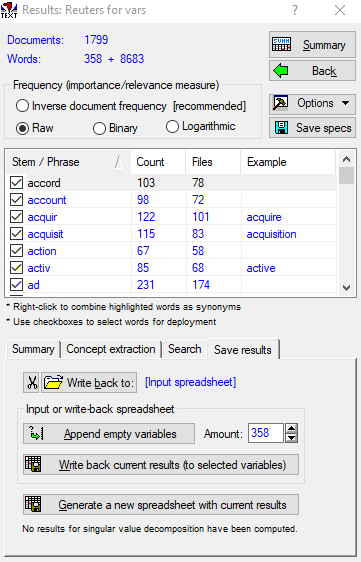
**Методы Text Mining**

**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.

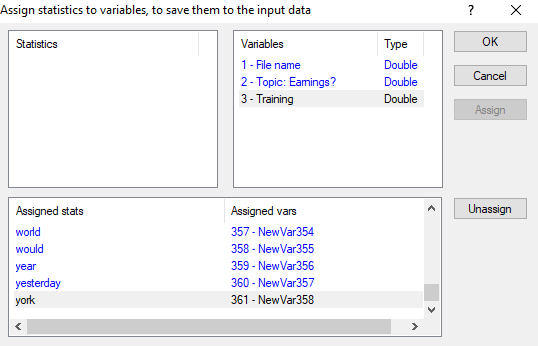
**Вариант 1**

**Ход работы:**

Проведём анализ данных согласно варианту:

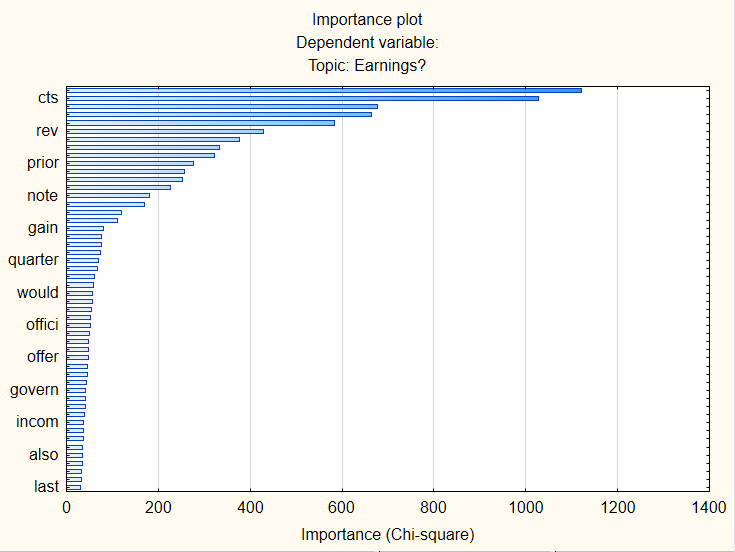


Далее сохраним извлечённых частот слов во входной файл:



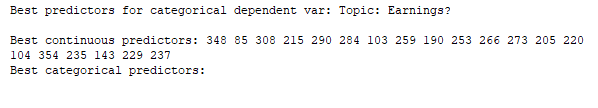
Эти простые шаги завершают конкретную часть этого анализа, посвященную интеллектуальному анализу текста. Остается создать хорошую модель для прогнозирования содержания новостей, чтобы мы могли автоматически классифицировать их.

Далее определим 50 лучших предикатов:

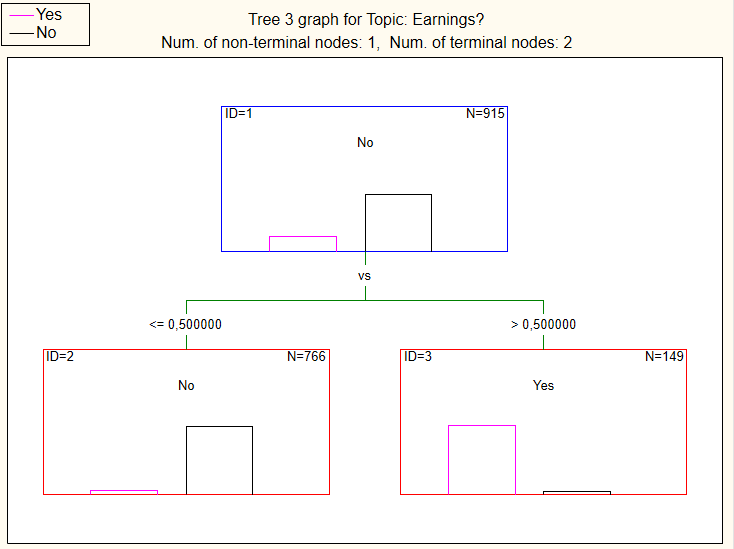


Судя по этому графику, может быть достаточно взять только первые 20 или около того предикторов для окончательного моделирования. Мы будем использовать 20 лучших переменных в качестве предикторов для дальнейшего построения модели, в частности, чтобы использовать деревья классификации и регрессии для построения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

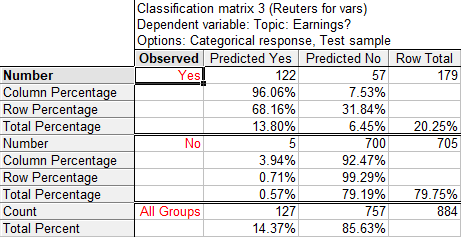
Получим эти 20 предикатов:



Далее определим по 20 лучшим предикатам древовидный график:



Далее вычислим следующую матрицу классификации:



Интерпретируем результаты

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) - 122;

- количество случаев, когда классификация с ответом Yes (Predicted Yes) правильная, т.е. соответствует эталонам с ответом Yes (Observed Yes) - 700;

- общее количество случаев успешной классификации (прогноза) - 122+700=822;

- количество случаев, когда классификация с ответом No (Predicted No) неправильная, т.е. соответствует эталонам с ответом Yes (Observed Yes) - ошибка 1 рода - 5;

- количество случаев, когда классификация с ответом Yes (Predicted Yes) неправильная, т.е. соответствует эталонам с ответом No (Observed No) - ошибка 2 рода - 57;

- общее количество случаев неуспешной классификации (прогноза) – 5+57=62;

- общее количество случаев (элементов выборки, Count) - 884.

Отсюда можно рассчитать точность прогноза:

P = (822 / 884) \* 100 % = 92,98 %

Общая ошибка прогноза:

P = (62 / 884) \* 100 % = 7,013 %

Ошибка прогноза 1 рода:

P = (5 / 884) \* 100 % = 0,565 %

Ошибка прогноза 2 рода:

P = (57 / 884) \* 100 % = 6,48 %

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

**Вывод**: изучил основы методов Text Mining (текстовой добычи), приобрёл навыки работы с методами Text Mining (текстовой добычи) в системе STATISTICA StatSoft, осуществил обработку методами Text Mining индивидуального набора данных и интерпретировал результаты.