

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ»
ФАКУЛЬТЕТ ПОВЫШЕНИЯ КВАЛИФИКАЦИИ И ПЕРЕПОДГОТОВКИ КАДРОВ
МЕЖДУНАРОДНЫЙ НАУЧНО-МЕТОДИЧЕСКИЙ ЦЕНТР

ПРОГРАММА ПРОФЕССИОНАЛЬНОЙ ПЕРЕПОДГОТОВКИ
«Большие данные и цифровой образовательный инжиниринг»
(288 часов)

Балыбердин Алексей Сергеевич
ИТОГОВАЯ АТТЕСТАЦИОННАЯ РАБОТА
НА ТЕМУ

«Применение алгоритмических и инструментальных средств машинного
обучения для модернизации учебной программы курса "Системы
автоматизированного проектирования в разработке технологического
оборудования" на основе анализа вакансий»

 Балыбердин А.С.

_____ Варфоломеев А.А.

Итоговая аттестационная работа защищена
«__» _____ 2021 г.

Оценка _____

Председатель аттестационной комиссии

_____ / _____

Содержание

Введение.....	3
Анализ рабочей программы инструментами Knime Analytics Platform	4
Аннотация рабочей программы дисциплины	21
Обсуждение и анализ полученных результатов	23
Заключение	24
Список использованных источников	25

Введение

Рабочая программа дисциплины для повышения качества образования студентов должна реализовать профессиональный стандарт, лежащий в основе учебного плана, и рекомендации работодателя. В частности, в работе рассматривается рабочая программа курса "Системы автоматизированного проектирования в разработке технологического оборудования" для направления 15.03.02 Технологические машины и оборудование.

Для оптимизации рабочей программы с требованиями работодателя в предлагаемой работе сравнивается текст рабочей программы или ее аннотация с описанием различных вакансий на <https://www.hh.ru/> а в качестве инструмента для анализа применяется Knime Analytics Platform.

Текст часто называют «неструктурированными» данными. Это относится к тому факту, что текст не имеет той структуры, которую мы обычно ожидаем от данных: таблицы записей с полями, имеющими фиксированное значение (по сути, коллекции векторов признаков), а также ссылки между таблицами. У текста, конечно, много структуры, но это языковая структура, предназначенная для потребления человеком, а не для компьютеров. Слова могут иметь разную длину, а текстовые поля могут содержать разное количество слов. Иногда порядок слов имеет значение, иногда нет. Люди пишут без грамматики, они неправильно пишут слова, они объединяют слова вместе, они непредсказуемо сокращают и расставляют произвольные знаки препинания. Текст может содержать синонимы (несколько слов с одинаковым значением) и омографы (одно написание используется для нескольких слов с разными значениями).

Общая стратегия интеллектуального анализа текста заключается в использовании простейшего (наименее затратного) метода, который работает. По сути, мы берем набор документов, каждый из которых представляет собой последовательность слов относительно свободной формы, и превращаем его форму векторов признаков. Каждый документ - это один экземпляр, но мы не знаем заранее, какими будут его функции. Подход, который использован в работе, называется «Bag of Words» или «мешком слов». Как следует из названия, подход состоит в том, чтобы рассматривать каждый документ как просто набор отдельных слов. Этот подход игнорирует грамматику, порядок слов, структуру предложения и (обычно) пунктуацию. Он обрабатывает каждое слово в документе как потенциально важное ключевое слово документа. Генерация представления проста и недорога, и, как правило, хорошо подходит для многих задач. В суть метода - каждое слово является токеном, и каждый документ представлен единицей (если токен присутствует в документе) или нулем (токен отсутствует в документе). Такой подход просто сокращает документ до набора содержащихся в нем слов.

Для того чтобы реализация метода была более качественной необходимо очистить текст от не несущих «нагрузку» терминов и знаков, т.е. исключить из текста знаки препинания, цифры, предлоги, союзы и т.д.

На рис. 1 представлена схема анализа рабочей программы и вакансий, с прогнозированием какая вакансия ближе к рабочей программе преподаваемой дисциплины.

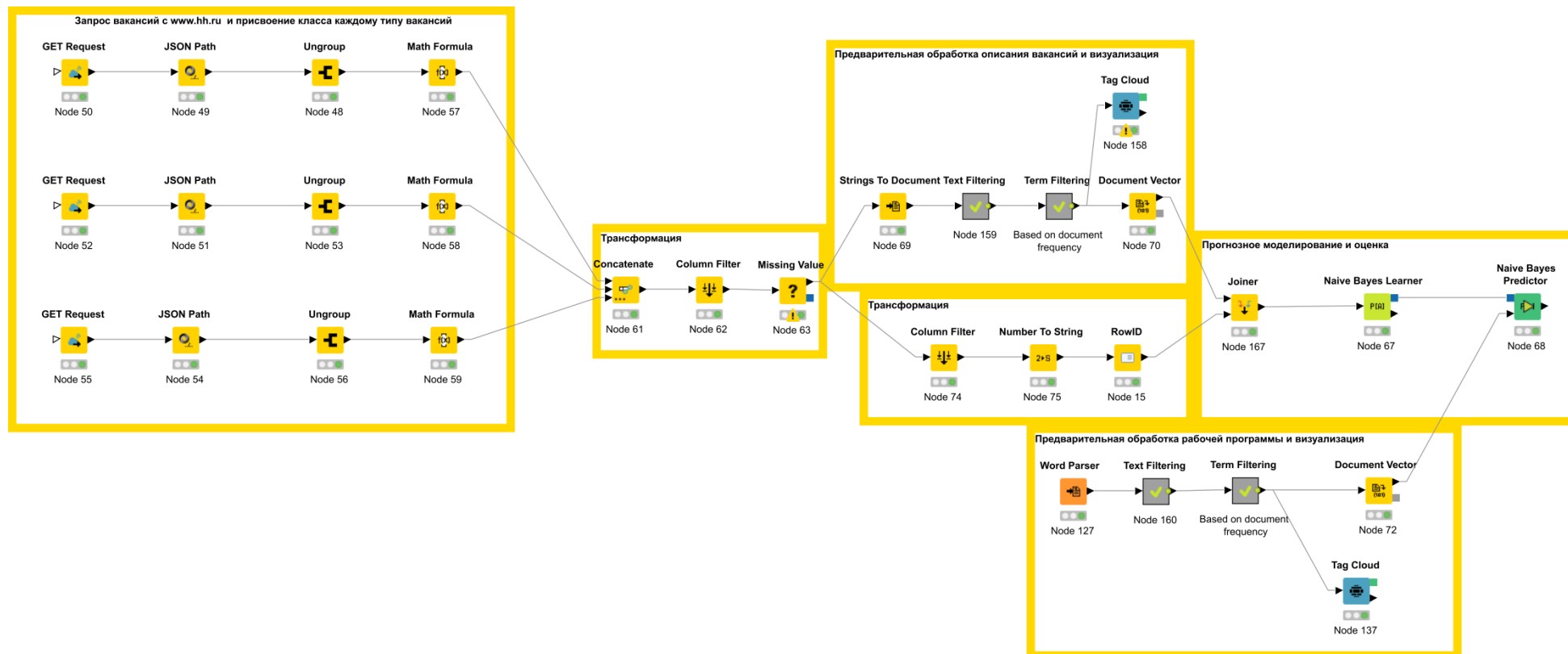


Рис. 1 Схема анализа рабочей программы и вакансий

Схема рис. 1 представляет собой последовательное выполнение каждого узла. Схема состоит из следующих блоков:

- Блок – Запроса вакансий с www.hh.ru и присвоением класса каждому типу вакансий;
- Блок – Трансформации полученной информации о вакансиях;
- Блок – Предварительной обработки описания и визуализации вакансий;
- Блок – Предварительной обработки рабочей программы и ее визуализации;
- Блок – Прогнозирования и оценки вакансий с рабочей программой.

Рассмотрим более подробно каждый блок.

Блок – Запроса вакансий с www.hh.ru и присвоением класса каждому типу вакансий:

Узел предназначен для отправки запроса get к веб-службе. В частности запросы выглядят следующим образом:

https://api.hh.ru/vacancies?area=1&per_page=100&only_with_salary=true&text=компьютерные&search_field=name

https://api.hh.ru/vacancies?area=1&per_page=100&only_with_salary=true&text=инженерANDтехнолог&search_field=name

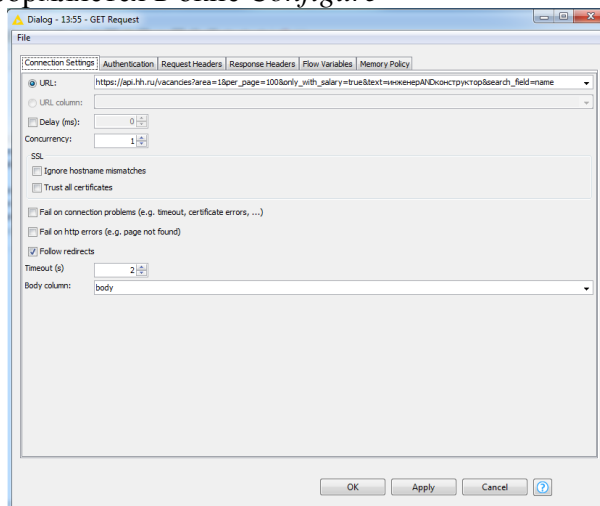
https://api.hh.ru/vacancies?area=1&per_page=100&only_with_salary=true&text=инженерANDконструктор&search_field=name

Запрос оформляется в окне *Configure*

GET Request



Node 55

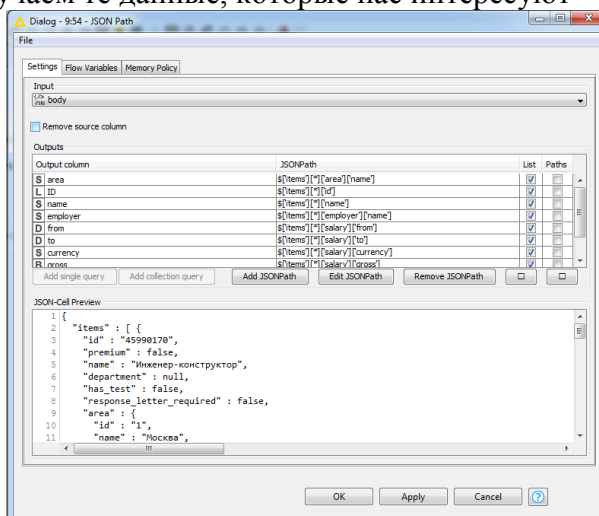


Узел предназначен для запросов *JSON*. *JSON* — текстовый формат обмена данными, основанный на *JavaScript*. Поэтому когда мы делаем запрос, то получаем те данные, которые нас интересуют

JSON Path



Node 54

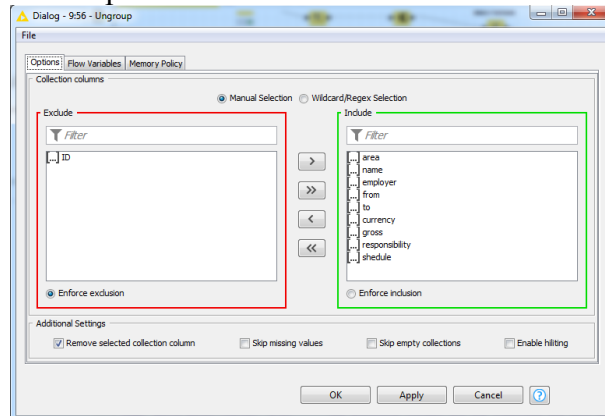


Разбивает список значений на строки со значениями в зависимости от выбранных переменных

Ungroup



Node 56

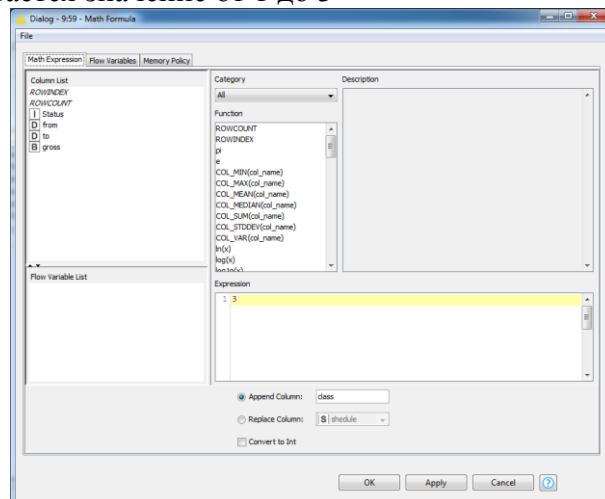


Это узел позволяет добавлять математические выражения после вычисления, которых добавляются в новый столбец или заменяют значения в старом. В данном случае создается столбец *Class* и каждому запросу присваивается значение от 1 до 3

Math Formula



Node 59



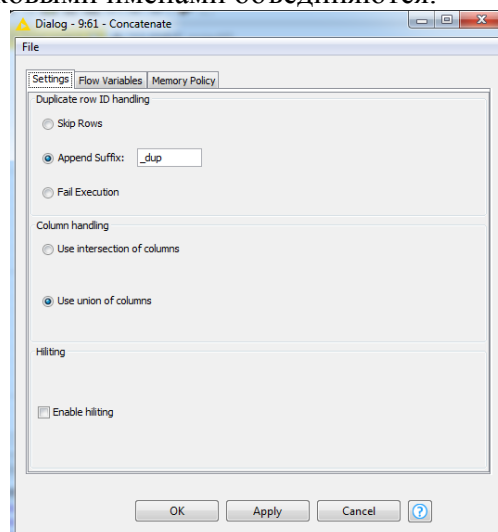
Блок – Трансформации полученной информации о вакансиях:

Узел, который объединяет три таблицы в одну. Столбцы с одинаковыми именами объединяются.

Concatenate



Node 61

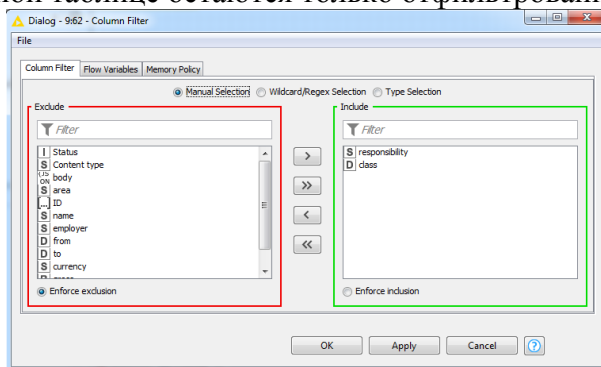


Column Filter



Node 62

Узел позволяющий фильтровать столбцы из входной таблицы. В выходной таблице остаются только отфильтрованные столбцы.

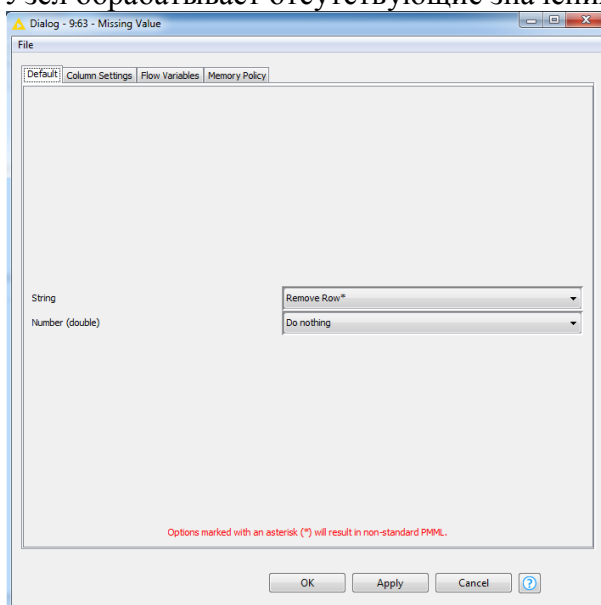


Узел обрабатывает отсутствующие значения во входной таблице.

Missing Value



Node 63



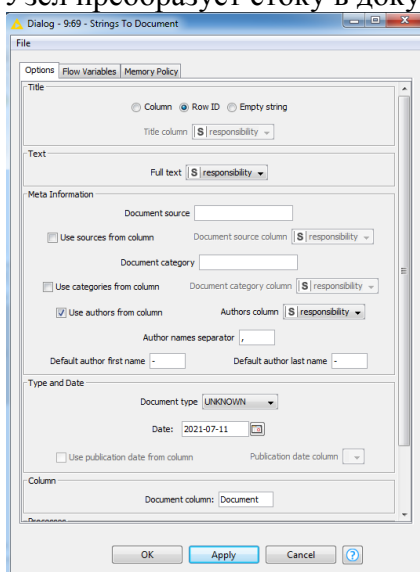
Блок – Предварительной обработки описания и визуализации вакансий:

Узел преобразует сток в документ.

Strings To Document

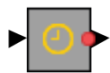


Node 69

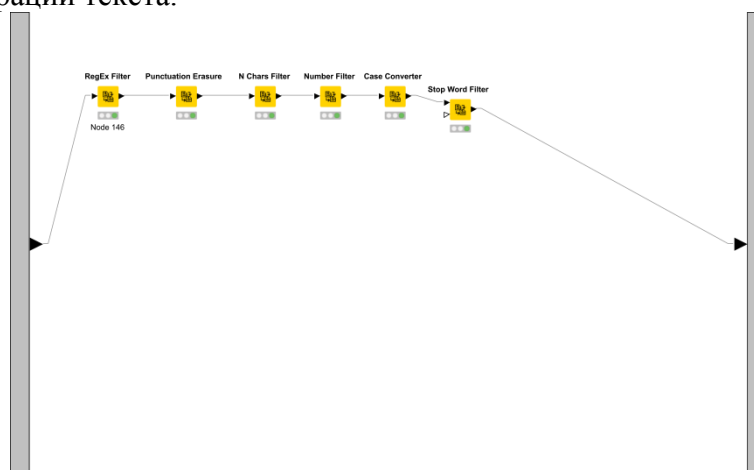


Объединяющий несколько узлов узел. В данном узле собраны узлы фильтрации текста.

Text Filtering



Node 159

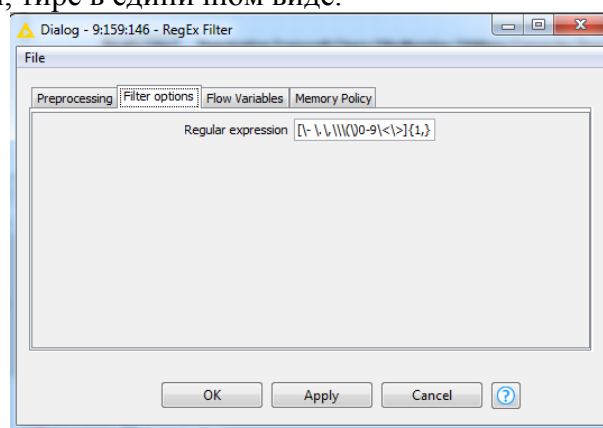


Узел, использующий регулярные выражения. Фильтрует все термины, содержащиеся во входном документе, которые соответствуют указанному регулярному выражению. В данном узле фильтруем знаки препинания, цифры, тире в единичном виде.

RegEx Filter

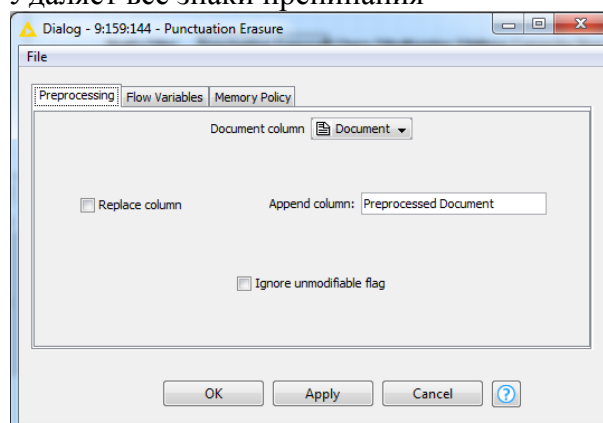


Node 146



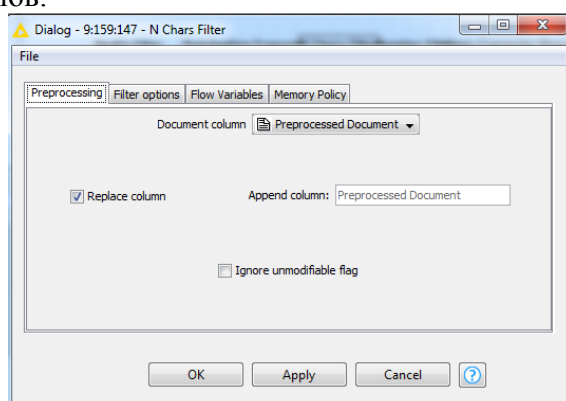
Удаляет все знаки препинания

Punctuation Erasure

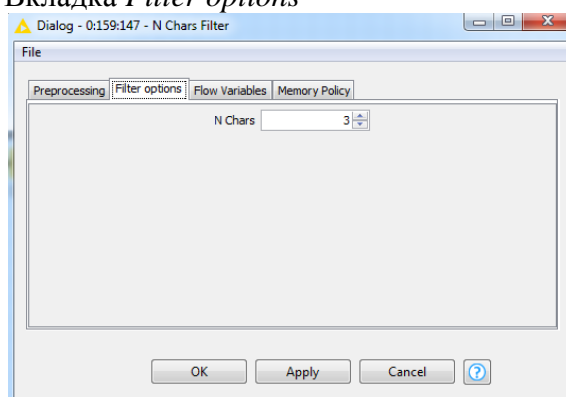


Фильтрует все термины, содержащиеся во входных документах, с количеством символов меньше указанного N. В данном случае меньше 3 СИМВОЛОВ.

N Chars Filter

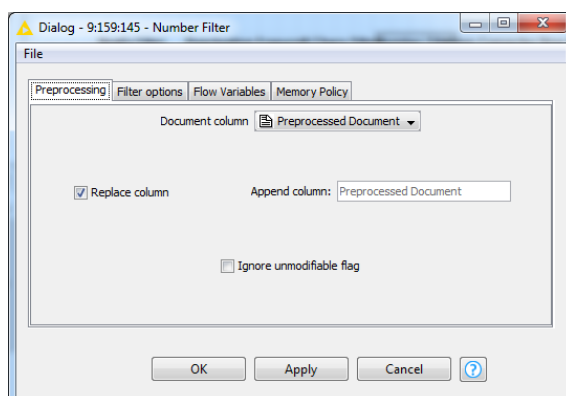


Вкладка *Filter options*

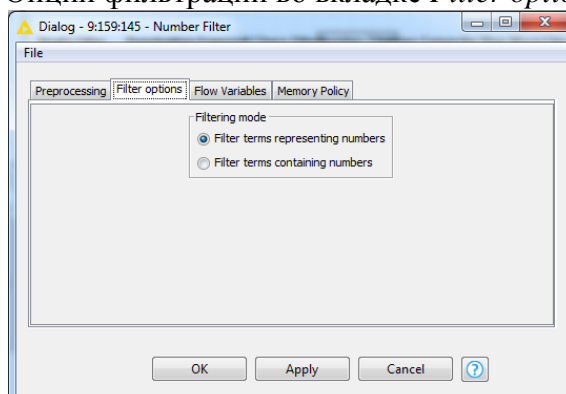


Фильтрует все элементы, содержащие цифры, знаки разделения, «+», «-».

Number Filter

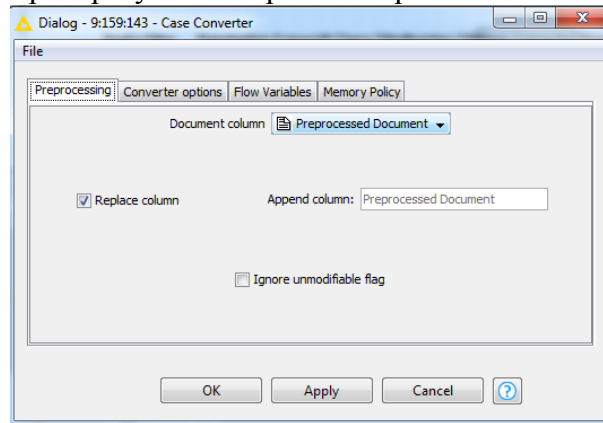


Опции фильтрации во вкладке *Filter options*

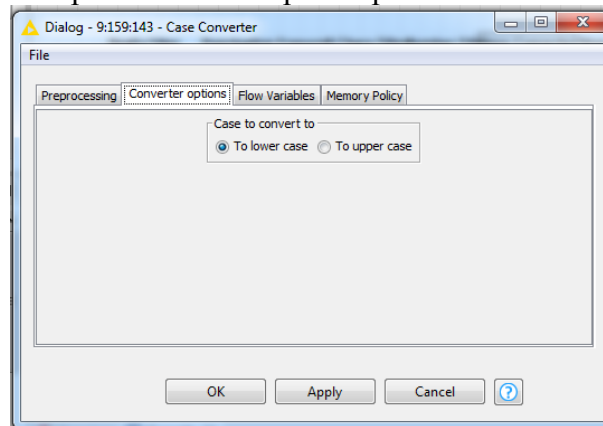


Преобразует все термины строчные.

Case Converter

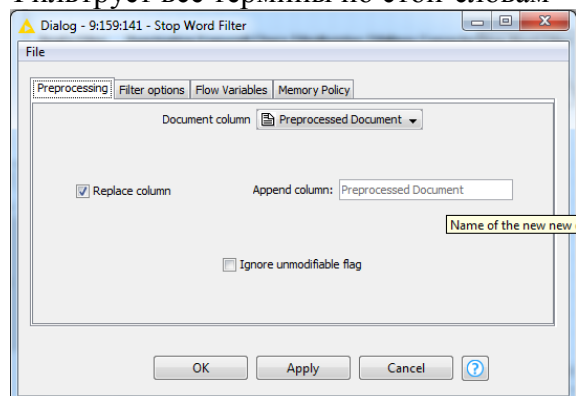


Настройка нижнего регистра

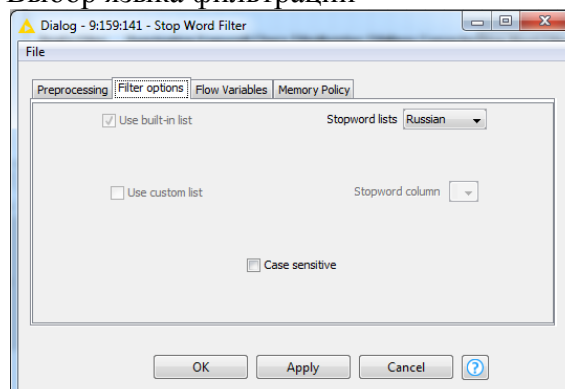


Фильтрует все термины по стоп-словам

Stop Word Filter




Выбор языка фильтрации



Объединяющий узел обработки текста на основе частотного анализа


Term Filtering



Based on document frequency



Extract Table Dimension



Extract number of documents

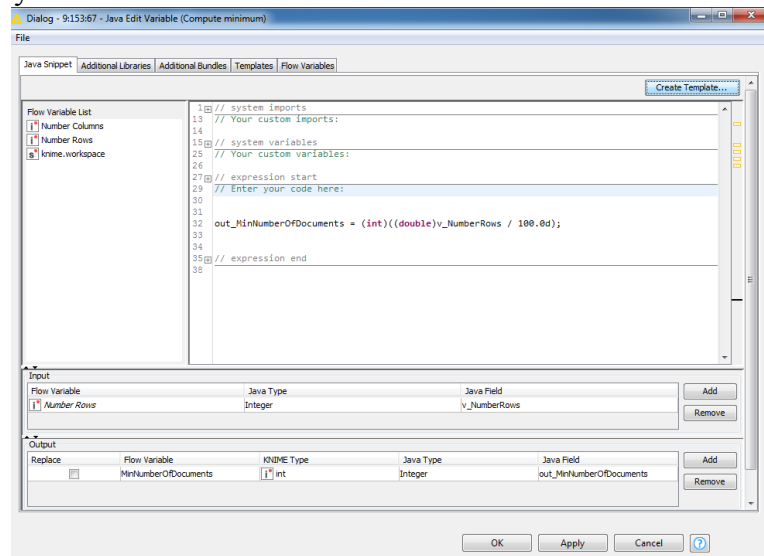
Узел, считающий количество строк и столбцов в исходной таблице

Узел обработки кода *Java*. В данном случае вычисляет минимальную частоту документов

Java Edit Variable

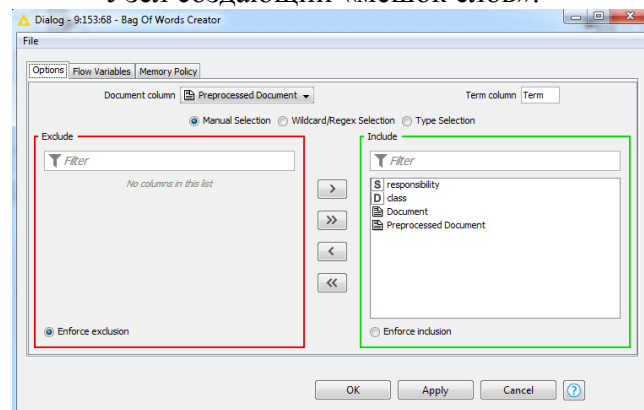


Compute minimum document frequency



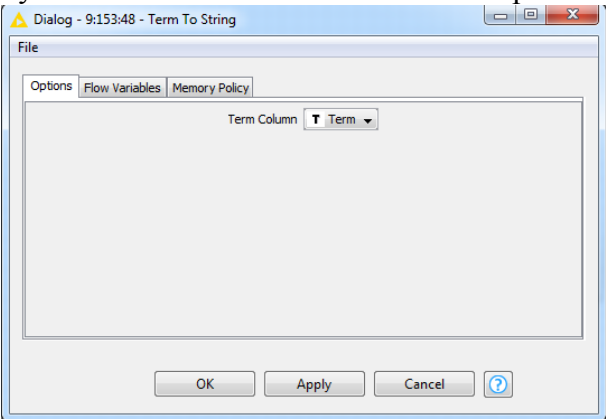
Узел создающий «мешок слов».

Bag Of Words Creator

Переводит полученные слова из «мешка слов» в строковые

Term To String

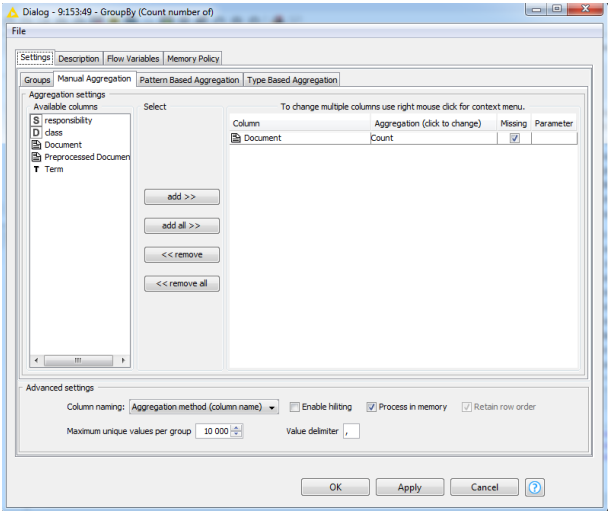
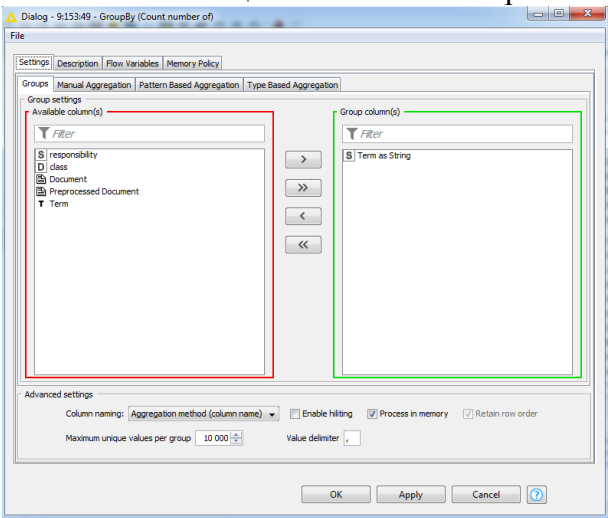


Считает количество каждого термина в таблице и формирует дополнительный столбец с количеством встречающихся терминов

GroupBy




Count number of documents each term occurs in

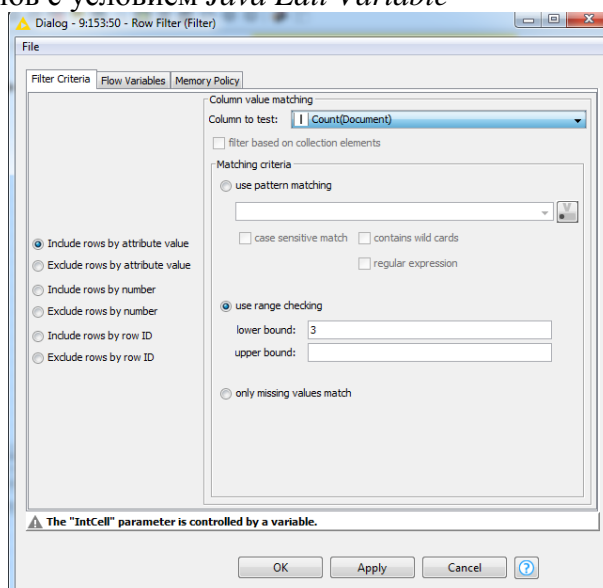


Узел фильтрует по минимальному количеству встречающихся терминов с условием *Java Edit Variable*

Row Filter




Filter terms / features

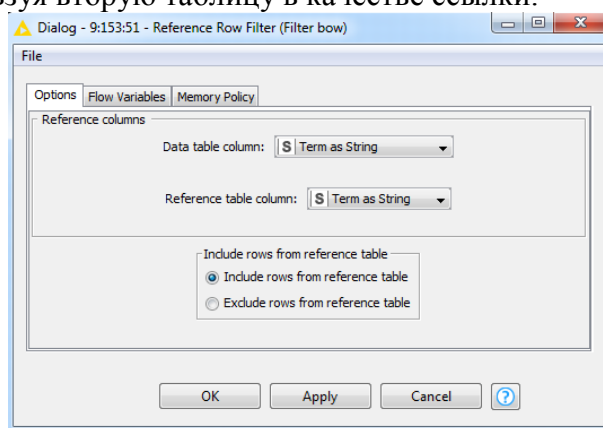


Этот узел позволяет фильтровать строки из первой таблицы, используя вторую таблицу в качестве ссылки.

Reference Row Filter

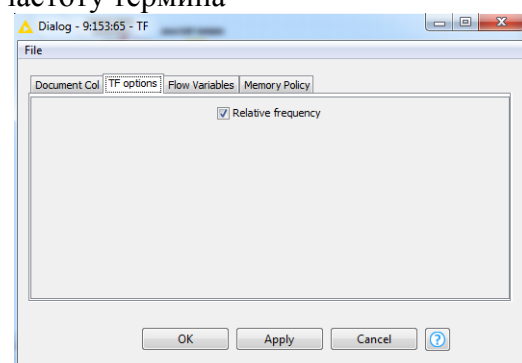
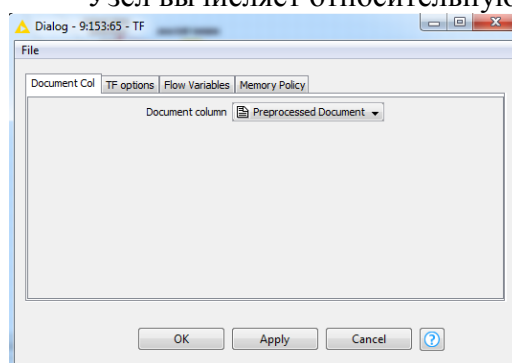


Filter bow



Узел вычисляет относительную частоту термина

TF

Представляет облако тегов. Облако тегов для вакансий представлено ниже

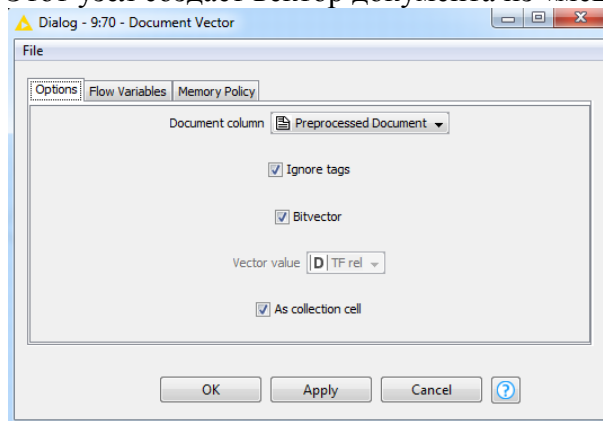
Tag Cloud



Node 158



Этот узел создает вектор документа из «мешка слов»

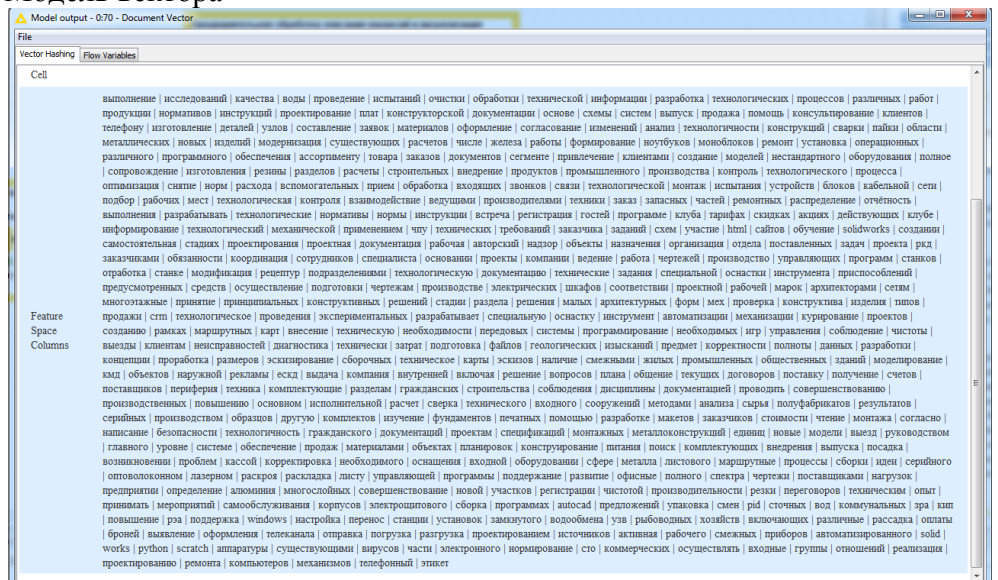


Document Vector



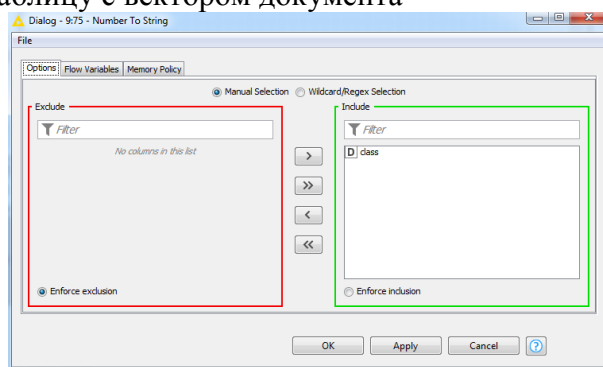
Node 70

Модель вектора



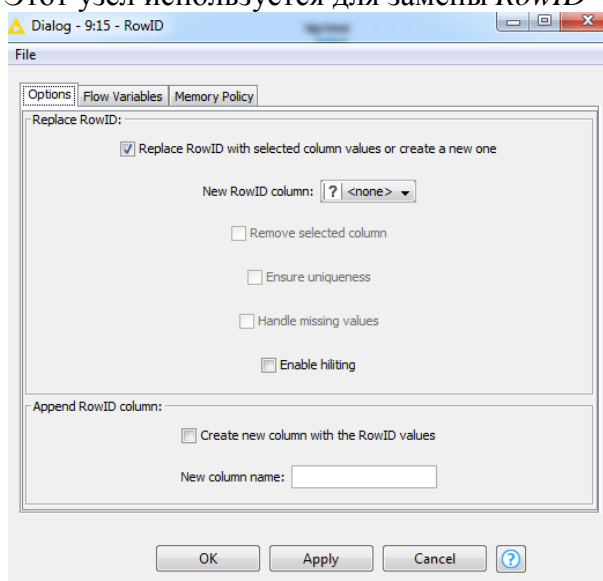
Узел, преобразующий числа в строковые для того чтобы объединить в одну таблицу с вектором документа

Number To String



Этот узел используется для замены *RowID*

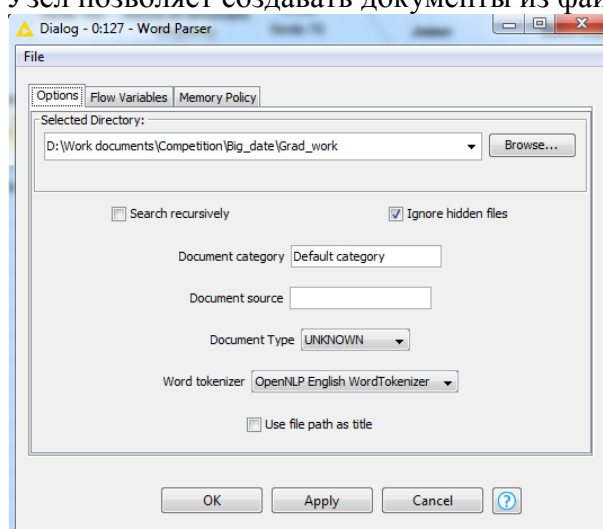
RowID



Блок – Предварительной обработки рабочей программы и ее визуализации.

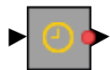
Узел позволяет создавать документы из файлов word

Word Parser

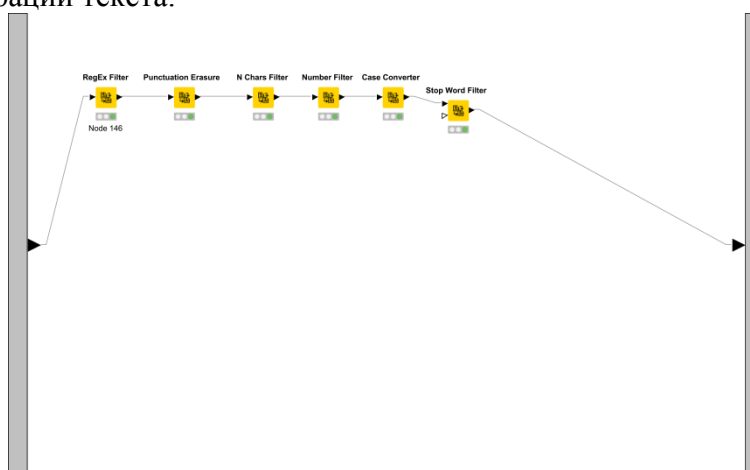


Объединяющий несколько узлов узел. В данном узле собраны узлы фильтрации текста.

Text Filtering

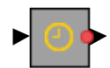


Node 159

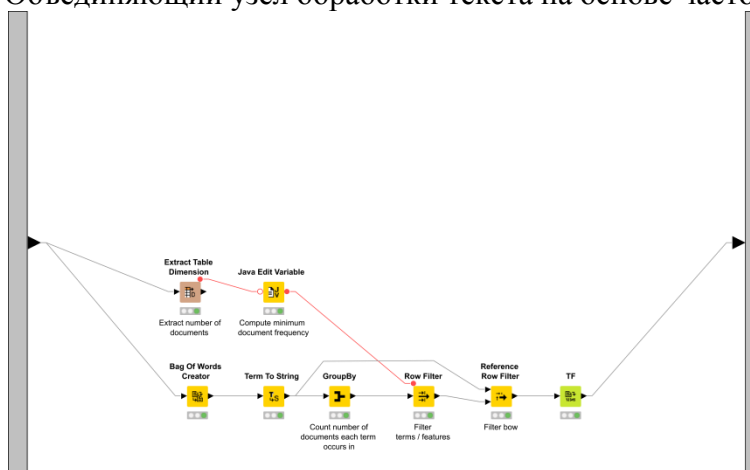


Объединяющий узел обработки текста на основе частотного анализа

Term Filtering



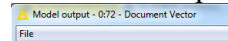
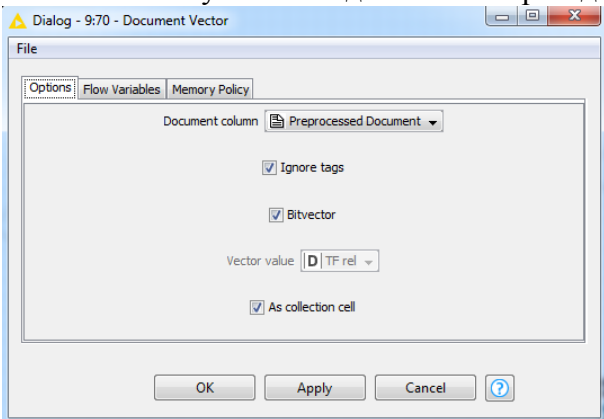
Based on document frequency



Представляет облако тегов. Облако тегов для рабочей программы представлено ниже



Этот узел создает вектор документа из «мешка слов»



Model output - 072 - Document Vector

File

Vector HashingFlow Variables

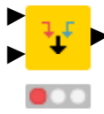
Parameter	Value
Ignore Tags	true
BitVector	true
Vector Value	null
As Collection	true
Cell	

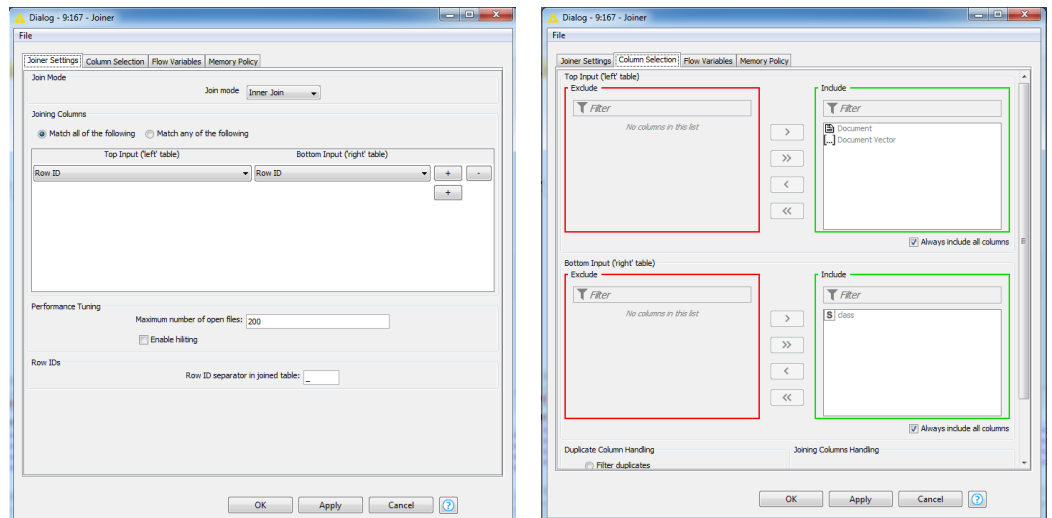
Feature Space Columns

аннотация | работей | программы | дисциплины | системы | автоматизированного | проектирования | разработке | технологического | оборудования | программе | бакалаврита | технологические | машины | оборудование | профит | аппараты | промышленной | экологии | квалификация | выпускника | бакалавр | выпускающая | кафедра | химических | заводов | цели | основания | цели | является | теоретическая | профессиональная | подготовка | студентов | области | графического | изображения | информации | получение | студентами | навыков | пользования | современных | компьютерных | технологий | подготовке | технической | технологической | документации | формирования | самостоятельной | работы | выработка | знаний | необходимых | студентам | выполнения | чтения | технических | чертежей | эскизов | деталей | составления | конструкторской | производства | содержание | принципы | задачи | системный | подход | создание | ассоциативных | моделирование | результате | обучающийся | должен | знать | основные | составляющие | аппаратной | программной | части | графических | станций | закона | компьютерного | построения | чертежа | основополагающие | требования | стандарты | единой | методы | обратных | пространственных | объектов | изображения | чертежа | прямых | плоскостей | кривых | линий | поверхностей | способы | преобразования | решения | чертежах | основах | метрических | позиционных | задачи | построение | чтение | сборочных | общего | вида | различного | уровня | сложности | назначения | рисунков | стандартных | разъемных | неразъемных | соединений | единиц | приение | конструкции | показанной | процессах | изготовления | возможностей | международных | стандартах | осуществлять | автоматизированное | проектирование | владеть | основными | методами | приемами | расчета | помощи | программ | оформлять | конструкторскую | сопровождающую | документацию | соответствия | ескд | возможностям | информационных | описания | принципами | создания | функционирования | возможностью | использования | современными | обработки | представления | навыками | современным | компьютерным | офисным | оборудованием


Блок – Прогнозирования и оценки вакансий с рабочей программой

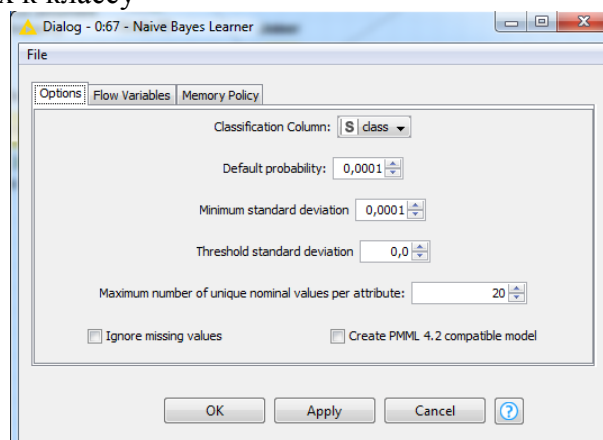
Узел, объединяющий таблицы

Joiner

Node 167



Узел создает байесовскую модель из заданных обучающих данных. Он вычисляет количество строк на значение атрибута для каждого класса для номинальных атрибутов и распределение Гаусса для числовых атрибутов. Созданная модель может быть использована в наивном байесовском предикторе для прогнозирования принадлежности неклассифицированных данных к классу

Naive Bayes Learner

Node 67

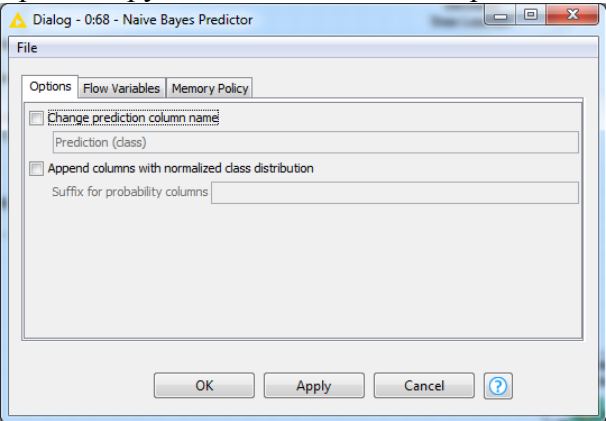


Прогнозирует класс для каждой строки на основе изученной модели

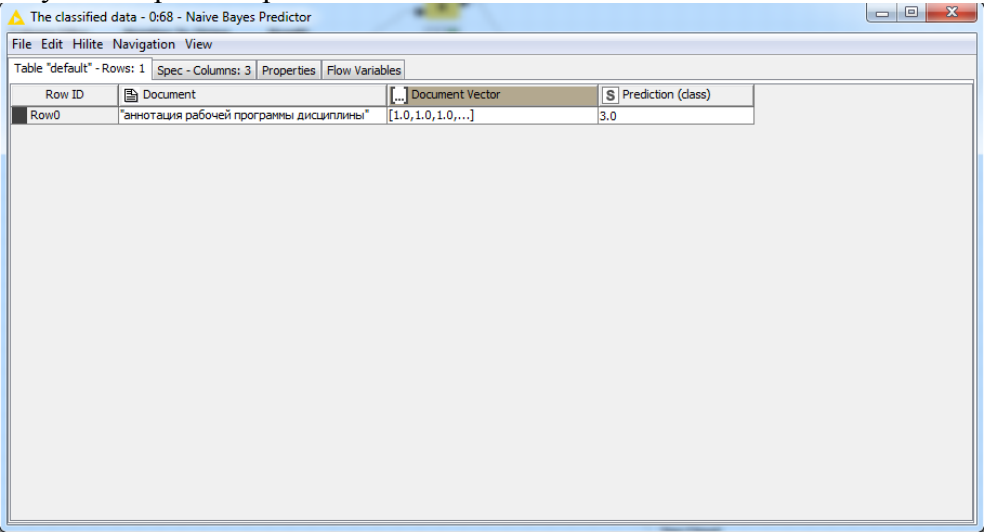
Naive Bayes Predictor



Node 68



Результат прогнозирования



Б1.В.07 Системы автоматизированного проектирования в разработке технологического оборудования

по программе бакалаврита: 15.03.02 Технологические машины и оборудование

профиль: Машины и аппараты промышленной экологии

квалификация выпускника: БАКАЛАВР

выпускающая кафедра: Оборудование химических заводов

Кафедра-разработчик рабочей программы: Оборудование химических заводов

1. Цели освоения дисциплины

Целями освоения дисциплины Системы автоматизированного проектирования в разработке технологического оборудования является теоретическая и профессиональная подготовка студентов в области графического изображения информации, получение студентами навыков пользования современными компьютерными технологиями при подготовке технической и технологической документации, формирования у студентов навыков самостоятельной работы, выработка знаний и навыков, необходимых студентам для выполнения и чтения технических чертежей, выполнения эскизов деталей, составления конструкторской и технической документации производства

2. Содержание дисциплины

Глава 1. Принципы и задачи проектирования. системный подход

Глава 2. Создание ассоциативных чертежей

Глава 3. 2D моделирование

Глава 4. 3d моделирование

3. В результате освоения дисциплины обучающийся должен:

Знать:

- Основные составляющие аппаратной и программной части современных графических станций
- Основные законы компьютерного построения чертежа;
- Основопологающие требования к конструкторской документации;
- Стандарты Единой системы конструкторской документации;
- Методы построения обратимых чертежей пространственных объектов;
- Изображения на чертеже прямых, плоскостей, кривых линий и поверхностей; способы преобразования чертежа;
- Способы решения на чертежах основных метрических и позиционных задач;
- Построение и чтение сборочных чертежей общего вида различного уровня сложности и назначения.
- Методы построения эскизов, чертежей и технических рисунков стандартных деталей,
- Разъемных и неразъемных соединений деталей и сборочных единиц;
- О принципе работы конструкции, показанной на чертеже;
- Об основных технических процессах изготовления деталей;
- О возможностях компьютерного выполнения чертежей;
- О международных стандартах.

Уметь:

- Осуществлять автоматизированное проектирование технологического оборудования;

- Владеть основными методами и приёмами расчета технологического оборудования при помощи программ автоматизированного проектирования.
- Оформлять конструкторскую и сопровождающую документацию в соответствии с ЕСКД.

Владеть: основными возможностями информационных технологий; методами описания информационных технологий; принципами создания и функционирования; возможностью использования информационных технологий; современными методами обработки и представления информации; навыками работы с современным компьютерным и офисным оборудованием

Обсуждение и анализ полученных результатов

Подводя итоги экспериментального исследования, необходимо отметить, что применение Байесовского классификатора позволило, оценить соответствие представленной рабочей программы с вакансиями с www.hh.ru.

Конечно, как и любой метод оценки, он не лишен недостатков и базируется на предположении, что одни слова чаще встречаются в тексте, на котором обучается модель, а другие - нет, и неэффективен, если данное предположение неверно. Т.е. если мы не подготовим текст перед обучением, наша модель будет работать не корректно.

Например, если в вакансии много лишнего текста не несущего смысл. Вычленить такой текст машинными методами будет в принципе не возможно. Поэтому для улучшения работы моделей обучения и поиска вакансий желательно, чтобы в вакансиях использовались ключевые слова.

Заключение

В работе рассмотрена оптимизация рабочей программы под рынок труда на основе применение алгоритмических и инструментальных средств машинного обучения. В результате анализа вакансий показано, что рабочая программа соответствует вакансии «инженер-конструктор», но может быть применима для вакансии «инженер-технолог». Конечно, у каждой вакансии есть своя специфика, а получаемые знания по учебному плану будут более общие.

Список использованных источников

1. Boris Mirkin. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. 2010.
2. James, Witten, Hastie, Tibshirani. An Introduction to Statistical Learning. 2013.
3. Hastie T., Tibshirani R, Friedman J. The Elements of Statistical Learning (2nd edition). Springer, 2009.
4. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006.
5. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. MIT Press, 2012.
6. Murphy K. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
7. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014.
8. Willi Richert, Luis Pedro Coelho. Building Machine Learning Systems with Python. Packt Publishing, 2013.
9. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval. Cambridge University Press, 2009
10. Foster Provost, Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc., 2013
11. Knime Hub [электронный ресурс]. URL: <https://hub.knime.com/> (дата обращения: 16.07.2021).
12. Knime Documentation [электронный ресурс]. URL: <https://docs.knime.com/> (дата обращения: 16.07.2021).