

Інтелектуальний аналіз природної мови за допомогою аналітичної платформи KNIME

Виконала: студентка групи ДА-62

Маркіна Софія Павлівна

Керівник:

Шаптала Роман Віталійович

Актуальність

Актуальність теми даної роботи полягає в тому, що на сьогоднішній день наявна досить велика кількість програмних засобів, що дозволяють реалізувати процес інтелектуального аналізу тексту, через це стає необхідність в дослідженні та аналізі існуючих середовищ, з метою визначення можливостей, якості та функціоналу вищезгаданих.

Мета роботи

Метою даної роботи є:

- вивчення та аналіз існуючих алгоритмів та методів для вирішення задачі інтелектуального аналізу даних;
- дослідження та тестування аналітичної платформи KNIME шляхом реалізації процесів інтелектуального аналізу даних на базі цього середовища;
- оцінка функціоналу середовища за якістю реалізації поставлених задач.

Завдання

Завданням даної роботи стало:

- дослідження функціональних можливостей аналітичної платформи KNIME для обробки природної мови;
- побудова робочих процесів для задач: витягу теми неструктурованого тексту, аналізу настроїв користувачів та побудови дерева рішень для визначення словникових конструкцій.

Використані документи

3617,"http://www.imdb.com/title/tt0210075/usercomments", "Girlfight follows a project dwelling New York high school girl from a sense of futility into the world of amateur boxing where she finds self esteem, purpose, and mu
3671,"http://www.imdb.com/title/tt0337640/usercomments", "Hollywood North is an euphemism from the movie industry as they went to Canada to make movies because of tax breaks and cheaper costs in a civilized city like T
3157,"http://www.imdb.com/title/tt0118225/usercomments", "Almost everyone is funny, where with many shows there are only a few funny characters. I will be sad to see this show end next year, but it will be going off the air as one of the best shows ever. That '70s
660,"http://www.imdb.com/title/tt0716825/usercomments", "9/10- 30 minutes of pure holiday terror. Okay, so it's not that scary. But it sure is fun. The Crypt Keeper (John Kassir) takes a tale of holiday FEAR, giving us all Christm
265,"http://www.imdb.com/title/tt0182225/usercomments", "A series of random, seemingly insignificant thefts at her sister's boarding house has Miss Lemon quite agitated. A ring, light bulbs, a rucksack, a lighter, a stethoscopi
4027,"http://www.imdb.com/title/tt0347779/usercomments", "A very good adaptation of the novel by Amrita Pritam. Urmila and Manoj Bajpai have given their best. There is a natural flair in the movie and I felt it right through. It lo
5820,"http://www.imdb.com/title/tt0298131/usercomments", "Although the beginning of the movie in New York takes too long, the movie is a must see for people who like this genre. When Hannah goes to Berlin to visit the old
10668,"http://www.imdb.com/title/tt0088915/usercomments", "As many reviewers here have noted, the film version differs quite a bit from the stage version of the story. I have never seen the stage version of the story, and the
1473,"http://www.imdb.com/title/tt0828154/usercomments", "Bear in mind, any film (let alone documentary) which asserts any kind of truth, will generate an adverse and proportional amount of cynicism, from those to whom
8337,"http://www.imdb.com/title/tt0110099/usercomments", "Being a big fan of the romantic comedy genre, and therefore having seen a large number of these films, it is rare that one strikes me as totally unique. For that matte
11217,"http://www.imdb.com/title/tt0076683/usercomments", "Bored with the normal, run-of-the-mill staple films to watch this Halloween that I've seen over and over again, I took a chance on The Sentinel, hoping it could get i
12389,"http://www.imdb.com/title/tt0090799/usercomments", "Care Bears Movie 2: A New Generation isn't at all a bad movie. In fact, I like it very much. Yes I
1212,"http://www.imdb.com/title/tt0092615/usercomments", "Deodato brings us some mildly shocking moments and a movie that doesn't take itself too s
5272,"http://www.imdb.com/title/tt0820111/usercomments", "Deranged and graphically gory Japanese film about little beings taking people over and turnin
9536,"http://www.imdb.com/title/tt0365960/usercomments", "Everyone knows about this "Zero Day" event. What I think this movie did that Elephant did
11782,"http://www.imdb.com/title/tt0071507/usercomments", "Expecting to see another Nunsploitation movie with a mean Mother Superior abusing and to
11469,"http://www.imdb.com/title/tt0406713/usercomments", "First things first! This isn't an action movie although there is a lot of action in it! I think you
9228,"http://www.imdb.com/title/tt0001032/usercomments", "Had this movie been made a few years later, I would have given it a lower score. However, for
9945,"http://www.imdb.com/title/tt0956331/usercomments", "I admit I had no idea what to expect before viewing this highly stylized piece. It could have be
5992,"http://www.imdb.com/title/tt0468458/usercomments", "I chose to watch this film at Tribeca based on Judd Hirsch and Scott Cohen and found it to b
4678,"http://www.imdb.com/title/tt0446345/usercomments", "I desperately need this on a tape, not a DVD, and soon! I have one nephew who is in the infan
3422,"http://www.imdb.com/title/tt0079672/usercomments", "I have spent the last week watching John Cassavetes films - starting with 'a woman under the
6823,"http://www.imdb.com/title/tt0117179/usercomments", "I just got this video used and I was watching it last night. The acting started out extremely ba
3976,"http://www.imdb.com/title/tt0795102/usercomments", "I just saw The Drugs Years on VH1 and I love it. I think it reflects the drug history very well an
10202,"http://www.imdb.com/title/tt0354690/usercomments", "I personally liked this movie and am alarmed at the rating's some people have given it. It is a
2844,"http://www.imdb.com/title/tt0108288/usercomments", "I saw the last five or ten minutes of this film back in 1998 or 1999 one night when I was cha
5169,"http://www.imdb.com/title/tt0101625/usercomments", "I saw this ages ago when I was younger and could never remember the title, until one day I v
10768,"http://www.imdb.com/title/tt0079672/usercomments", "I spent 5 hours drenched in this film. Nothing I have ever seen comes close to the delicious t
6827,"http://www.imdb.com/title/tt0324046/usercomments", "I was pretty young when this came out in the US, but I recorded it from TV and watched it ov
8693,"http://www.imdb.com/title/tt0117093/usercomments", "I've just finished listening to the director's commentary for this film, and I think the one big

Текст п'єси «Лісова пісня» у форматі EPUB

Лісова пісня Леся Українка

Драма-феєрія в 3-х діях

СПИС ДІЯЧІВ "ЛІСОВОЇ ПІСНІ"

ПРОЛОГ

"Той, що греблі рве". Русалка.
Потерчата (двос). Водяник.

ДІЯ I

Дядько Лев. Перелесник.
Лукаш. Пропасниця (без мови).
Русалка. Потерчата.
Лісовик. Куць.
Мавка.

ДІЯ II

Мати Лукашева. Килина.
Лукаш. Русалка.
Дядько Лев. "Той, що в скалі
Мавка сидить".
Русалка Польова. Перелесник.

ДІЯ III

Мавка. Хлопчик.
Лісовик. Лукаш.
Куць. Діти Килинині
Злидні. (без мови).
Мати Лукашева. Доля.
Килина. Перелесник.

ПРОЛОГ

Старезний, густий, предковичний ліс на Волині. Посеред лісу простора галява з плакучою березою і з великим прастарим дубом. Галява скраю переходить в куля та очерети, а в одному місці в яро-зелену драговину — то береги лісового озера, що утворилося з лісового струмка. Струмок той вбігає з гущавини лісу, впадає в озеро, потім, по другім боці озера, знов витікає і губиться в хащах. Саме озеро — тиховоде,

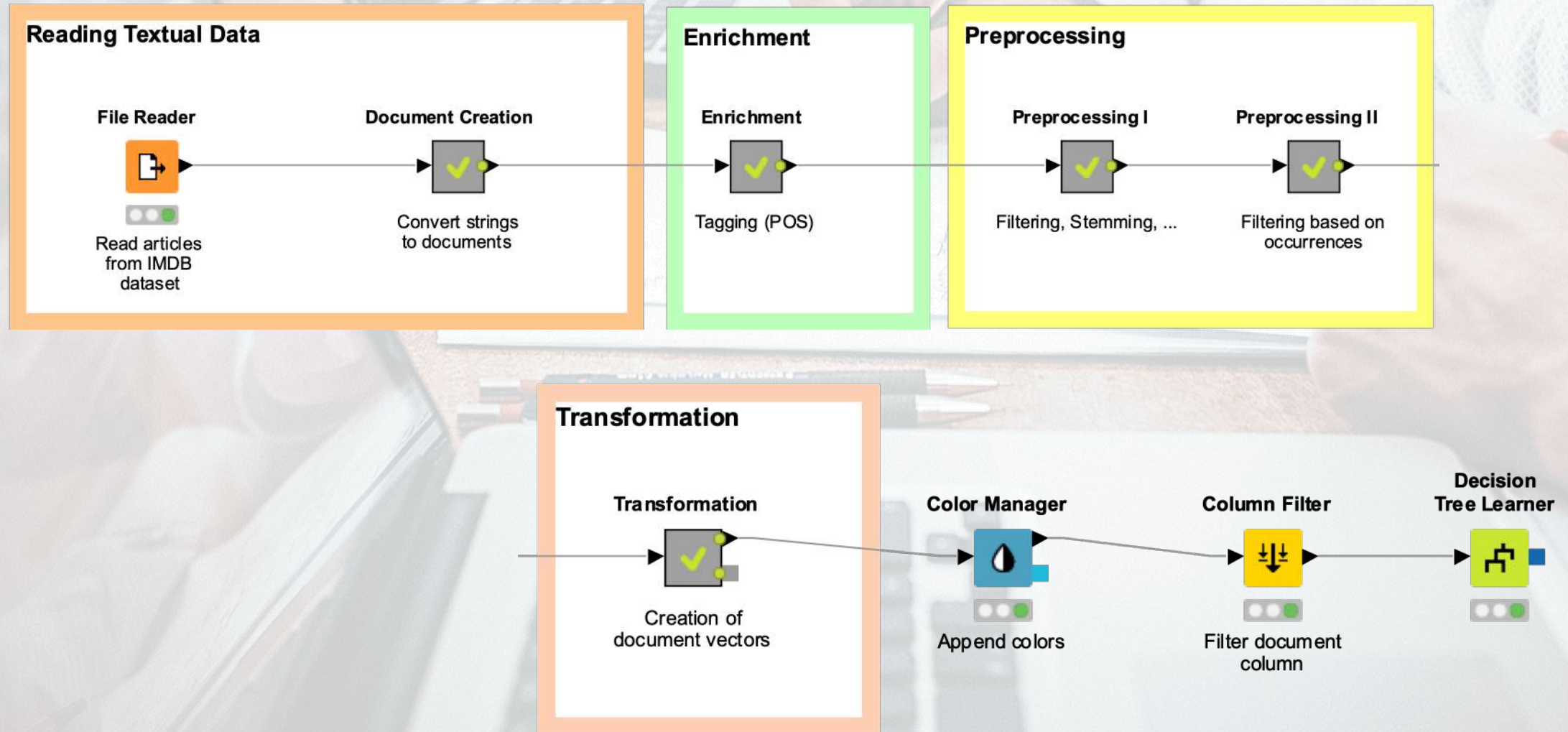
База даних рецензій сайту IMDB у вигляді csv - таблиці

Робочий процес у KNIME

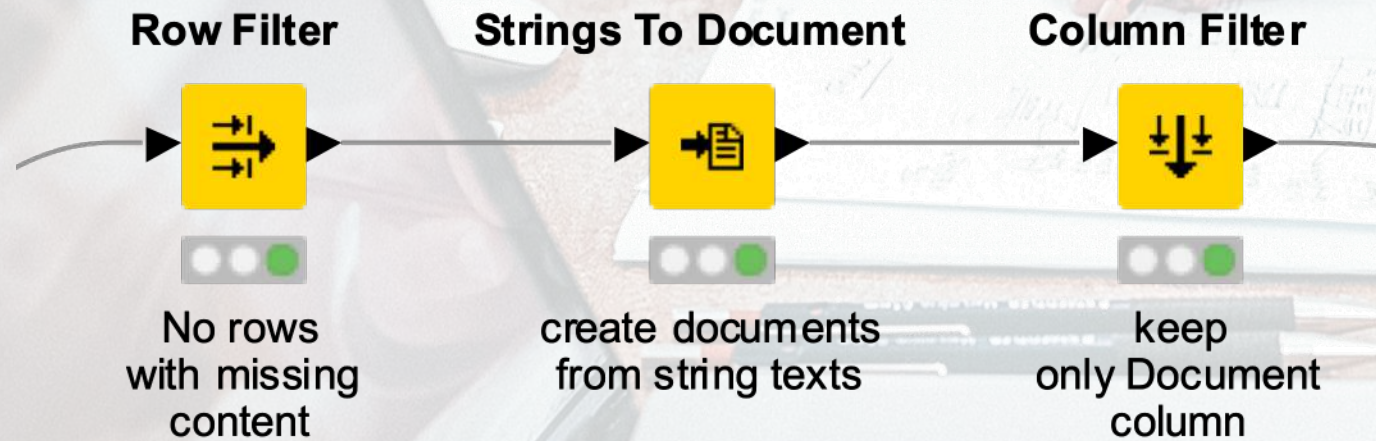
The image shows the KNIME Analytics Platform interface with several key components highlighted by yellow callouts:

- KNIME Explorer:** Located in the top-left pane, it displays a hierarchical tree view of the workspace, including folders like 'My-KNIME-Hub', 'EXAMPLES', 'LOCAL', and 'My_First_Project'.
- Workflow Coach:** Located in the middle-left pane, it provides 'Recommended Nodes' such as Joiner, Column Filter, Row Filter, Partitioning, GroupBy, Missing Value, and Statistics.
- Node Repository:** Located in the bottom-left pane, it lists various node categories like IO, Manipulation, Views, Analytics, DB, and Other Data Types.
- Workflow Editor:** The central workspace where a workflow named 'My First Workflow' is being built. The workflow consists of a 'File Reader' node (labeled 'Read sales data'), followed by a 'Column Filter' node (labeled 'Include country, date and amount'), and a 'Row Filter' node (labeled 'Exclude rows with country "unknown"'). The output of the Row Filter is split into two visualization nodes: a 'Stacked Area Chart' (labeled 'Sales over time') and a 'Pie/Donut Chart' (labeled 'Sales per country').
- Description:** A pane on the right showing the configuration and help text for the 'File Reader' node.
- KNIME Hub Search:** A search bar at the bottom right for finding workflows, nodes, and other resources.
- Outline:** A pane at the bottom left showing a hierarchical overview of the workflow structure.
- Console:** A pane at the bottom right displaying the KNIME Console output, which includes a welcome message and copyright information for KNIME AG.

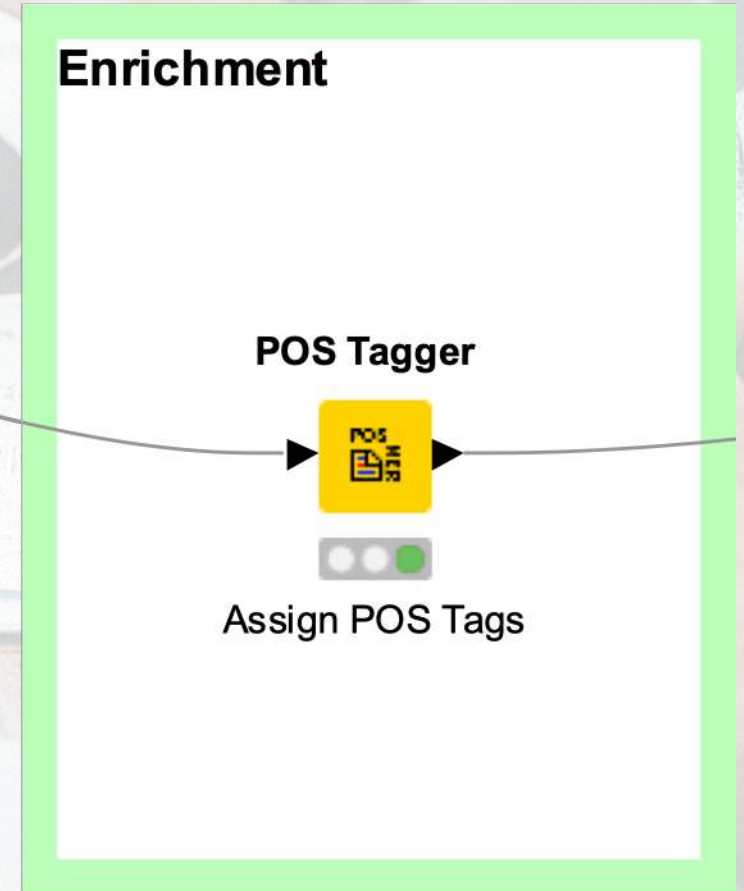
Decision tree: загальний робочий процес



Decision tree: зчитування даних та збагачення

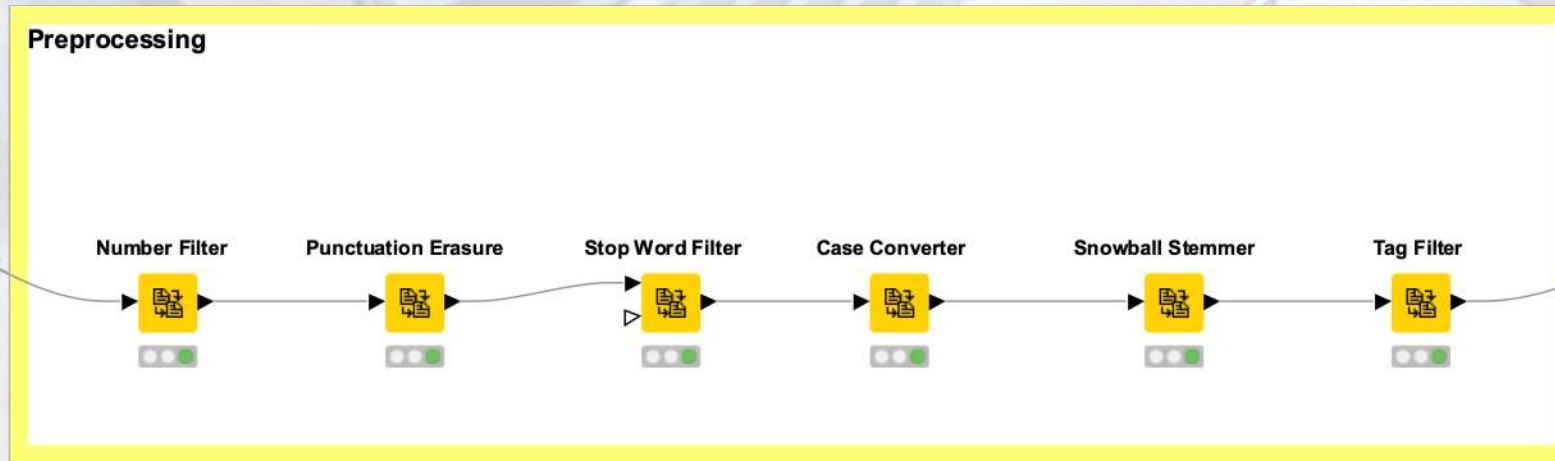


Етап зчитування даних

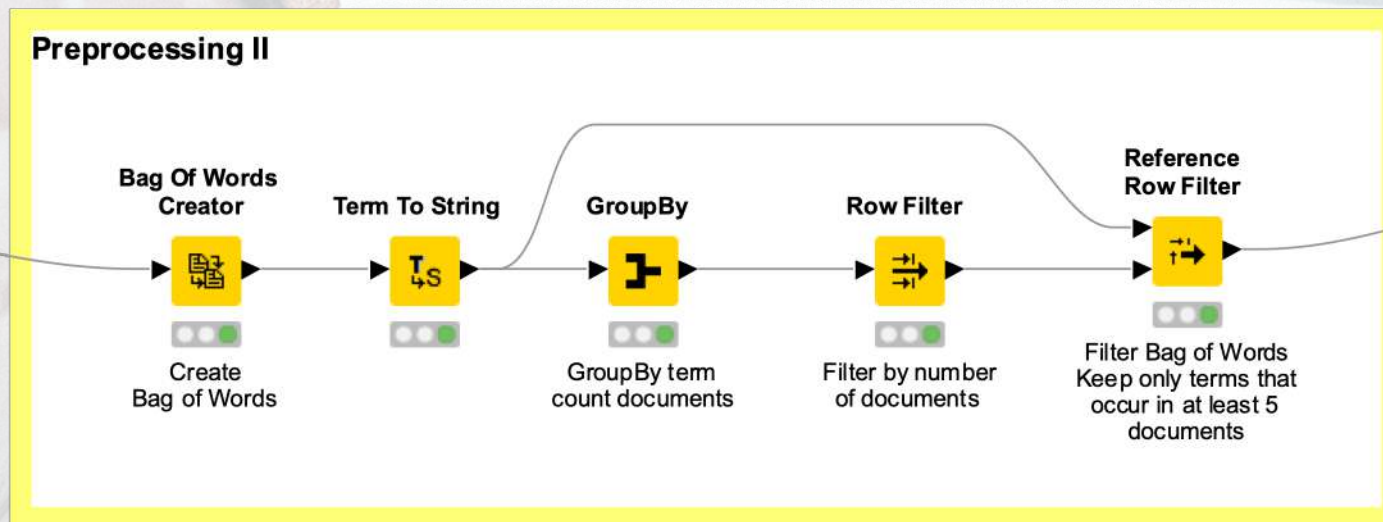


Етап збагачення

Decision tree: попередня обробка

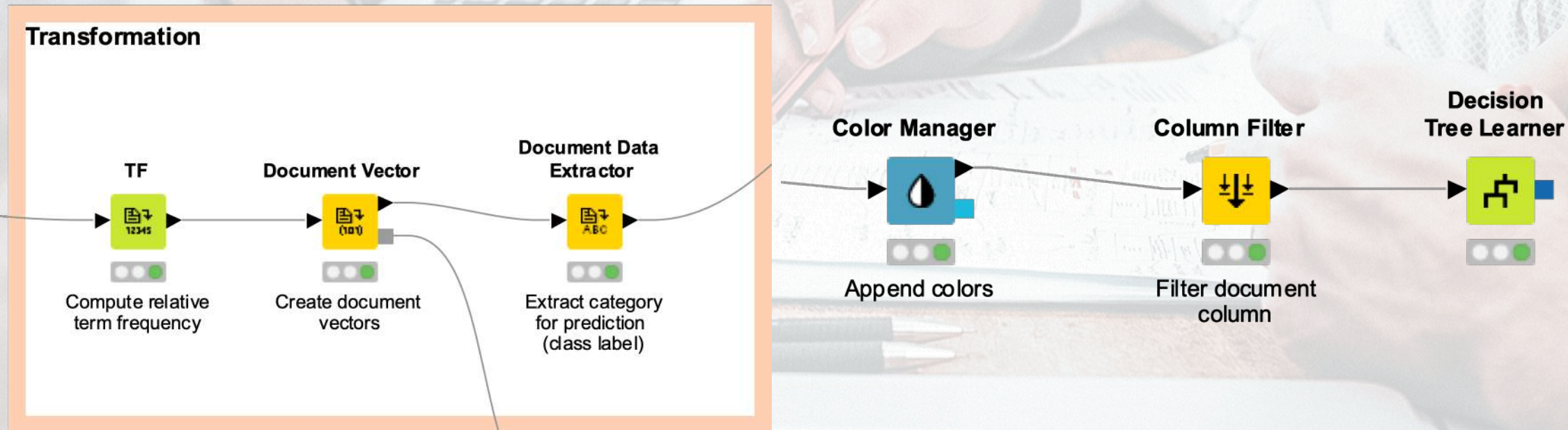


Перший етап попередньої обробки



Другий етап попередньої обробки

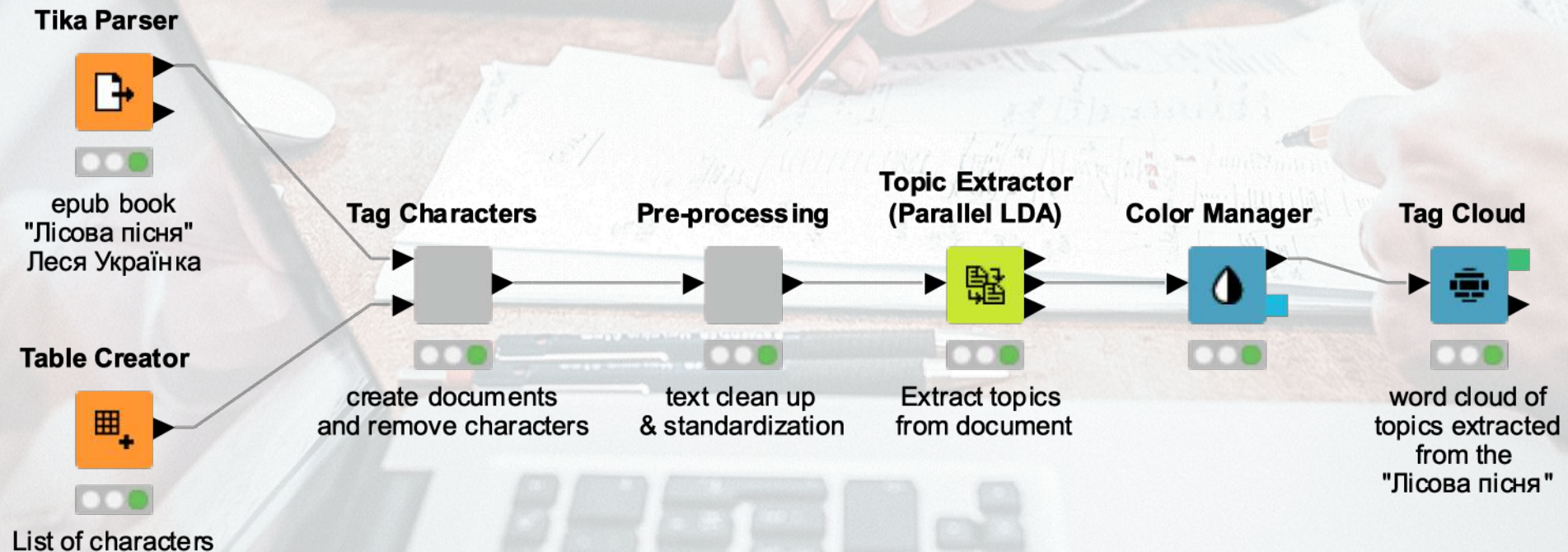
Decision tree: частоти, векторизація та виведення результатів



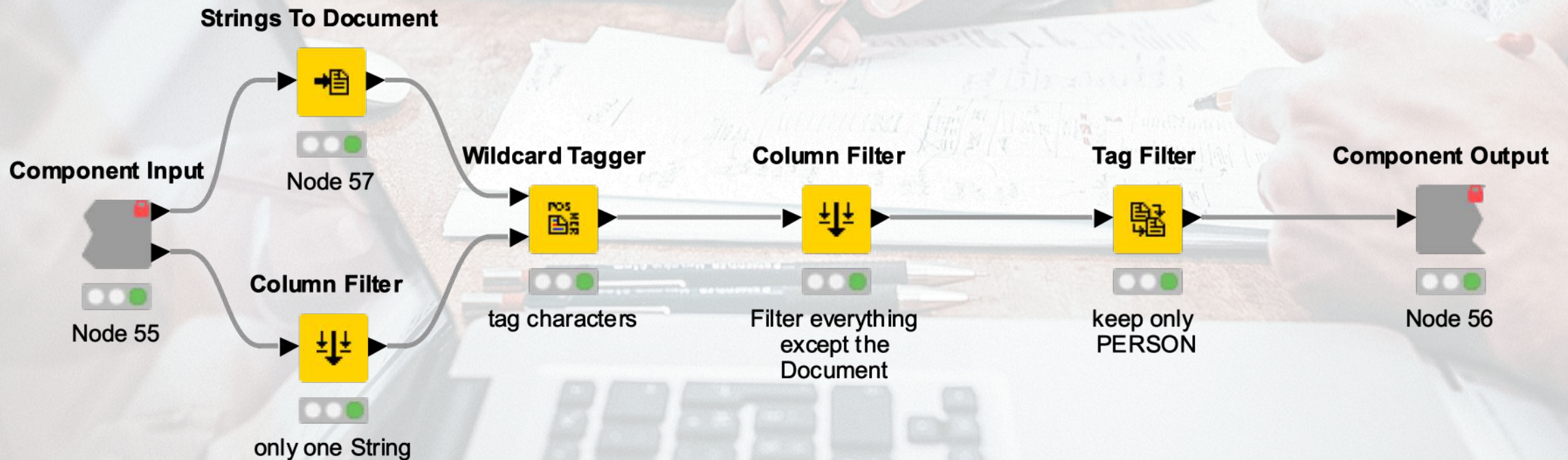
Етап трансформації

Розподілення кольорів, фільтрація за відповідними рядками та побудова дерева рішень

Topic Detection: загальний робочий процес



Topic Detection: вилучення з тексту дійових осіб і стоп-слів



Topic Detection: Topic Extractor (Parallel LDA)

Dialog - 2:756 - Topic Extractor (Parallel LDA) (Extract topics)

Options Flow Variables Memory Policy

Document column: Seed:

No of topics: No of words per topic:

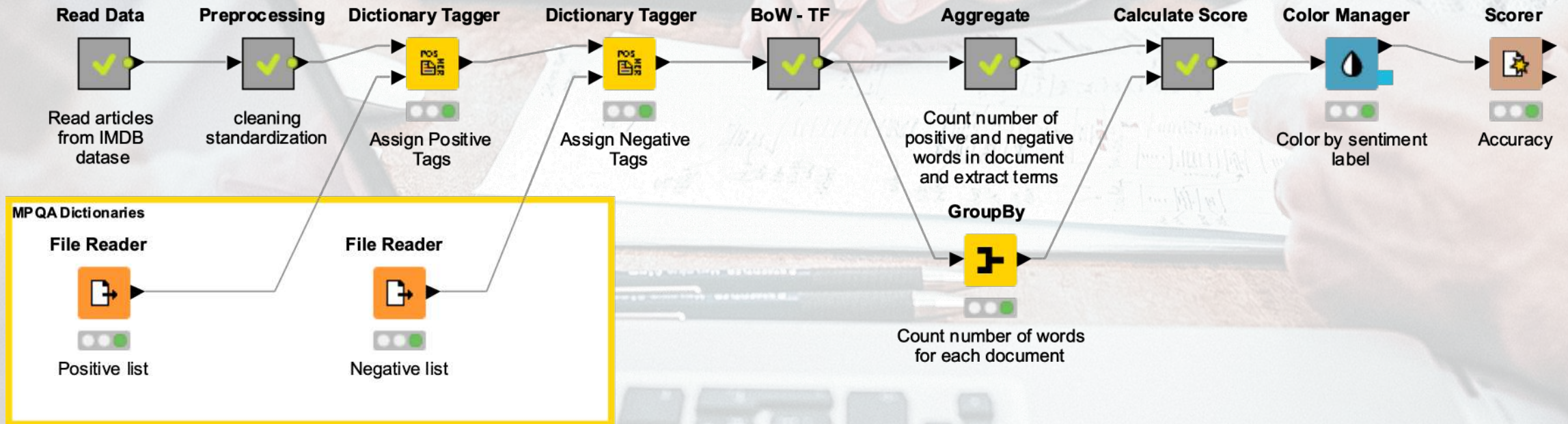
Alpha: Beta:

No of iterations: No of threads:

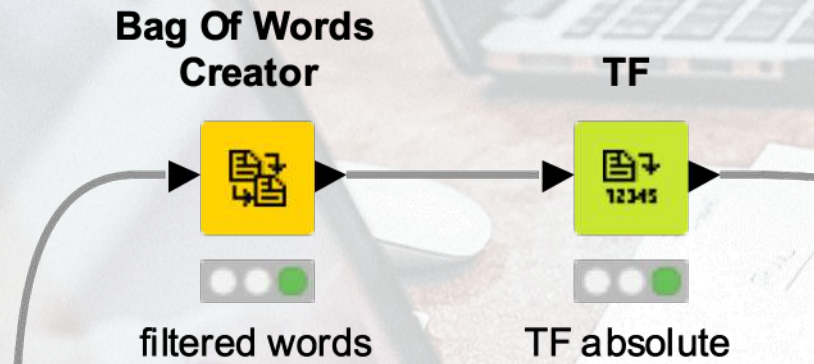
OK Apply Cancel ?

Діалогове вікно з
відповідними
налаштуваннями

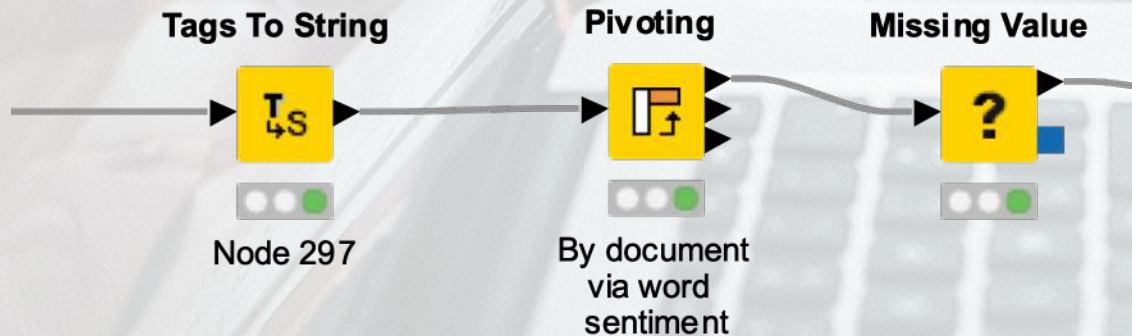
Sentiment analysis на базі лексичного підходу



Sentiment analysis: BoW, частота, підрахунок кількості слів

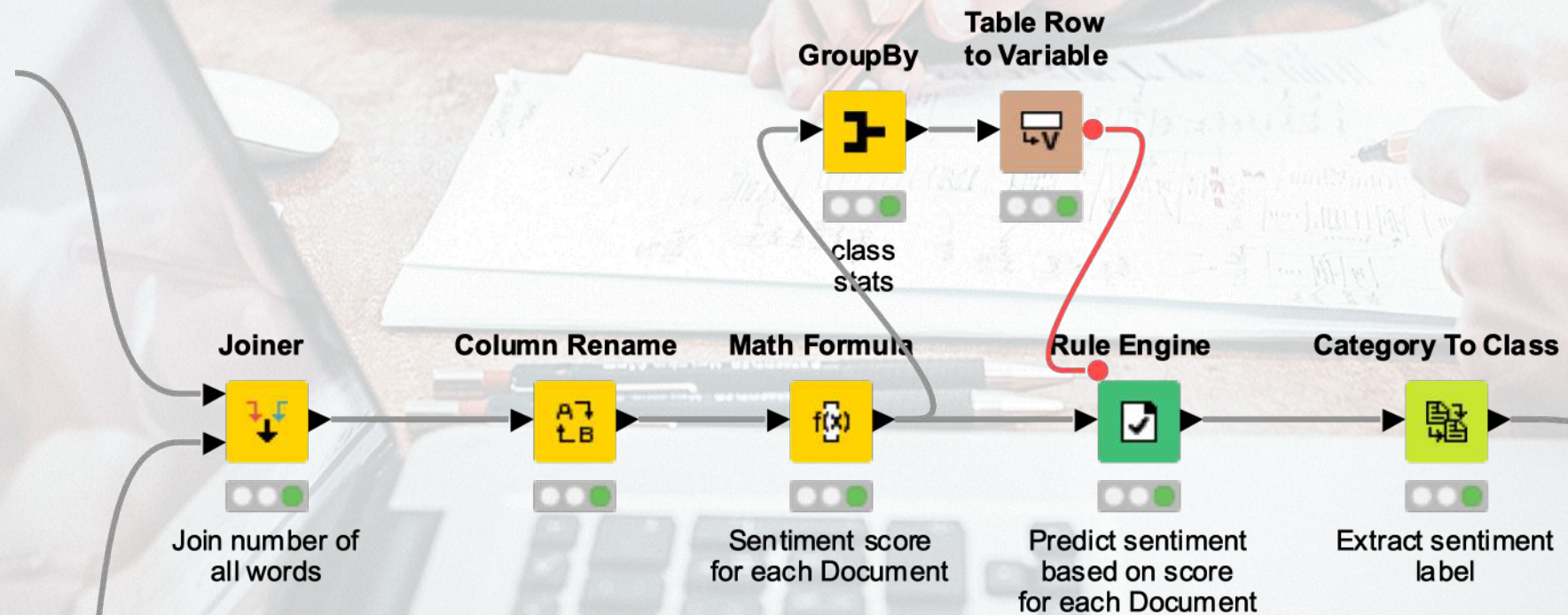


Вузли обчислення частоти терміну та створення сумки слів



Підрахунок термінів відповідно до тегів сентиментів

Sentiment analysis: обчислення оцінки документу за сентиментами



Sentiment analysis: результати обробки

Accuracy statistics - 3:310 - Scorer (Accuracy)

File Hilite Navigation View

Table "default" - Rows: 3 Spec - Columns: 11 Properties Flow Variables

| Row ID | TrueP... | FalseP... | TrueN... | False... | Recall | Precisi... | Sensiti... | Specifity | F-me... | Accur... | Cohen... |
|---------|----------|-----------|----------|----------|--------|------------|------------|-----------|---------|----------|----------|
| POS | 675 | 325 | 639 | 361 | 0.652 | 0.675 | 0.652 | 0.663 | 0.663 | ? | ? |
| NEG | 639 | 361 | 675 | 325 | 0.663 | 0.639 | 0.663 | 0.652 | 0.651 | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 0.657 | 0.314 |

Статистика точності для усіх стовпців

Confusion Matrix - 3:310 - Scorer (Accuracy)

File Hilite

| Sentiment ... | POS | NEG |
|---------------|-----|-----|
| POS | 675 | 361 |
| NEG | 325 | 639 |

Correct classified: 1,314 Wrong classified: 686

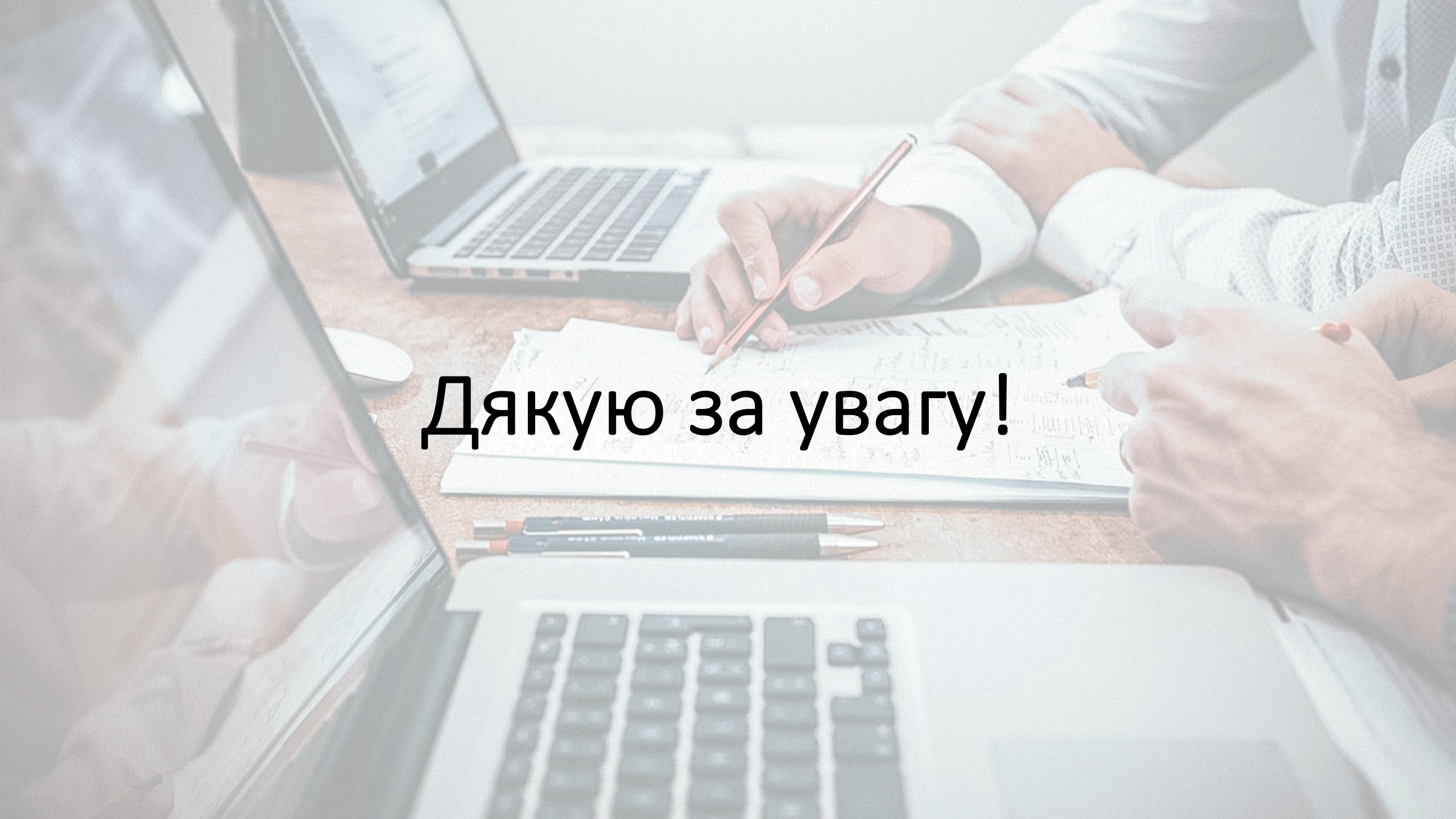
Accuracy: 65.7 % Error: 34.3 %

Cohen's kappa (κ) 0.314

Загальна статистика за усіма параметрами

Висновки

- Досліджено методи та алгоритми обробки неструктурованого тексту.
- Вивчено середовище та його функціональні можливості.
- Реалізовано процес інтелектуального аналізу природної мови на базі аналітичної платформи KNIME з використанням розширення Text Processing.

A person is sitting at a desk, writing in a notebook with a red pencil. The desk is cluttered with two laptops, several pens, and a mouse. The person is wearing a light-colored, patterned shirt. The background is slightly blurred, showing a window with a view of a building. The text "Дякую за увагу!" is overlaid in the center of the image.

Дякую за увагу!