

# Recuperación y clasificación en textos

Gerardo Fernández Rodríguez

Manuel Payán Cabrera

# Introducción

- Recuperación de información en la web
- Clasificación de textos recuperados

# Recuperación de información en la web

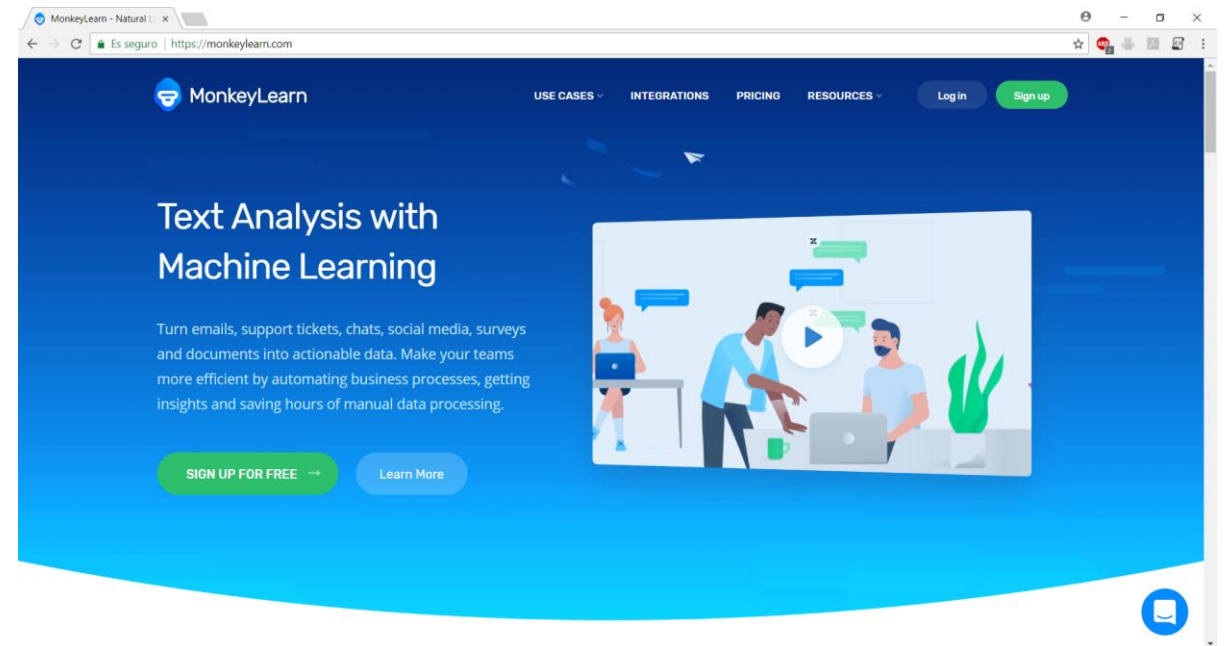
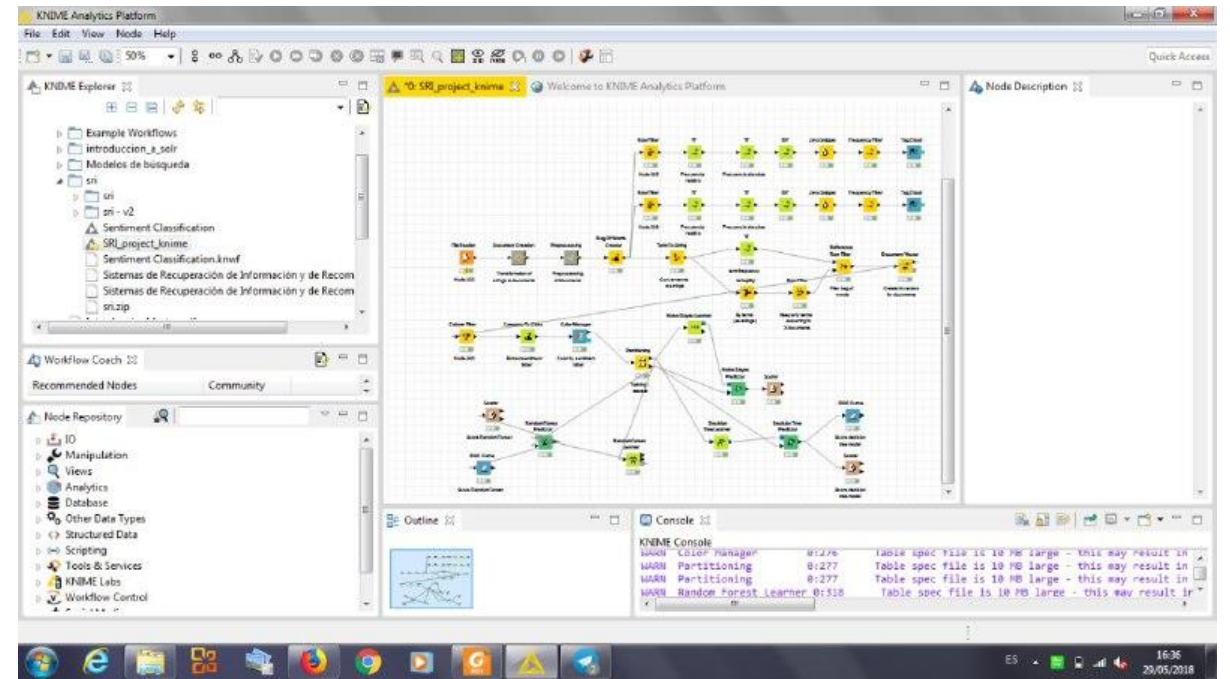
- TripAdvisor
- Crawler (Python+Xpath)
- Resultado obtenido

The screenshot shows the TripAdvisor page for NH Collection Granada Victoria. The page includes a header with the TripAdvisor logo and navigation links. The main content area displays the hotel name, a rating of 4.5 stars, and a user review. The review is by josemiguel2resviera, dated 3 weeks ago, and has a rating of 5 stars. The review text is: "Si un viajero busca un hotel con perfecta ubicación, personal dulce y diligente, y extraordinarias instalaciones, este es, sin duda el que están buscando! Nos trataron fabulosos! Nos sentimos como en casa! Altamente recomendable! No se lo pierdan. Mas". The review is highlighted with a red box. Other elements highlighted with red boxes include the hotel name, the number of people viewing the hotel (8 personas están mirando este hotel), and the review title "Excelentes habitaciones, servicios y atención".

```
def parse_review(self, response):
    item = HotelSentimentItem()
    item['hotel'] = response.xpath('//span[@class="ui_header h2"]/text()').extract()[0]
    item['title'] = response.xpath('//div[contains(@class,"quote")]/h1/text()').extract()[0]
    item['content'] = response.xpath('//div[@class="entry"]/p/text()').extract()[0]
    item['stars'] = response.xpath('//span[contains(@class, "ui_bubble_rating")]/@alt').extract()[0]
    return item
```

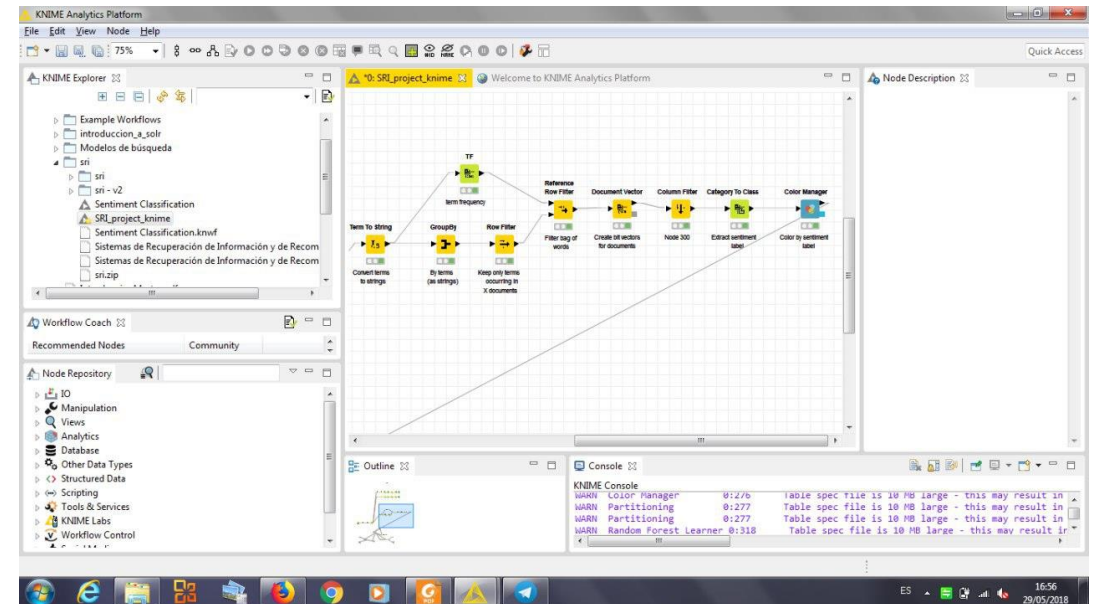
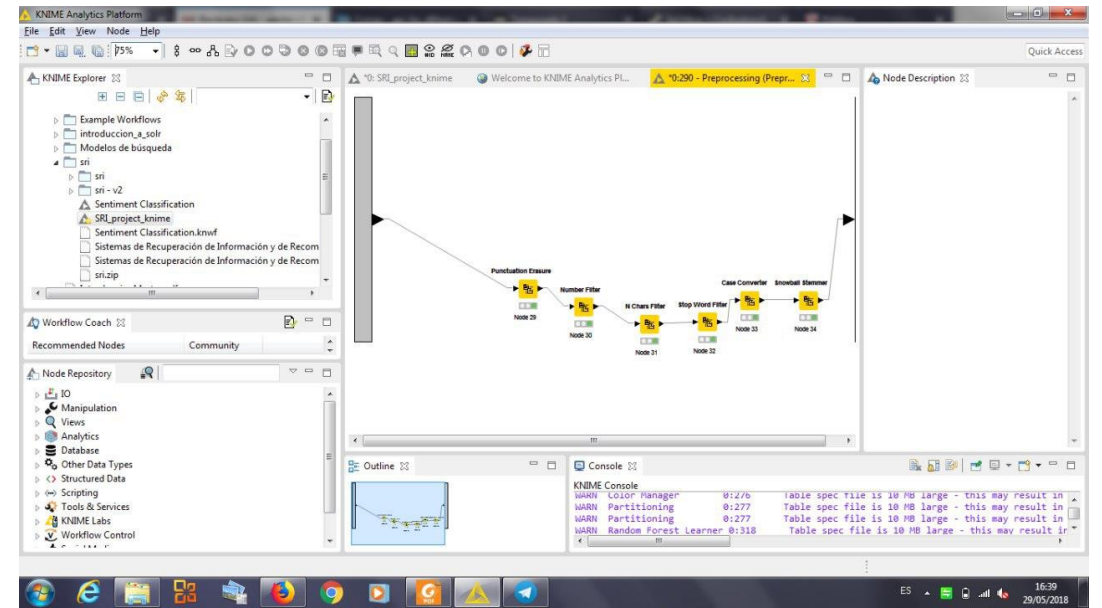
# Clasificación de textos recuperados

- KNIME
- MonkeyLearn



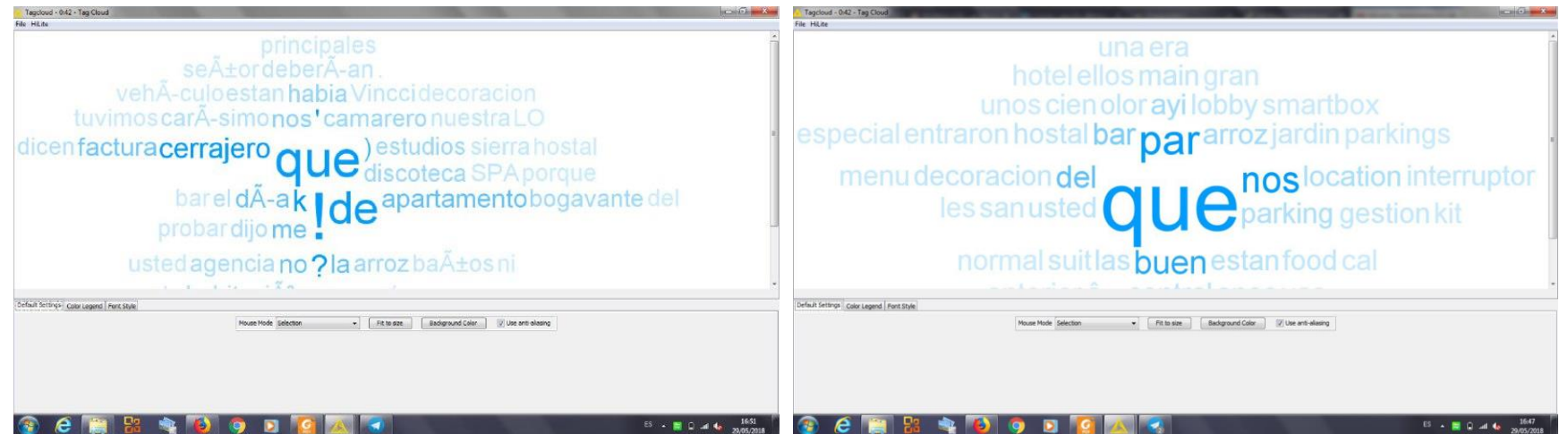
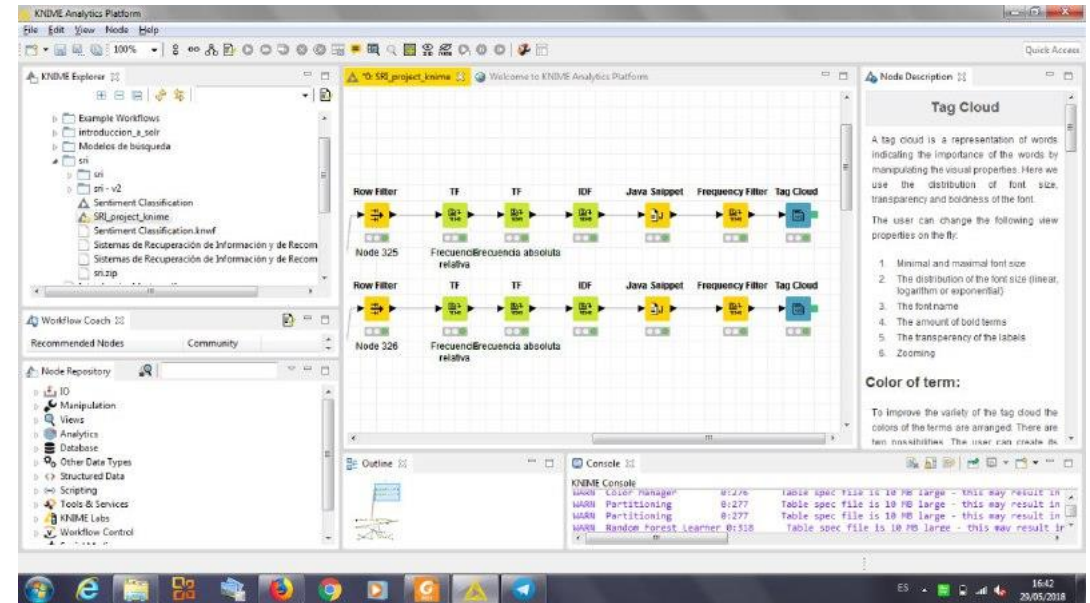
# KNIME: Preprocesamiento

- Punctuation erasure
- Number filter
- Nchar filter
- Stop words
- Case converter
- Stemmer
- Bag of Words
- Resultado



# KNIME: Tag Cloud

- Motivación
- Comparativa: antes y después del preprocesamiento





# KNIME: clasificación de textos

- Naive Bayes
- Decision Tree
- Random Forest

## Naive Bayes

La matriz de confusión obtenida tras aplicar este clasificador es:

	NEG	POS
NEG	11	213
POS	37	669

Por su parte los resultado de accuracy y kappa son:

- Accuracy: 0.73
- Kappa: -0.005

## Decision Tree

Usa un árbol de decisión con un motor de inferencia C4.5 para clasificar los comentarios que se han obtenido. Los resultados son los siguientes:

	NEG	POS
NEG	97	127
POS	109	597

- Accuracy: 0.746
- Kappa: 0.287

## Random Forest

Es una generalización sobre el anterior donde se lanzan distintos árboles de decisión y se van conjuntando sus resultados

	NEG	POS
NEG	47	177
POS	0	706

- Accuracy: 0.81
- Kappa: 0.287

# MonkeyLearn

- Análisis de textos con Machine Learning
- ElasticSearch

The screenshot displays the MonkeyLearn interface for a text classification project. At the top left, a decision tree shows a 'Root' node branching into 'bad' and 'good' nodes. To the right, a statistics panel shows 33,143 subtree samples and 0 category samples. Below this, three donut charts represent Accuracy (67%), Precision (--), and Recall (--). A keywords panel shows a word cloud with terms like 'gran', 'lun', 'sercotel', 'lun', 'gran', 'afectu', 'eduard', 'futura', 'visit', 'atent', 'juan', 'lopez', 'meli', 'gran', 'hotel', 'daur', 'chavel', 'esper', 'futura', 'vais', 'ocasion', 'juan', 'lopez', 'bont', 'gran', 'hotel', 'alixar', 'daur', 'cuv', 'atent', 'juan', 'ocasion', 'juan', 'lopez', 'lun', 'veri', 'futura', 'vais', 'esper', 'futura', 'ban', 'arab'. Below the keywords, a text input field contains the sentence: 'Estamos muy contentos con la elección del hotel. El servicio ha sido muy bueno y el buffet libre espectacular!'. A 'Submit' button is visible. The bottom section shows the 'SRI\_project' project details, including 'Sandbox', 'Live', and 'API' tabs. The 'Sandbox Parameters' section includes 'Language' (Spanish), 'N-gram range' (Unigrams, Bigrams and Trigrams), 'Normalize weights' (checked), 'Preprocess social media' (unchecked), 'Filter stopwords' (checked), and 'Use whitelist' (unchecked). The 'Text' input field is empty. The 'HTTP/1.1 200 OK' response shows the classification result: 'text': 'Estamos muy contentos con la elección del hotel. El servicio ha sido muy bueno y el buffet libre espectacular!', 'external\_id': null, 'error': false, 'classifications': [{ 'tag\_name': 'good', 'tag\_id': 54971364, 'confidence': 0.631, 'parents': [] }].



# Conclusiones

- Necesidad de preprocesamiento
- Clasificación supervisada vs clasificación no supervisada