Recuperación y clasificación en textos

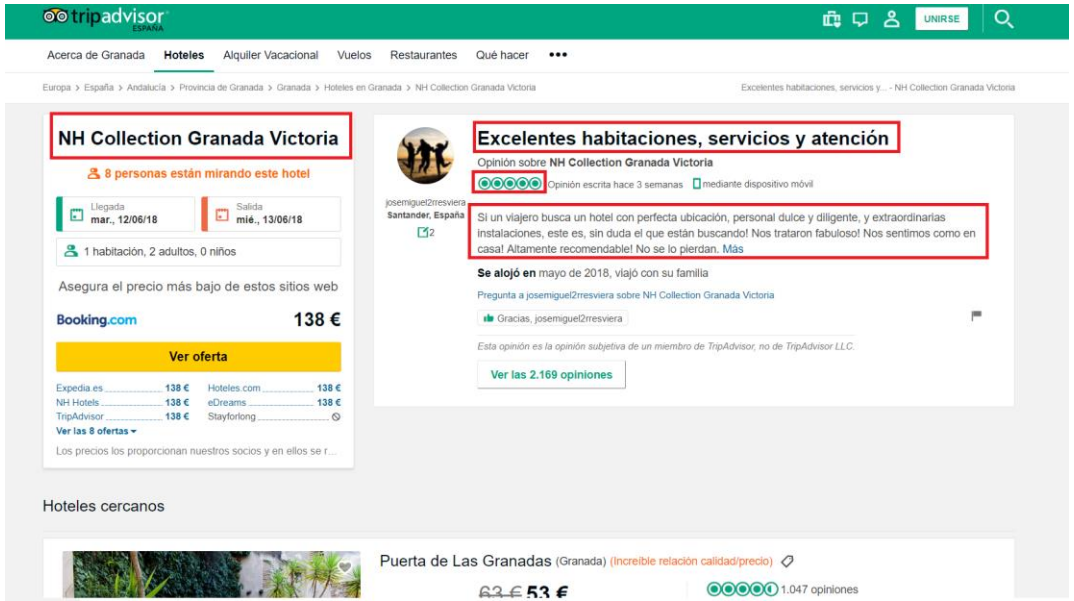
Gerardo Fernández Rodríguez

Manuel Payán Cabrera

Introducción

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

Recuperación de información en la web

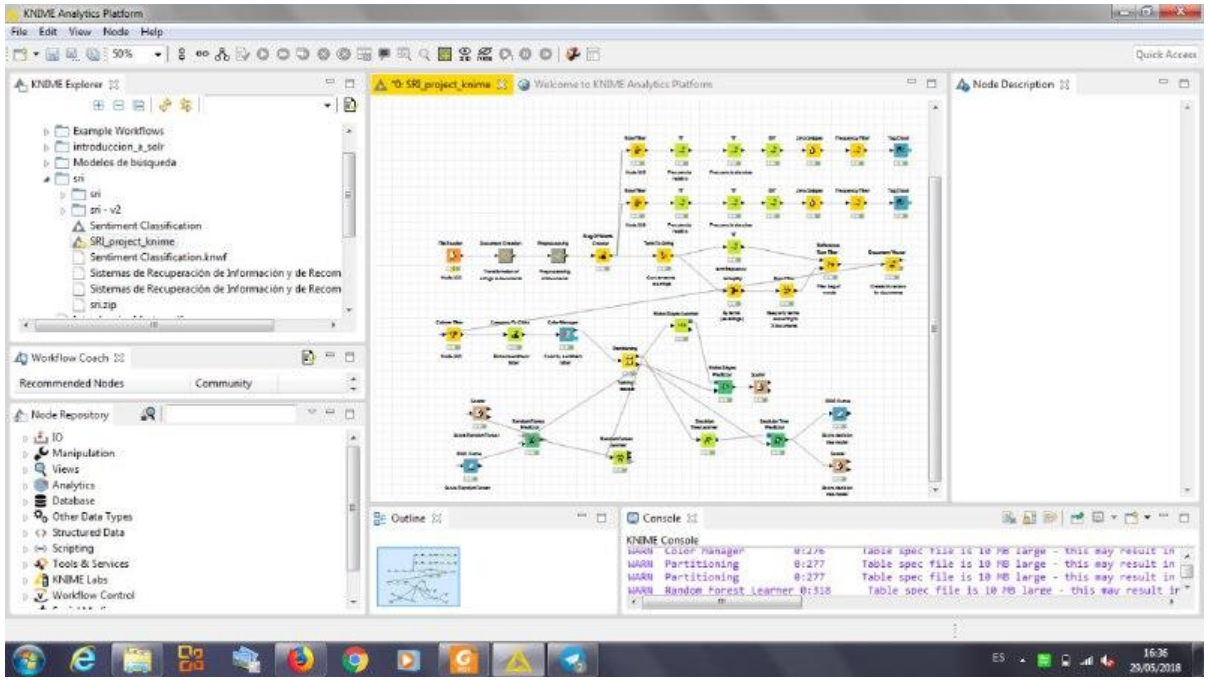


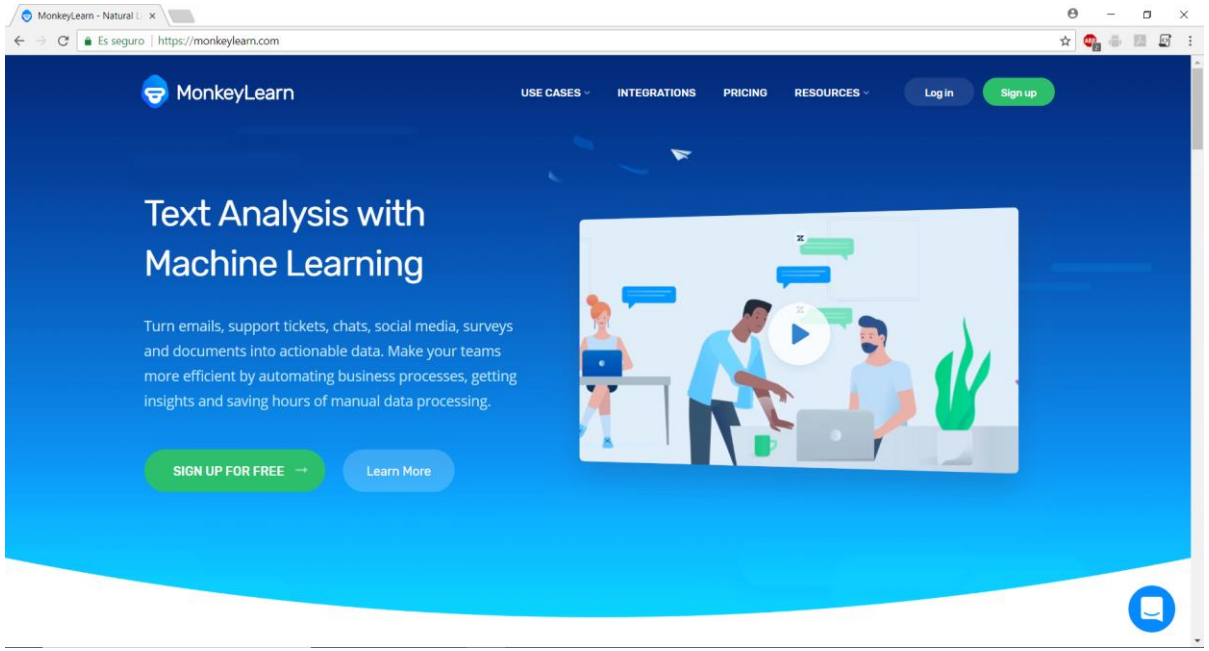
- TripAdvisor
- Crawler (Python+Xpath)
- Resultado obtenido

```python
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

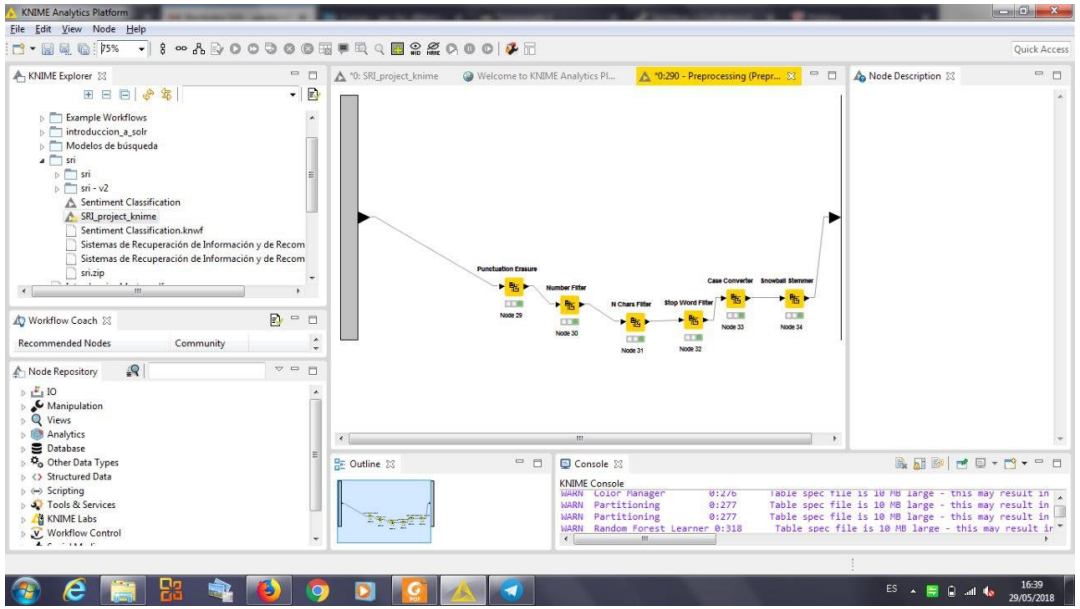
- KNIME
- MonkeyLearn

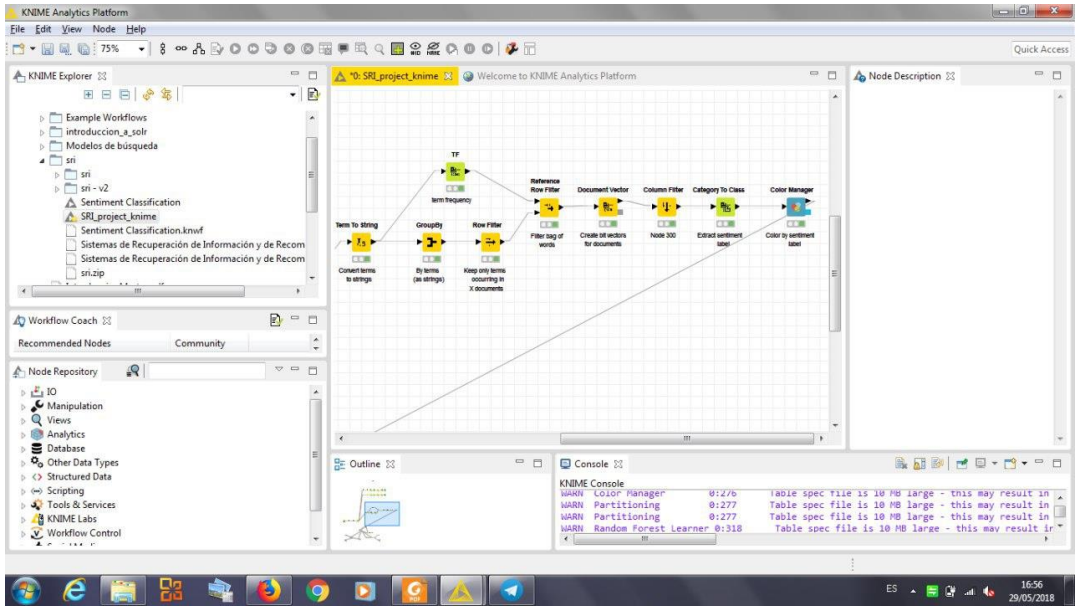




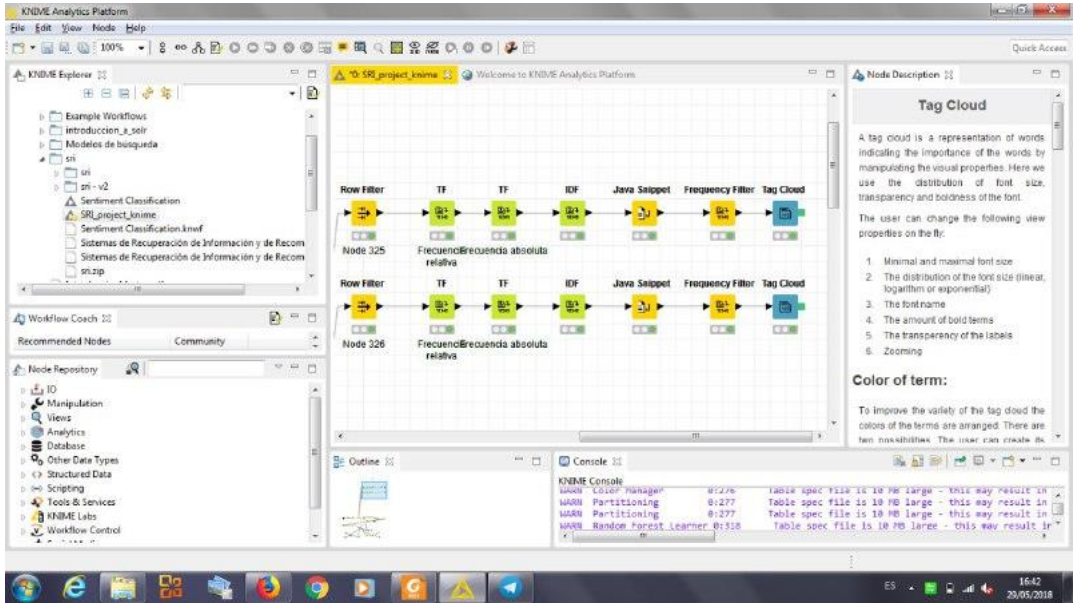
KNIME: Preprocesa-miento

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

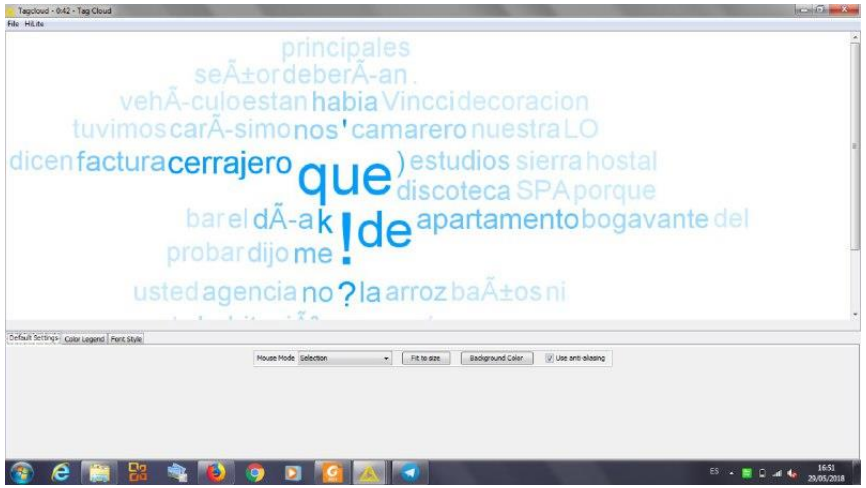


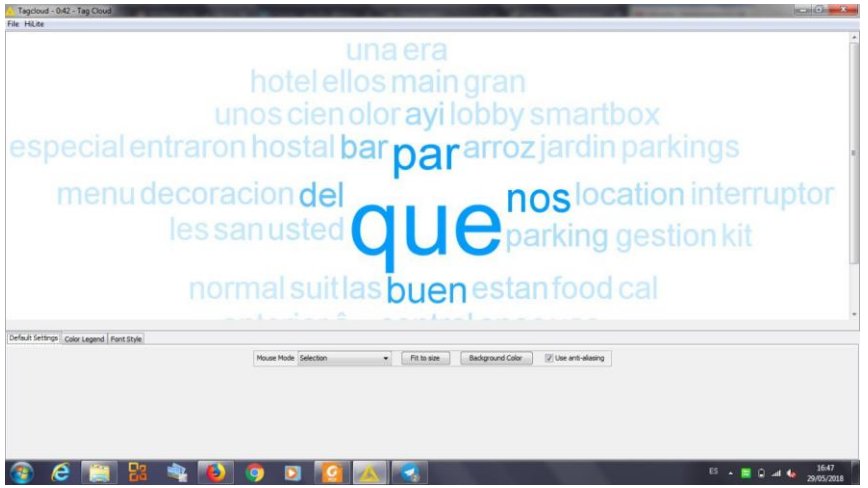


KNIME: Tag Cloud



- 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

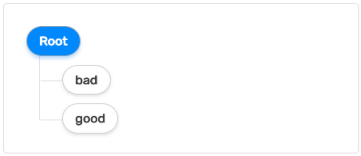
Random Forest

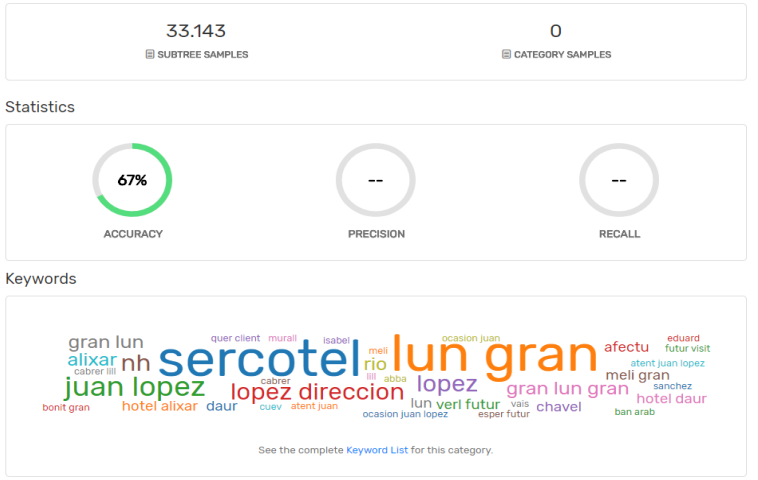
Es una generalización sobre el anterior donde se lanzan distintos árboles de decisión y se van conjuntando sus resultados

	NEG	POS
NEG	47	177
POS	0	706

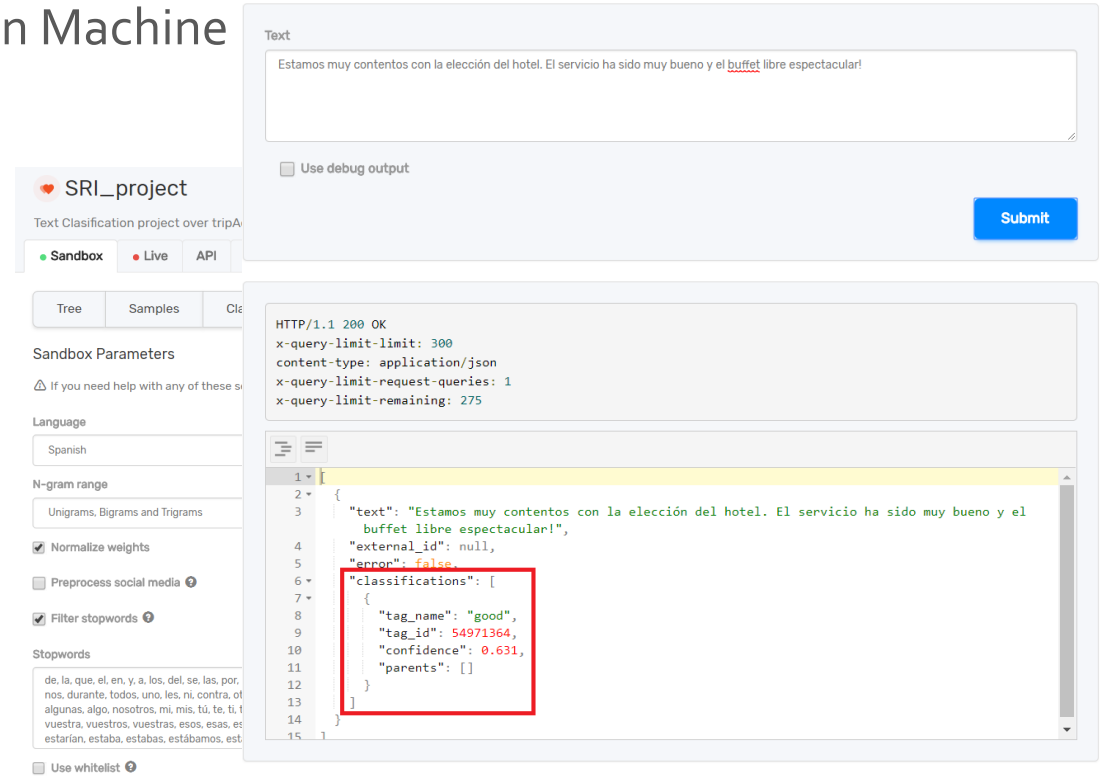
- Accuracy: 0.81
- Kappa: 0.287

MonkeyLearn





- Análisis de textos con Machine Learning
- ElasticSearch



Conclusiones

- Necesidad de preprocesamiento
- Clasificación supervisada vs clasificación no supervisada