



UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE CIENCIAS QUÍMICAS, FÍSICAS Y MATEMÁTICAS

DIRECCIÓN DE LA UNIDAD DE LA UNIDAD DE INVESTIGACIÓN-FCQFM

Procesamiento de Lenguaje Natural (NLP) en el drama Ollantay



Chullo Oscar¹ Zuñiga Arturo²

¹E.P. de Matemática, UNSAAC, Cusco, Perú; ²E.P. de Matemática, UNSAAC, Cusco, Perú

¹ 193029@unsaac.edu.pe; ² arturo.zuniga@unsaac.edu.pe

Resumen

El propósito del siguiente póster es dar a conocer el **Procesamiento de Lenguaje Natural (NLP)** para el procesamiento de textos en el campo del **aprendizaje supervisado**, mostrándose una aplicación en el texto del drama Ollantay, en este se definen los **token** y **n-gramas** como unidades de análisis para conocer a rasgos generales un análisis preliminar del texto.

Palabras Clave: Procesamiento de Lenguaje Natural, aprendizaje supervisado, token, n-gramas.

Introducción

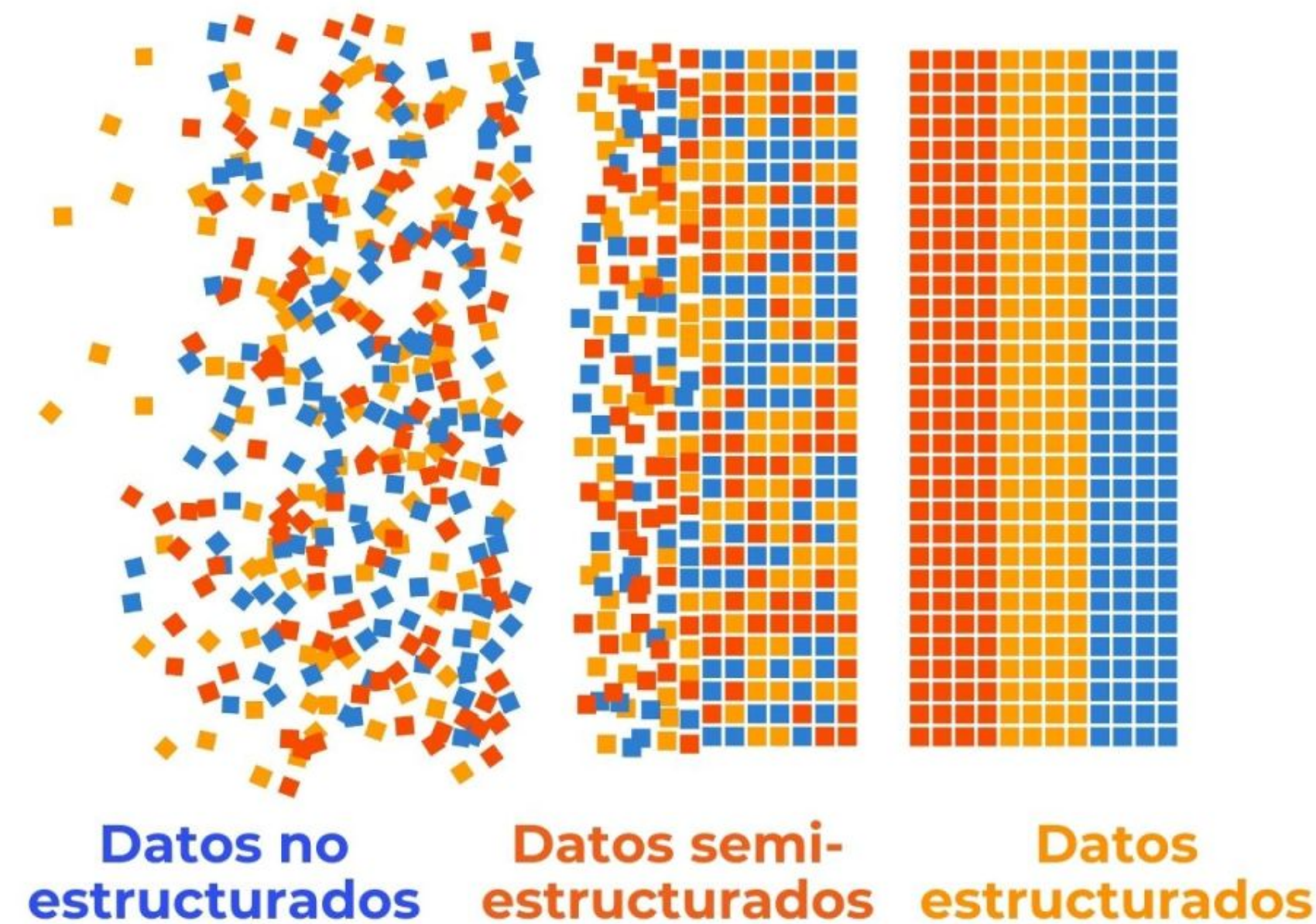
El **lenguaje natural** es un lenguaje desarrollado y evolucionado por los humanos a través del uso y comunicación “natural”. Tal lenguaje puede ser expresado de manera escrita, verbal o incluso con signos.(Sarkar 2016)

El **procesamiento del lenguaje natural (NLP)** es un campo de estudio que se centra en la interacción entre el lenguaje natural y la computación, implica el uso de técnicas computacionales para analizar, comprender y generar texto o voz en lenguaje natural.

Metodología

Para realizar los análisis en los textos (datos no estructurados, ver **Fig. 1**), se definen los **token** como unidad significativa de los textos, en términos de la minería de textos, un **token** puede almacenar una sola **palabra**, una conjugación de **n-gramas**, una **oración** o un **párrafo**.(Silge and Robinson 2017)

Figura 1: Tipos de datos

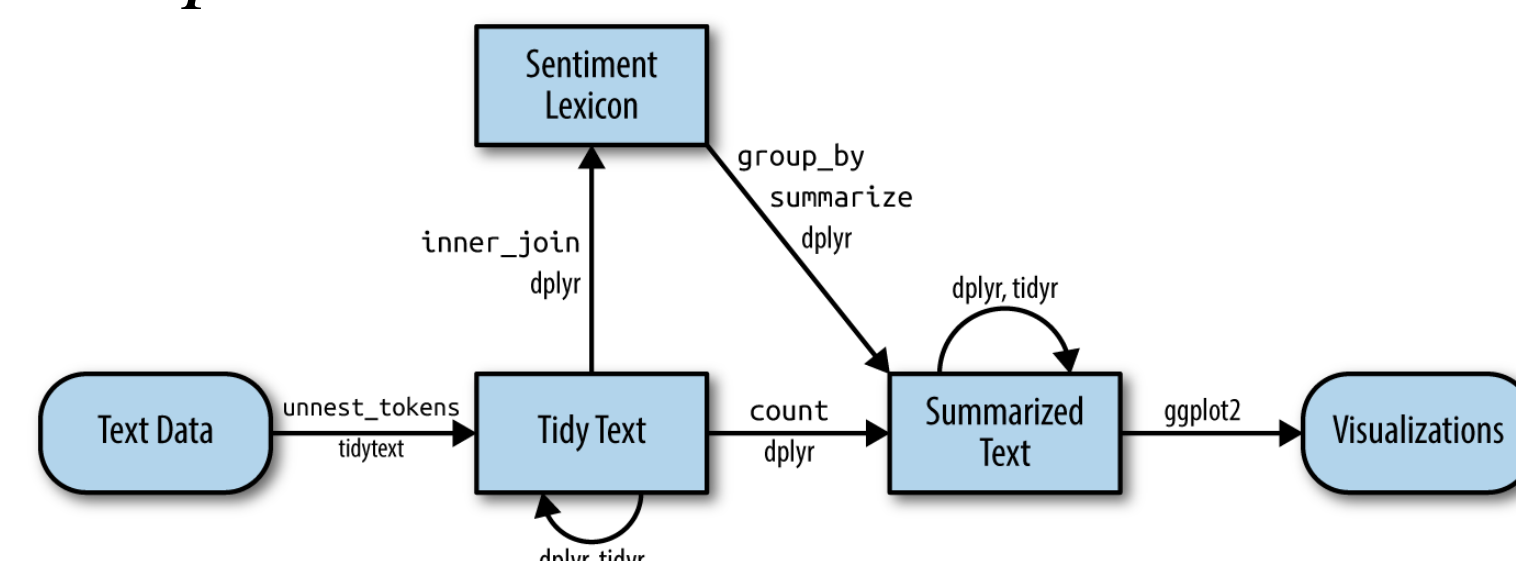


Estructurar los datos de texto significa que se ajustan a los principios de datos ordenados y se pueden manipular con un conjunto de herramientas consistentes:

1. **Cadena:** Como vectores de caracteres.
2. **Corpus:** Conjunto de cadenas sin procesar.
3. **Matriz de documentos y términos:** Describe un corpus con una fila para cada documento y una columna para cada término.

Si queremos analizar un texto, se puede realizar el análisis de **token** por palabras individuales o por grupos de palabras, tomando en consideración el **token**, se puede evaluar un texto de forma sentimental (ver **Fig. 2**), para así clasificar si el **token** es **positivo** o **negativo**, o considerar otros criterios de evaluación.

Figura 2: Diagrama de flujo de un análisis de texto para el análisis de sentimientos



En caso se quiera hacer un análisis para evaluar los **token** y su relación entre sí, se recurre a los **n-gramas** para examinar que **token** tienden a seguir a otros, dependiendo del número de **token** se emplean los sufijos: **Bi**, **Tri**, **Tetra**, etc.

Bi-grama

$$X \longrightarrow Y$$

Tri-grama

$$X \longrightarrow Y \longrightarrow Z$$

Para realizar un análisis de textos, en el software R se puede acceder a la colección **Proyecto Gutenberg** mediante el paquete **gutenbergr**, del cual se ha extraído la versión traducida al Inglés del drama **Ollantay**. El cual se consideró como la base de datos textuales a analizar mediante el **NLP**.

Resultados

Análisis de sentimientos

El drama **Ollantay** nos presenta una narrativa sobre el amor, la justicia y la lucha social. Con un análisis de **token**, se obtuvieron los resultados que se muestran en **Fig. 3** y **Fig. 4**.

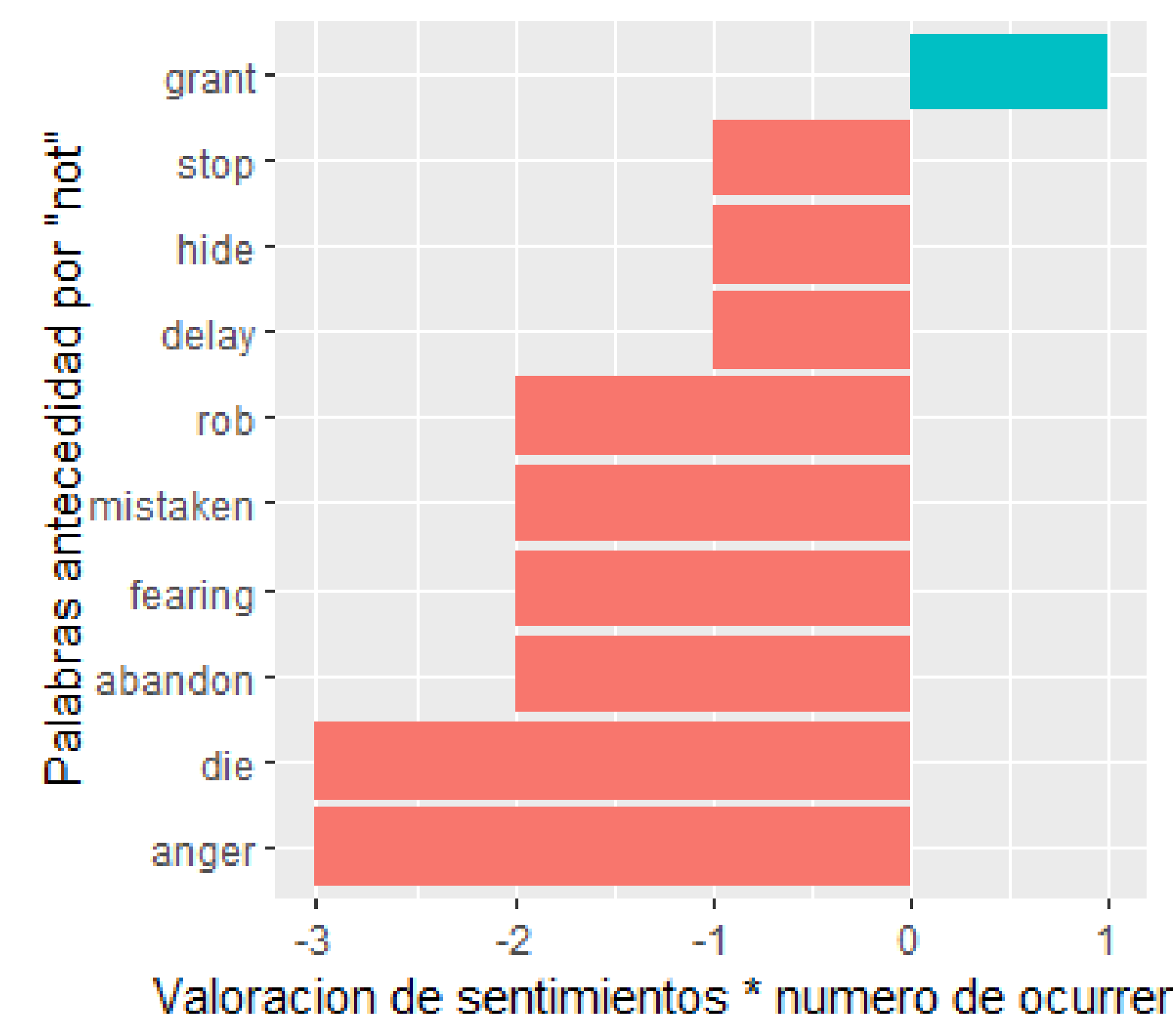
Figura 3: Clasificación de sentimientos

negative



positive

Figura 4: Valoración de sentimientos con bi-gramas



Análisis de n-gramas Se han establecido los **token** con todas las palabras de la obra, al cual se hizo un análisis por **Bi-gramas** (ver **Fig. 5**) y **Tri-gramas** (ver **Fig. 6**)

Figura 5: Análisis correlacional de palabras con bi-gramas

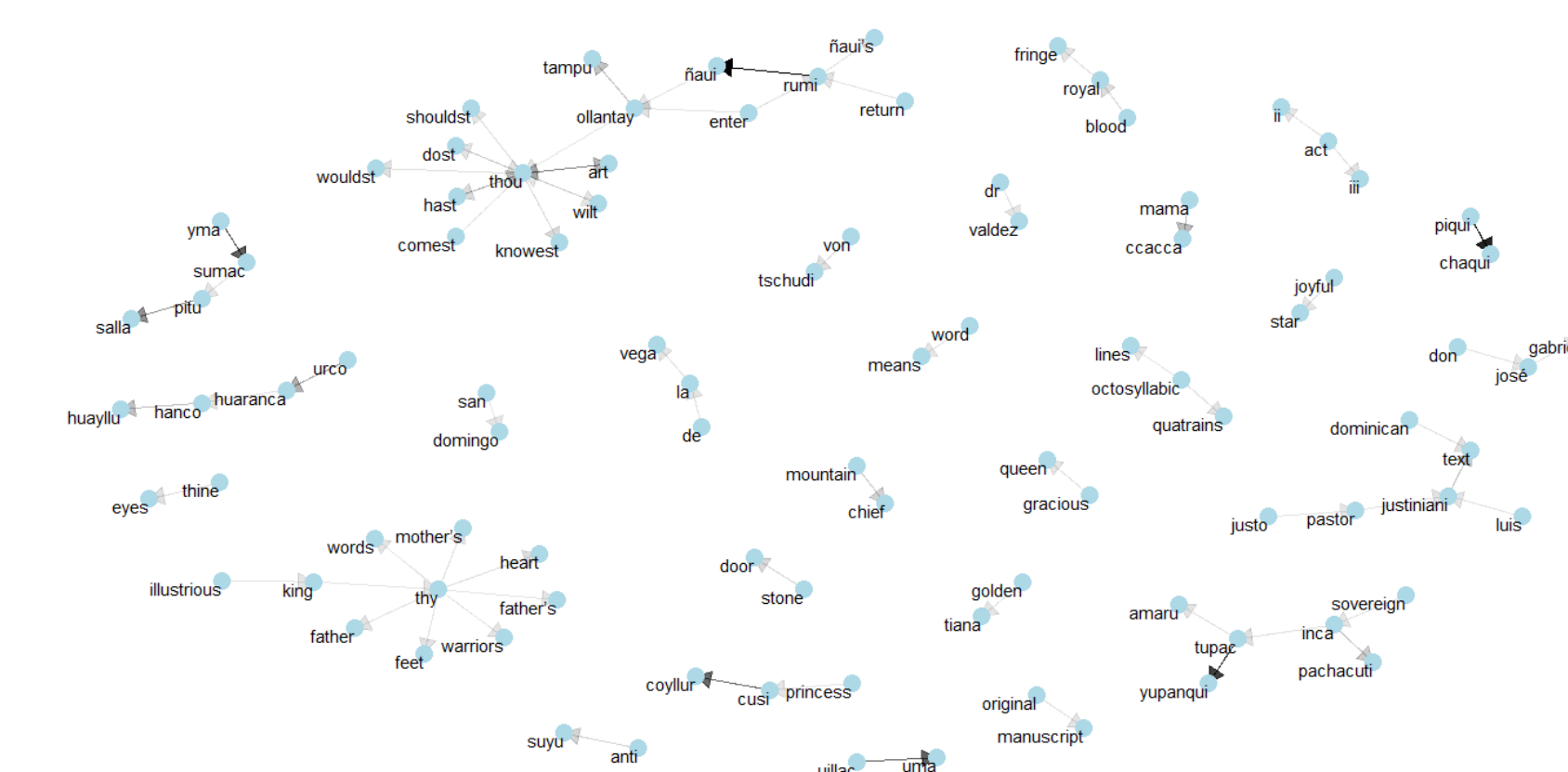
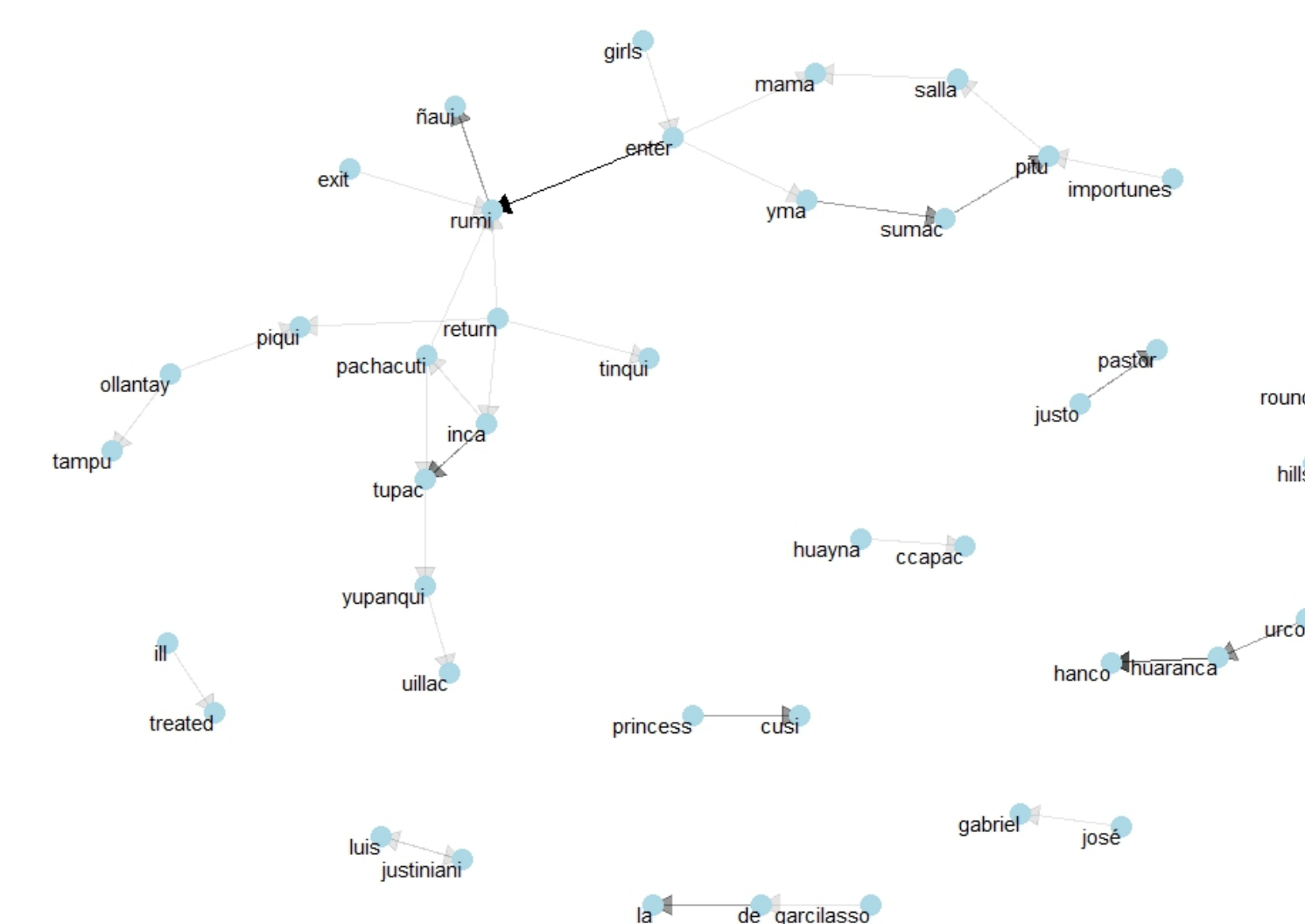


Figura 6: Análisis correlacional de palabras con tri-gramas



Conclusión

1. Con el análisis de sentimientos, se pudieron identificar sentimientos **positivos** y **negativos**, el cual guardan relación con la idea general de la obra. Sin embargo, cabe resaltar que un análisis con bi-gramas es efectivo para identificar adecuadamente los sentimientos.
2. En el análisis de los n-gramas se ha podido ver que un análisis con bi-gramas es más explicativo que por tri-gramas.

Referencias

- Sarkar, Dipanjan (2016). *Text Analytics with Python*. R package version 1.43. URL: <https://github.com/dipanjanS/text-analytics-with-python>.
- Silge, Julia and David Robinson (2017). *Text Mining with R*. URL: <https://www.tidytextmining.com/>.