UNED

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# DH@Madrid Summer School 2016

## Análisis computacional del soneto del Siglo de Oro – Parte 2

Boria Navarro Colorado – borja@dlsi.ua.es

erc European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos

Madrid, 1 de julio de 2016

DiXiT
Digital Scholarly Editions
Initial Training Network

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Índice

- Primera parte:
  - Introducción: marco metodológico.
  - Compilación y anotación de un corpus de sonetos del Siglo de Oro (SdO).
- Segunda parte:
  - Modelos computacionales para el análisis métrico y semántico.
  - Práctica: análisis del corpus a gran escala con buscador (web).

erc European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Análisis a gran escala

- ¿Cuáles son los gustos métricos generales de la sonetística áurea? ¿Son constantes o varían a lo largo del periodo?
- ¿Hay patrones métricos característicos de algún autor, escuela, periodo...?, ¿contrastan unos con otros? ¿Hay diferencias rítmico-métricas entre el Renacimiento y el Barroco? ¿Hay gustos comunes?
- ¿Se utilizan patrones métrico similares para tratar temas similares? ¿Hay un ritmo específico para el poema amoroso? ¿Y para el panegírico? ¿Hay relación entre la métrica y la semántica?
- ...

erc European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Análisis de la métrica del soneto del Siglo de Oro mediante estadística descriptiva

erc European Research Council

# Análisis métrico

- Análisis gustos métricos:
  - Frecuencias métricas.
  - Por épocas.
  - Evolución temporal.
  - Patrones métricos especiales.

Navarro Colorado (2016) “Hacia un análisis distante distante del endecasílabo áureo” *Rhythmica*.

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

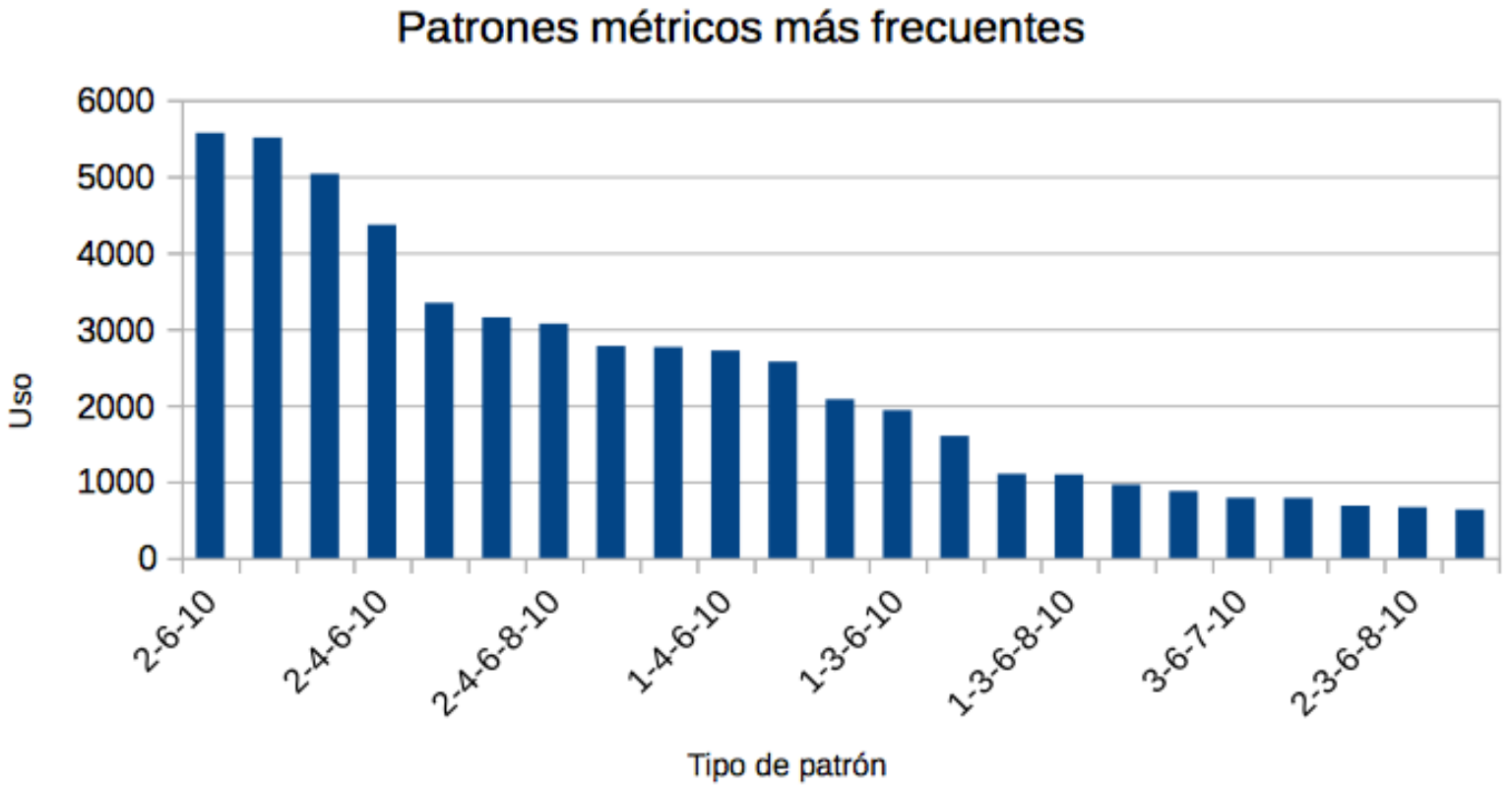
# Frecuencias métricas

| Posición | Patrón | Patrón Numérico | Cantidad | Frecuencia (%) |
|---|---|---|---|---|
| 1 | -+---+---+- | 2-6-10 | 5568 | 7,83 |
| 2 | -+-+---+-+- | 2-4-8-10 | 5506 | 7,74 |
| 3 | --+--+---+- | 3-6-10 | 5031 | 7,07 |
| 4 | -+-+-+---+- | 2-4-6-10 | 4365 | 6,14 |
| 5 | ---+-+---+- | 4-6-10 | 3342 | 4,70 |
| 6 | -+---+-+-+- | 2-6-8-10 | 3151 | 4,43 |
| 7 | -+-+-+-+-+- | 2-4-6-8-10 | 3069 | 4,31 |
| 8 | ---+---+-+- | 4-8-10 | 2776 | 3,90 |
| 9 | +--+---+-+- | 1-4-8-10 | 2763 | 3,88 |
| 10 | +--+-+---+- | 1-4-6-10 | 2717 | 3,82 |
| 11 | --+--+-+-+- | 3-6-8-10 | 2572 | 3,62 |
| 12 | ---+-+-+-+- | 4-6-8-10 | 2079 | 2,92 |
| 13 | +-+--+---+- | 1-3-6-10 | 1936 | 2,72 |

erc European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Frecuencias métricas




European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

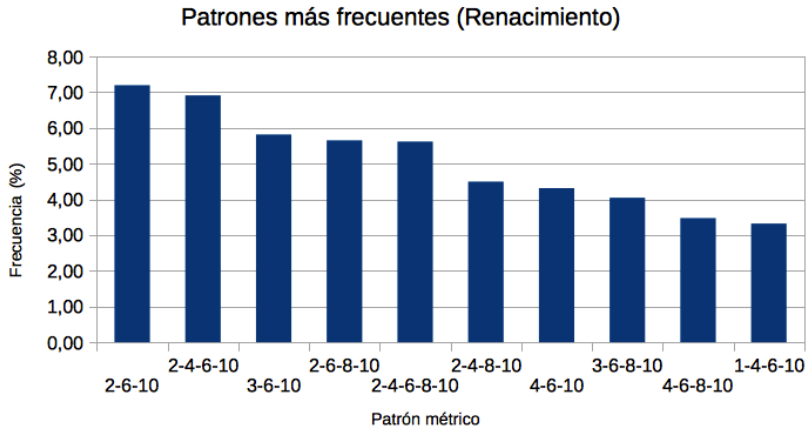
# Frecuencias métricas

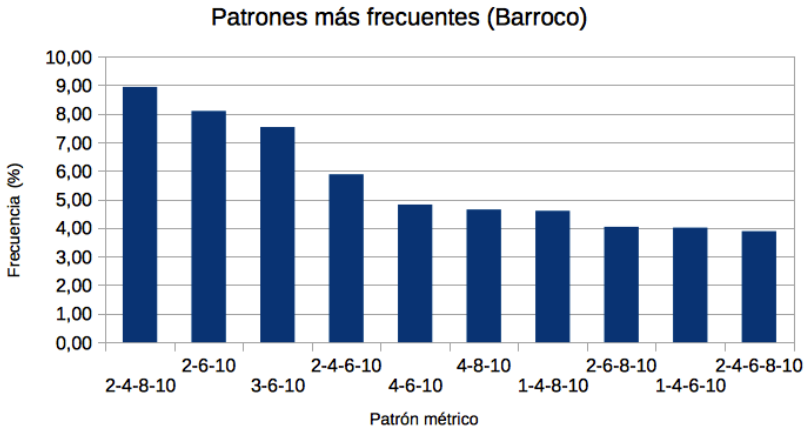
- Los tres patrones más frecuentes corresponden a los tres tipos básicos de endecasílabo:
  - Heroico: 2, 6, 10
  - Sáfico: 2, 4, 8, 10
  - Melódico: 3, 6, 10
- Hasta la posición 20, re-elaboraciones.
- Resto (fq < 1,5%):
  - Antirrítmicos, pocos apoyos métricos, erróneos.

erc European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Frecuencias métricas




erc European Research Council

# Frecuencias métricas

- Renacimiento:
  - Preferencia por patrones en 6.
  - Uso de patrones con 4 y 5 apoyos métricos.
- Barroco:
  - Preferencia por patrones en 4 y 8.
  - Preferencia por patrones con sólo 3 apoyos.

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

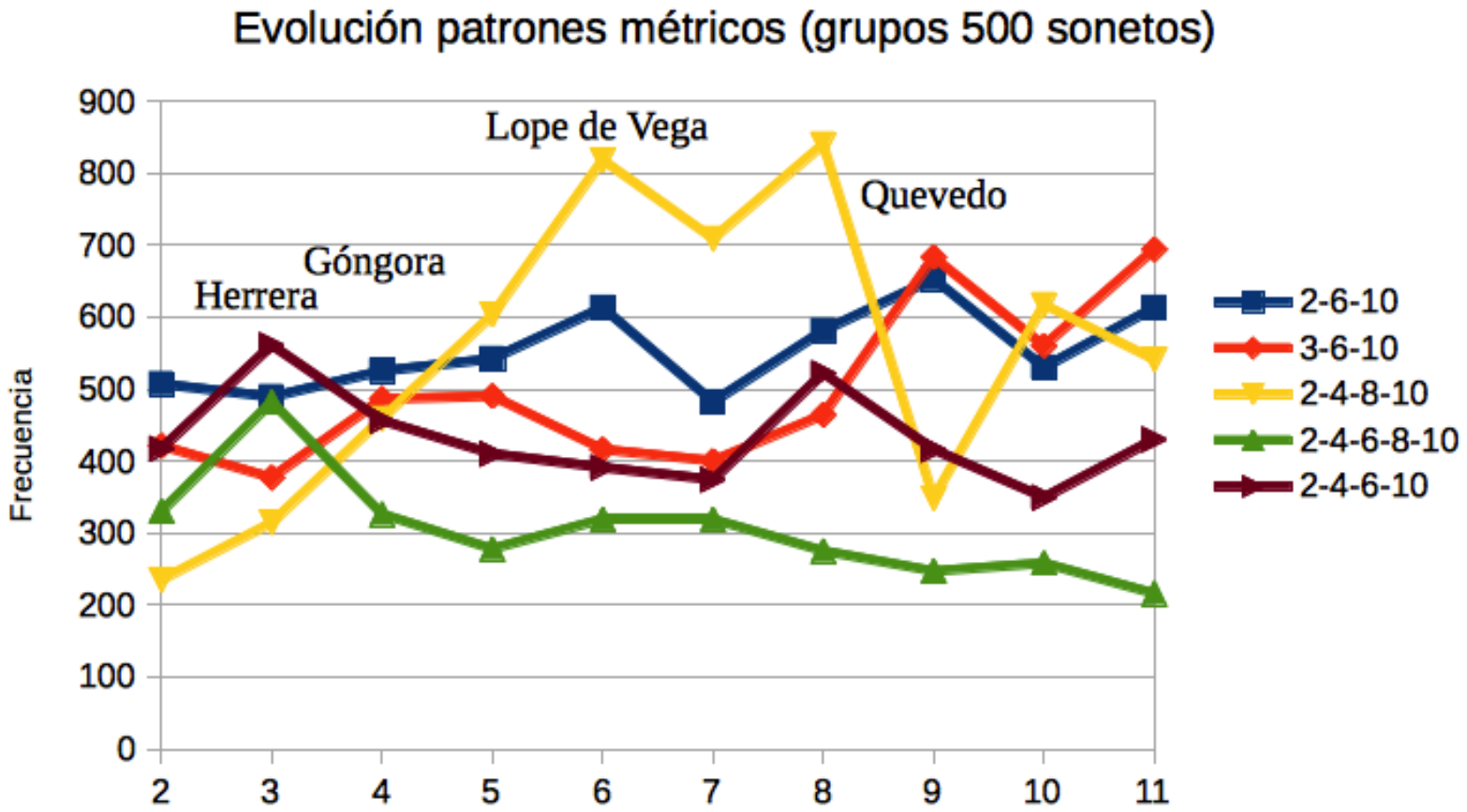
# Evolución temporal

- Problema de representatividad:
  - Garcilaso infra-representado (38 sonetos)
  - Lope de Vega supra-representado (+1000 sonetos).
- Organización temporal por fecha de nacimiento/fallecimiento de cada autor.
- Grupos de 500 versos.

erc European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Evolución temporal




European Research Council

# Evolución temporal

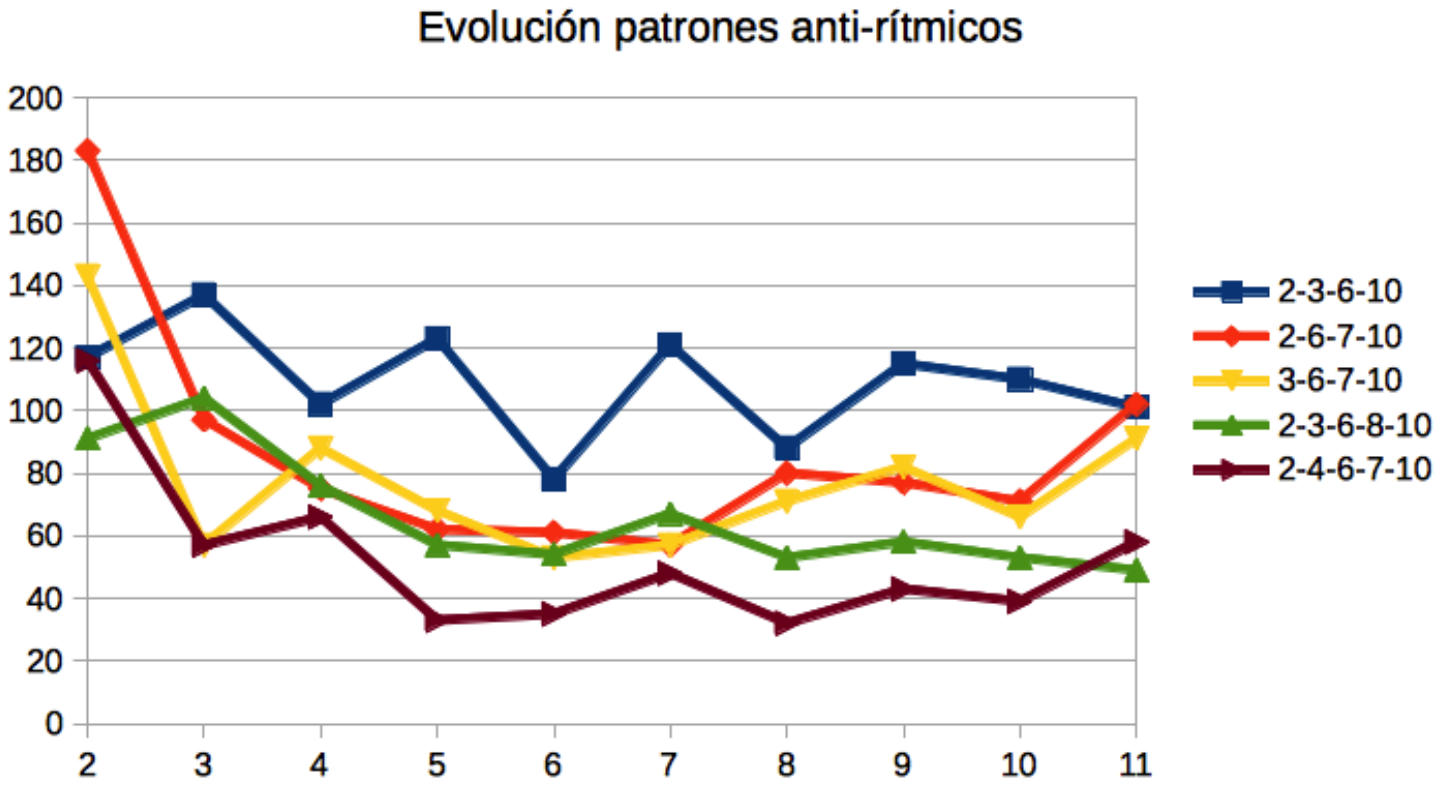
- 2ª generación Renacimiento (gr. 3 - Herrera):
  - Gusto por patrones con apoyos métricos explícitos (4 ó 5).
- 1ª generación Barroco (grs. 4-8):
  - Gusto creciente por metro sáfico (4, 8, 10).
  - Preferencia por patrones con pocos apoyos métricos.

# Evolución temporal

- 2ª generación Barroco (gr. 9 - 11):
  - Quevedo: vuelta a gustos métricos renacentistas (Garcilaso).
  - Otros autores (Tirso de Molina, Soto de Rojas, etc.) continúan la tradición barroca de Lope de Vega.
  - Patrón melódico (ritmo ternario: 3, 6, 10).

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Antirritmia




erc European Research Council

# Antirritmia

- Preferencia antirritmia en 2-3.
- Primer renacimiento: preferencia acusada por antirritmia en 6-7.
  - Hipótesis: rítmica similar al verso medieval del s. XV.
  - Lectura con cesura: ritmo similar al dodecasílabo 7+5.
  - Íñigo López de Mendoza: preferencia por endecasílabos con apoyo métrico en 7.

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE iNNOVACIÓN EN HUMANIDADES DIGITALES

# Conclusiones

- Punto de vista diferente: distante, a gran escala.
  - Datos generales descriptivos de todo el periodo.
- Corpus representativo hecho literario:
  - Amplio: autores
  - Anotado: métrica
- Método: estadística descriptiva.

erc European Research Council

# ¿Cómo se ha hecho?

- Cálculo de frecuencias con Python:
  - Cargar y analizar XML
  - Extracción patrones métrico y cálculo de frecuencias
  - [Opción alternativa] Analizar XML con R:
    - Cfr. Jockers (2014), *Text Analysis with R for Students of Literature*, págs. 89 y ss.

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE iNNOVACIÓN EN HUMANIDADES DIGITALES

# ¿Cómo se ha hecho?

- Dividir corpus en subcorpus (por época, temporal, por autor, etc.)  y extraer datos de cada uno.
- Ejercicio práctico: uso de buscador web con GUI.

erc European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Métodos avanzados: Análisis semántico del soneto del Siglo de Oro mediante espacios semánticos vectoriales

erc European Research Council

# Semántica distribucional

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Significado contextual de las palabras:
  - "You shall know a word by the company it keeps" (Firth 1957)
  - "Words will occur in similar contexts if and only if they have similar meanings" (Harris 1951)

erc European Research Council

# Espacio semántico vectorial

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Representación vectorial.
  - Espacio Semántico Vectorial:
    - Matriz Palabras X Contexto
  - Un vector representa el significado de una palabra.
  - Cada vector está formado por las frecuencias de aparición del vector en cada contexto.

erc European Research Council

# Espacio semántico vectorial

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Similitud:
  - Permite realizar cálculos de similitud.
  - Cuanta mayor similitud haya entre dos palabras, más semejantes serán sus significados.

erc European Research Council

# LDA Topic Modeling

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- El significado (distribucional) de una palabra está representado por *topics*.
- Cada palabra se asigna a cada topic en función de:
  - Cuántas veces se ha asignado la palabra al *topic* previamente.
  - El principal *topic* del texto donde aparece la palabra.

—

Blei, D.M., (2012) "Probabilistic Topic Models"

erc European Research Council

# LDA Topic Modeling

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Al final del proceso:
  - Extracción de los principales *topics* del corpus.
  - Cada *topic* es un conjunto de palabras representativas.

erc European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# LDA Topic Modeling

| "Genetics" | "Evolution" | "Disease" | "Computers" |
|---|---|---|---|
| human | evolution | disease | computer |
| genome | evolutionary | host | models |
| dna | species | bacteria | information |
| genetic | organisms | diseases | data |
| genes | life | resistance | computers |
| sequence | origin | bacterial | system |
| gene | biology | new | network |
| molecular | groups | strains | systems |
| sequencing | phylogenetic | control | model |
| map | living | infectious | parallel |
| information | diversity | malaria | methods |
| genetics | group | parasite | networks |
| mapping | new | parasites | software |
| project | two | united | new |
| sequences | common | tuberculosis | simulations |

erc European Research Council

# LDA Topic Modeling

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

| | | |
|---|---|---|
| 0 | 2,5 | ¡oh hay siempre agua canta visto he suena dijo ¡qué cantar ¡y tienen buena hora ve otro flores cuatro |
| 1 | 2,5 | alma caminos españa buen eres guerra paz melancolía fuerte ¿quién fue castilla ríos quiere caballero guarda rincón manchego mujeres |
| 2 | 2,5 | sol viejo sueño jardín agua sed pasa ceño brilla sombras guiomar noble (a primavera tristeza memoria arco arena boca |
| 3 | 2,5 | sombra luna noche sueños yo voz vino dulce humilde estrellas fiesta espada calle quimera señora florida fiebre soy fantasma |
| 4 | 2,5 | ser sino mundo poeta sí pensamiento pensar conciencia otro puede pretende cada metafísica propia sea lógica real decir lírica |
| 5 | 2,5 | verde río flor alto sierra azul campo vi monte piedra nadie nube santa vii tren encina roca romero encinar |
| 6 | 2,5 | poeta mairena tiempo ha arte recuerdo verso barroco rima imágenes conceptos función espíritu tanto estrofa español queda artista versos |
| 7 | 2,5 | tierra campo hacia campos duero montes cielo verdes oro nieve soria luz ramas fría río castilla álamos sol grises |
| 8 | 2,5 | fuente agua tarde clara vieja piedra yo triste aire verano pena cristal alegría historia fue amores labios silencio copla |
| 9 | 2,5 | día vida tiempo don plaza olivares maestro mal lluvia toda mil infantil paso mancha negros cuerpo tic-tic divino quisiera |
| 10 | 2,5 | corazón mar dios ha señor dice niño hizo aguarda hace espera fe esperanza ilusión cabeza mares vida ¡ay gota |

erc European Research Council

# TM en los sonetos del Siglo de Oro

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Tipos de *topics*:
  - Temas tradicionales.
  - Rimas.
  - Relaciones simbólicas.
  - Ruido.

--

Navarro Colorado, B. (2015) "A computational linguistic approach..."

erc European Research Council

# TM en los sonetos del Siglo de Oro

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Tipos de *topics*:
  - Temas tradicionales.

| Topic | Tema |
|---|---|
| amor fuerza desdén arco niño cruel ciego flecha fuego ingrato sospecha | Amor no correspondido |
| hoy yace sepulcro fénix mármol polvo ceniza ayer guarda muerta piedad cadáver | Funeral, panegírico. |
| españa rey sangre roma imperio grande baña valor extraña reino carlos hazaña engaña saña bárbaro | Caída imperio (ruinas). |

erc European Research Council

# TM en los sonetos del Siglo de Oro

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Tipos de *topics*:
  - Rimas.

| Topic | Tema |
|---|---|
| cielo suelo velo vuelo celo alto consuelo recelo hielo mortal desvelo acá muestra desconsuelo tierra allá cubre duelo eterna | - |
| gloria memoria victoria historia eterna mayor inmortal triunfo será honor | - |
| olvido perdido sentido vencido rendido atrevido querido ofendido nacido oído conocido tenido escondido vestido agradecido merecido venido esclarecido podido | . |

erc European Research Council

# TM en los sonetos del Siglo de Oro

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Tipos de *topics*:
  - Relaciones simbólicas.

| Topic | Símbolo |
|---|---|
| **río** agua fuente frío corriente **cristal** desvarío tajo curso aguas betis brío pasa corre ninfas albedrío estío ribera margen | agua – cristal (Petrarca) |
| hermosa rosa blanco **flor** color labios rosas **cristal** blanca púrpura **perlas** rostro boca **nieve** abril frente **clavel** beldad risa | Descripción dama |

erc European Research Council

# Aplicar Topic Modeling

- Entrada: corpus plano.
- Herramientas:
  - MALLET:
    - http://mallet.cs.umass.edu/
  - Topic Modeling Tool (para corpus en inglés):
    - https://code.google.com/archive/p/topic-modeling-tool/
  - Gensim para Python
    - https://radimrehurek.com/gensim/
  - R
    - https://cran.r-project.org/web/packages/topicmodels/index.html

# Aplicar Topic Modeling

- Proceso:
  - Cargar corpus (creación matriz)
  - Extraer *topics*
- Con MALLET:
  - http://programminghistorian.org/lessons/topic-modeling-and-mallet
  - http://www.dlsi.ua.es/~borja/riilua/6.TopicModeling_v02.pdf
- Con R:
  - Jockers, M.L., (2014) *Text Analysis with R for Students of Literature*, págs. 135 y ss.
    - http://link.springer.com/10.1007/978-3-319-03164-4

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

# Conclusiones

- Modelos Semánticos Vectoriales permiten análisis semánticos a gran escala.
  - Semántica distribucional
- LDA Topic Modeling caso concreto de Modelo Semántico Vectorial.
  - Extrae *topics* de amplios corpus.
    - Cada *topic* es un conjunto de palabras relacionadas
  - Análisis de *topics* permite detectar aspectos semánticos recurrentes.

erc European Research Council

# ¿Dudas?

- …

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

erc European Research Council

# Práctica 2

- Análisis métrico y semántico del corpus con interfaz web.
  - http://goldenage.cervantesvirtual.com/
  - Análisis métrico: preguntas concretas para familiarizarse con la interfaz.
  - Análisis métrico-semántico (desarrollo):
    - Seleccionar un topic
    - Buscar sus patrones métricos característicos (si los tiene) y analizarlos.

# Bibliografía


- Blei, D.M., (2012) "Probabilistic Topic Models", *Communications of the ACM*, 55 (4), págs. 77–84.
- Firth, J. R. (1957), *Papers in Linguistics. 1934-1951*. Oxford University Press.
- Harris, Z. (1951), *Structural Linguistics*. University of Chicago Press, Chicago.
- Jockers, M (2014) *Text Analysis with R for Students of Literature*, Springer.
- McCallum. A. (2002) "Mallet: A machine learning for language toolkit". http://mallet.cs.umass.edu
- Navarro Colorado, B. (2014) "Recursos Informáticos para la investigación literaria". *Máster en Estudios Literarios* (MAESL), Universidad de Alicante. www.dlsi.ua.es/~borja/riilua/
- Navarro Colorado, B. (2015) "A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects", *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, págs. 105–113, Denver (Colorado), June 4, 2015.
- Navarro Colorado, B. (2016) "Hacia un análisis distante distante del endecasílabo áureo" *Rhythmica*, 14.



Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES



erc European Research Council

# Gracias

Contacto:

Borja Navarro Colorado
Dto. Lenguajes y Sistemas Informáticos
Universidad de Alicante
borja@dlsi.ua.es
@bncolorado