Hércules challenge: Extracción de tópicos de ROs

Jose Emilio Labra Gayo
Alejandro González Hevia

Universidad de Oviedo

Hercules Challenge

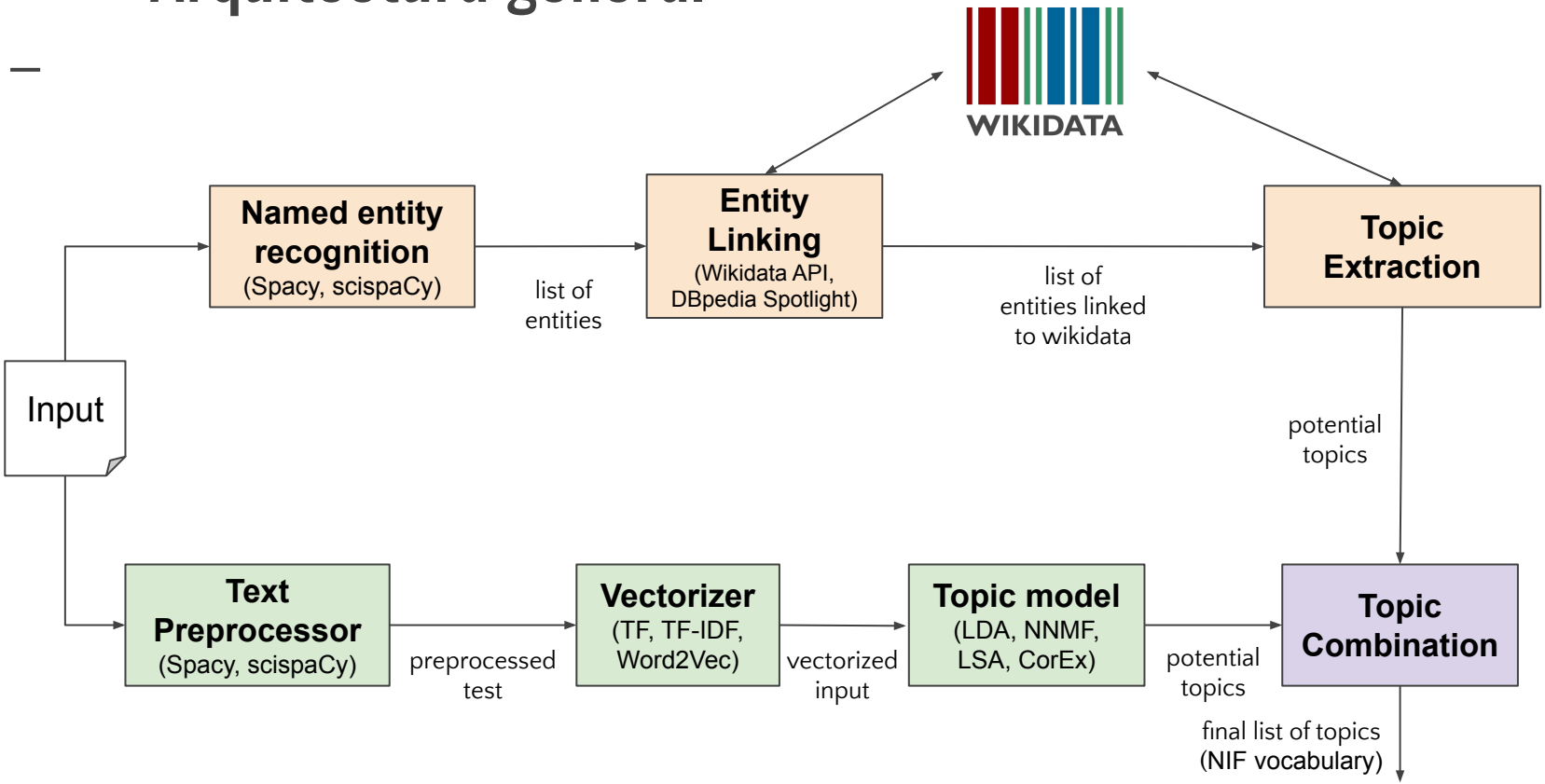
- Parte de la evaluación del lote I del proyecto EDMA (Enriquecimiento de Datos y Métodos de Análisis)
- Extracción de tópicos de distintos Research Objects
- 1 track por cada tipo:
  - Publicaciones científicas
  - Protocolos experimentales
  - Repositorios de código
- Presentación de resultados en formato Open Linked Data
- Tópicos enlazados a ontologías

Propuesta realizada

Propuesta realizada por Izertis-WESO

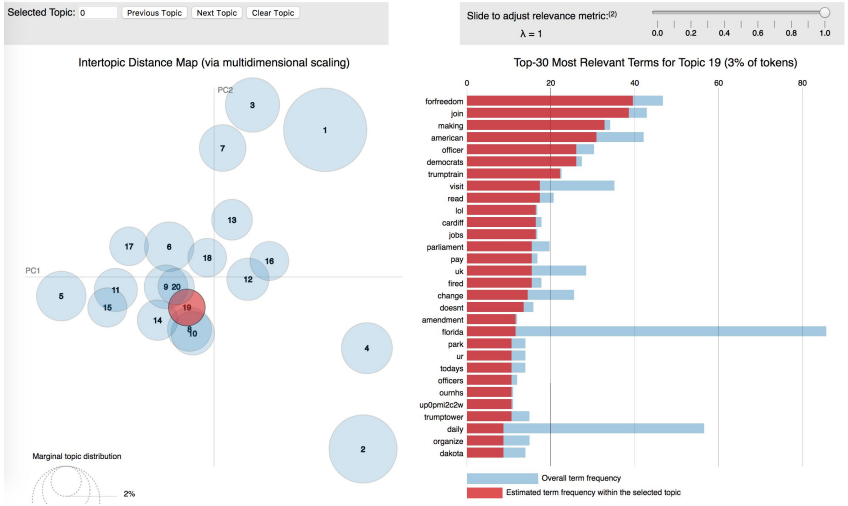
- Más información en http://edma-challenge.compute.weso.network/
- Incluye enlace a paper con resultados
- Demo interactiva
- Resultados
- Repositorio (público)
    - Código fuente del paper y librerías comunes
    - https://github.com/weso-edma/hercules-challenge-common
    - Instrucciones para reproducir mediante Jupyter Notebook

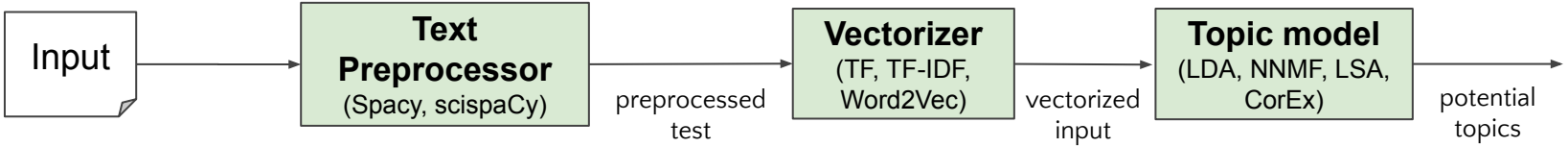
Arquitectura general



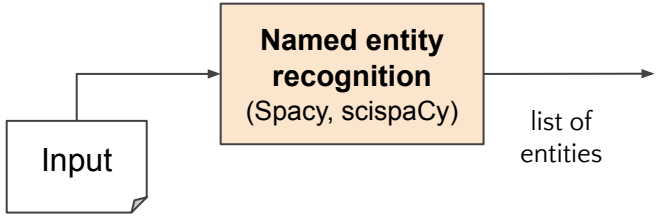
Extracción de tópicos “clásica”

- Preprocesado de texto
  - Tokenización
  - Lematización
  - Eliminar palabras vacías
- Vectorización de los tokens
- Aplicación de modelos de extracción de tópicos





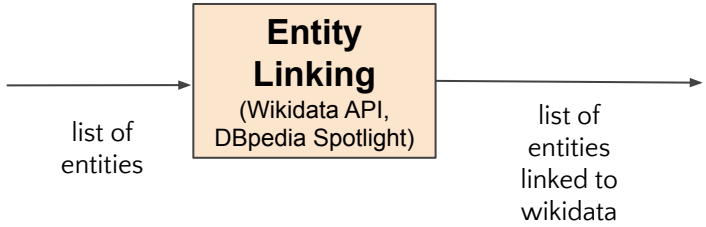
Reconocimiento de entidades



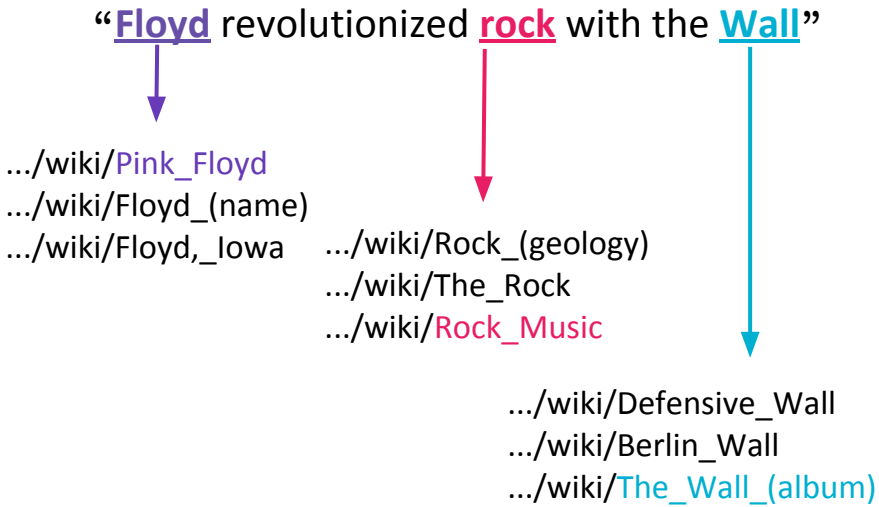
- Uso de SpaCy y ScispaCy
- Agrupación de entidades por tipo
- Diferentes modelos a evaluar
- Parámetro de mínimo número de ocurrencias
  - Ajustado para cada track

Introduction The fungus Alternaria NORP brassicicola causes black spot disease and is an economically important seed-borne pathogen of Brassicaceae GPE species. This necrotrophic ORG fungus strongly depends on seed transmission process for its long-term survival and dispersal (van den Bosch et al., 2010 DATE). However, fungal and plant factors that impact seed transmission are still poorly described. Such knowledge is crucial to propose strategies for improving the seed health, which remains a major issue for seed companies. Recent studies conducted with the Arabidopsis ORG thaliana / A. brassicicola pathosystem PERSON showed that the level of susceptibility of the fungus to water stress strongly influenced its seed transmission ability. For instance, two CARDINAL osmosensitive fungal mutants, defective for the class III Histidine ORG kinase (HK) AbNik1 ORG (Pochon et al., 2013 DATE) and the MAP ORG kinase AbHog1 PERSON (unpublished result), respectively, were highly jeopardized in their ability to colonize seeds. Consistently, Iacomi-Vasilescu et al ORG . (2008 DATE) had previously reported that A. brassicicola PERSON spontaneous phenylpyrrole resistant mutants with non-functional class III HK ORG were found to be strongly impaired in their ability to infect radish seeds in field conditions, indicating that a functional high osmolarity pathway is required for efficient infection of seeds. Pochon PERSON et al. (2013 DATE) also

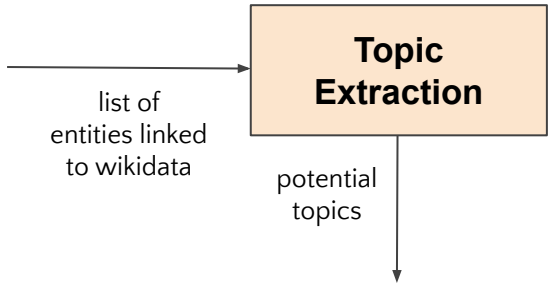
Entity linking



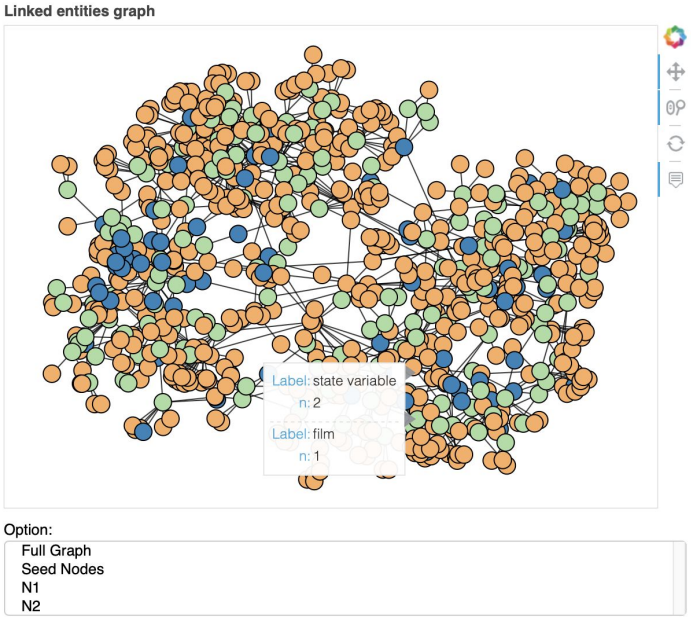
- Desambiguación de entidades y linkeado a ontologías
- Varias opciones barajadas
  - DBPedia Spotlight
  - Uso de la API de Wikidata
  - Herramientas de linkeado automático a Wikidata
- SpaCy proporcionará soporte a Wikidata en la versión 3.0!!
  - Entrenamiento de modelo con datos de Wikipedia.
  - Más información: https://medium.com/@mgalkin/spacy-irl-2019-and-wikidata-based-ner-64a799c17823



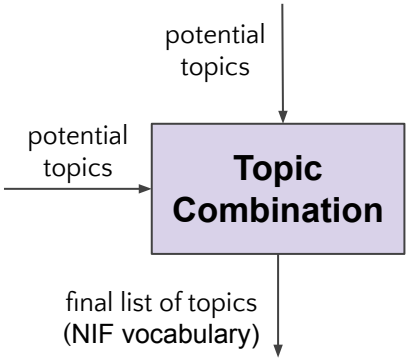
Extracción de tópicos



- Partimos de las entidades reconocidas previamente.
- Expansión de un conjunto de propiedades preseleccionadas:
  - P279, P910, P2579…
- Selección del subgrafo conectado más grande.
- Cálculo de entidades más relevantes del grafo
  - Information centrality
  - Eigenvector centrality
  - Betweenness
  - …



Combinación de tópicos



- En este punto entran 2 listas de tópicos de cada “sub-pipeline”.
- Cada tópico tiene una puntuación de confianza del sistema que lo produjo.
- Varias opciones para la selección de tópicos finales:
  - Ajuste manual de pesos
  - Sistema de aprendizaje automático
    - Requiere más datos
    - Mejor opción para el sistema final

Publicaciones

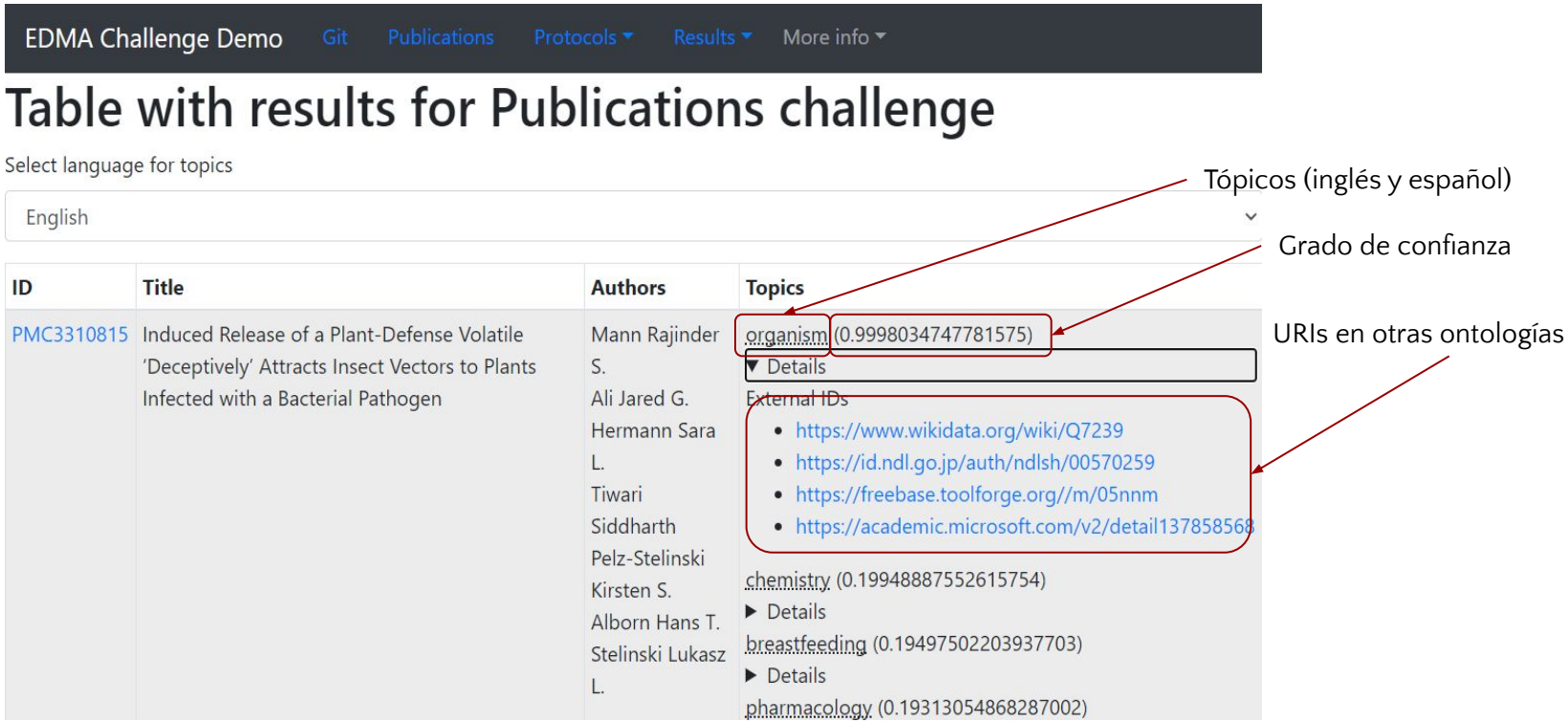
- 125 artículos de Agricultura de PMC.

- Particularidades
  - Características de entrada
    - Título
    - Abstract
    - Contenido
  - Fase extra para asociar tópicos por autor.

- Repositorio git (público):

https://github.com/weso-edma/hercules-challenge-publications

Publicaciones (Resultados)



Protocolos

- 100 protocolos experimentales de Bio-Protocol.

- Particularidades
  - Características de entrada
    - Título
    - Abstract
    - Lista de materiales y equipo
    - Procedimiento
  - Textos más cortos y con “pasos” sueltos sin conexión.

- Repositorio git (público): https://github.com/weso-edma/hercules-challenge-protocols

Protocolos (Generación de resúmenes)

- Se utiliza el abstract como baseline
- 2 tipos de resúmenes generados
  - Resúmenes abstractivos
    - Permite reusar texto existente y generar texto nuevo
    - Modelos BART pre-entrenados y ajustados a este dataset con un subconjunto aislado de protocolos
      - Facebook BART-cnn
      - Distilbart-cnn
      - Distilbart-xsum
  - Resúmenes extractivos
    - Identificar secciones importantes del texto a resumir
    - Técnica basada en TF

Protocolos (Resultados)

EDMA Challenge Demo | Git | Publications | Protocols | Results | More info

Protocols - topic extraction challenge

Select language for topics

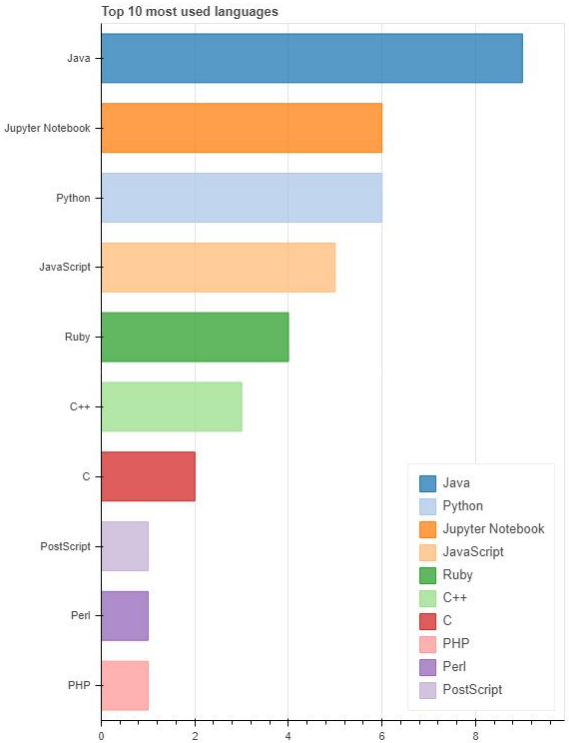
English

Title	Authors	Topics
Scratch Wound Healing Assay	Yanling Chen	software (0.19620991253644315) ▼ Details External IDs • https://www.wikidata.org/wiki/Q7397 • https://id.ndl.go.jp/auth/ndlsh/00684642 • https://freebase.toolforge.org//m/01mf0 • https://www.jstor.org/topic/computer-software • https://meshb.nlm.nih.gov/record/ui?ui=D012984 • https://academic.microsoft.com/v2/detail2777904410 • http://vocabularies.unesco.org/thesaurus/concept6081 chemistry (0.19507246376811593) ▶ Details science (0.19495944380069524) ▶ Details research (0.19462116830537884)

Repositorios Git

- 50 repositorios extraídos de GitHub.
- Características de entrada:
  - Descripción del repositorio
  - README (si lo tienen)
  - Mensajes en commits
  - Lista de nombres de los ficheros fuente
- Especial consideración a la desambiguación
  - DBpedia Spotlight
- Repositorio git (público)

https://github.com/weso-edma/hercules-challenge-git



Repositorios Git (Resultados)

memetools	tkuhn	Java (182983) Shell (3020)	catalogue (0.28273809523809523) ▶ Details communication medium (0.2714285714285714) ▶ Details publication (0.26988636363636365) ▶ Details mass media (0.26912181303116145) ▼ Details External IDs • https://www.wikidata.org/wiki/Q11033 • https://id.ndl.go.jp/auth/ndlsh/00567519 • https://www.jstor.org/topic/mass-media • http://vocabularies.unesco.org/thesaurus/concept487 • https://meshb.nlm.nih.gov/record/ui?ui=D008402 • https://academic.microsoft.com/v2/detail558299567 written work (0.268361581920904) ▶ Details work (0.26536312849162014)

Formato de salida

- Varias alternativas
  - SKOS
  - LEMON
  - Otros vocabularios…
- Formato de salida
  - Linked Open Data
  - Vocabulario NIF
- Resultados en JSON, JSON-LD, TTL, CSV

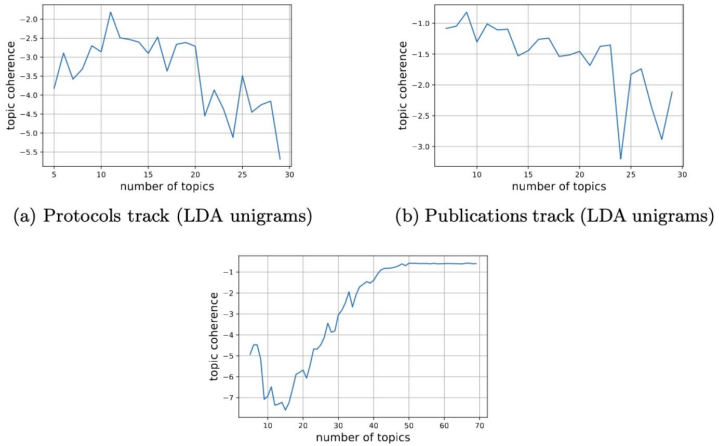
```
@prefix edma: <http://edma.org/challenge/> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif: <https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>
.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

edma:216602979 nif:isString "LIkelihood Ratio Interpretation ..." ;
    nif:predominantLanguage "en" ;
    nif:sourceURL <https://www.github.com/cmungall/LIRICAL> ;
    nif:topic [ a nif:annotation ;
            rdfs:label "statistics"@en, "estadística"@es ;
            rdfs:comment "study of ..."@en, "estudio de..."@es ;
            itsrdf:taIdentRef <http://id.nlm.nih.gov/mesh/E05.318.740>,
                              <http://id.nlm.nih.gov/mesh/H01.548.832>,
                              <https://www.wikidata.org/wikidata/Q12483> ;
            nif:confidence 2.094723e-01
     ], . . .
```

Evaluación (I)

- Métricas no supervisadas
  - Evaluar la distribución de tópicos obtenida “estadísticamente”.
  - P.ej. Topic coherence, perplexity.

- Métricas supervisadas
  - Comparar tópicos obtenidos con unos tópicos “objetivo”.
  - Muchas métricas existentes (exhaustividad, puntuación f1…)
    - No se tienen en cuenta matches “similares semánticamente”.
    - Solución propuesta: Métricas de similitud semántica.



(a) Protocols track (LDA unigrams)

(b) Publications track (LDA unigrams)

(c) Git track (NMF bigrams)

Fig. 4: Topic coherence evolution with number of topics

Evaluación (II)

Publicaciones

- Ground truth: Categorías filtradas de EuropePMC API
- Mejores resultados: Unigramas, modelo LDA (9 componentes), λ=0.8

Repositorios git

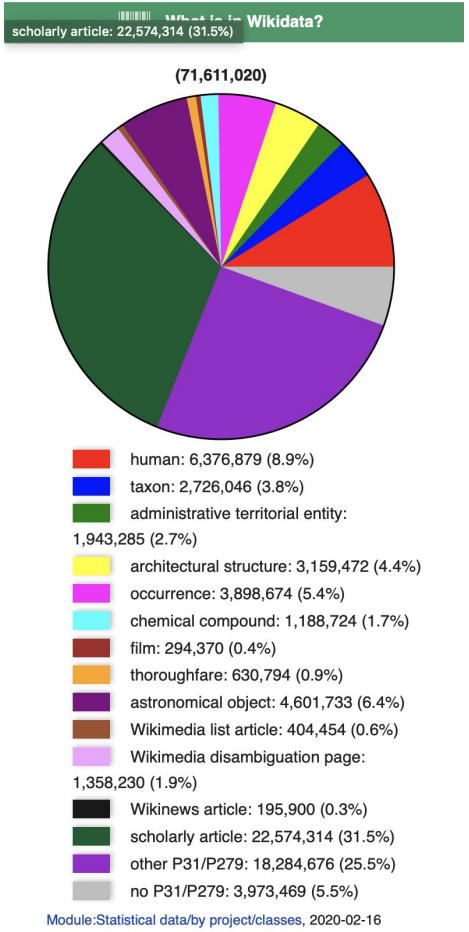
- Ground truth: Anotación manual de tópicos
- Datasets anotados manualmente (3 personas)
- Mejores resultados: Bigramas, modelo NMF con 50 componentes, λ=0,9

Protocolos

- Ground truth: subjects de los protocolos con filtrado manual
- Mejores resultados: Unigramas, modelo LDA con 11 componentes y λ=0,8

Trabajo futuro

- Mejorar técnicas de enlazado a entidades.
  - Mejoras de “desambiguación colectiva”.
  - Uso de Spacy+Wikidata.
- Rendimiento
  - Indexado de ontologías en sistemas tipo ElasticSearch.
  - Creación de subconjuntos de Wikidata.
- Mejoras de relevancia
  - Algoritmos de centralidad de grafos.
  - Implementar propuestas de Hulpus et al.



https://www.wikidata.org/wiki/Wikidata:Statistics

Demo presentada

Información disponible en:
http://edma-challenge.compute.weso.network/

- Paper (PDF)
- Demo interactiva
- Tablas de resultados
- Enlaces a repositorios (abiertos) con código fuente
- Instrucciones para reproducir mediante Jupyter Notebook

Referencias

- [Hulpus et al] **Unsupervised graph-based topic labelling using dbpedia.**
- [Budanitsky, A., Hirst, G.] Semantic distance in Wordnet: An experimental, application-oriented evaluation of five measures.
- [Chen et al] A survey on the use of topic models when mining software repositories.
- [JL. Martinez et al] **Information extraction meets the semantic web: a survey.**
- [Hellman S., Lehmann J.] Integrating NLP using linked data.
- [Chabchoub et al] Ficlone: improving dbpedia spotlight using named entity recognition and collective disambiguation.
- [Bhatia el at] Automatic labelling of topics with neural embeddings.

Fin de la presentación