

Máster en Ciencia de Datos

PROYECTO DE INVESTIGACIÓN PARA LA CIENCIA DE DATOS

1. Definición del objetivo y aportación

Lara Quijano Sánchez



Universidad Autónoma
de Madrid

Skills a trabajar y adquirir

- ❑ Identificar y seleccionar los métodos, técnicas y herramientas de ciencia de datos más apropiados para la resolución de los problemas abordados.
- ❑ Participar liderar proyectos de innovación e investigación en el ámbito de la ciencia de datos
- ❑ Identificar y manejar adecuadamente fuentes de información
- ❑ Capacidad de trabajar en equipo
- ❑ Organización de tiempo
- ❑ Reparto de tareas
- ❑ Diseñar, desarrollar y transferir los resultados de proyectos de investigación utilizando una metodología adecuada
- ❑ Redactar memorias y artículos científicos y técnicos

Un paper debe tener las siguientes secciones

- ❑ **Introducción:** Incluye motivación, necesidad del estudio, aportación científica, novedad que se incluye
- ❑ **Revisión del estado del arte:** qué se ha hecho hasta ahora, carencias identificadas
- ❑ **Metodología:** Descripción de algoritmos a usar, datasets disponibles o creados, limpieza y tratamiento, variables a utilizar en el modelo, modelo desarrollado
- ❑ **Resultados:** experimentos llevados a cabo, descripción, resultados, comparación con el estado del arte
- ❑ **Conclusiones:** insights, reusabilidad, interés, transferencia, limitaciones y trabajo futuro

Motivación

- ❑ No empezar desde 0: qué puedo reusar?
- ❑ Ver si merece la pena: ya estaba hecho? Que puedo mejorar?
- ❑ Ver si es realizable: tengo datos? Tengo recursos?

Definición de la tarea a Realizar

¿Qué quiero aprender/ inferir?

Clasificar, predecir, agrupar?

¿Qué se ha hecho ya al respecto?:

Pasos

1. Buscar bibliografía
2. Identificar similitudes en
 1. Dominio y datasets
 2. Atributos de entrada
 3. Técnicas de modelado
3. Identificar novedad y aportación

1. Similitudes en dominio

- ❑ Ya se ha hecho un proyecto igual/similar? En cuyo caso se me ocurren incluir algo más o diseñar un modelo distinto para mejorar lo hecho. Ver pasos 2 y 3
- ❑ Se predice lo mismo pero usando otros datos o fuentes ej:
 - ❑ Influencers en tweets vs Instagram/ fake news en twitter/periódicos, diferentes técnicas, sería lo mismo las conclusiones, se tienen otros datos? Longitud de texto? Qué es diferente?
 - ❑ Idioma, solo hay técnicas en inglés en el caso de textos? Es lo mismo para otros idiomas
 - ❑ Dominio,
 - ❑ Ej: predicción de gustos usando diferentes fuentes, dominio turismo info en redes sociales diferentes, info de webs como yelp, tripadvisor, google maps, info de páginas concretas de turismo/restaurantes, estudios estadísticos, creación de app concreta de “rastreo”/gustos. De nuevo, diferentes fuentes aportan datos/inputs distintos? Relevantes? Deriva en mejoras o insights conclusiones nuevas?
 - ❑ Diferente dominio, Ej. Cálculo del tie strength usando info de redes sociales vs cálculo usando movimientos y transferencias bancarias para estudiar afinidad de personas vs clientes

DataSets Existentes

☐ Hay datasets existentes del dominio que quiero estudiar?

☐ Si:

☐ Son públicos?

☐ Si:

☐ Disponen de todos los datos de entrada que necesito? Sino puedo completarlo?

☐ Es reproducible? No faltan datos del dataset original? (ej twitter borra tweets)

☐ No:

☐ Puedo pedirselos a los autores? Licencia? Dinero?

☐ No:

☐ Puedo conseguir datos?

☐ Conozco fuentes/empresas/datos públicos?

☐ Puedo crear el mío propio usando scrapping u otras técnicas?

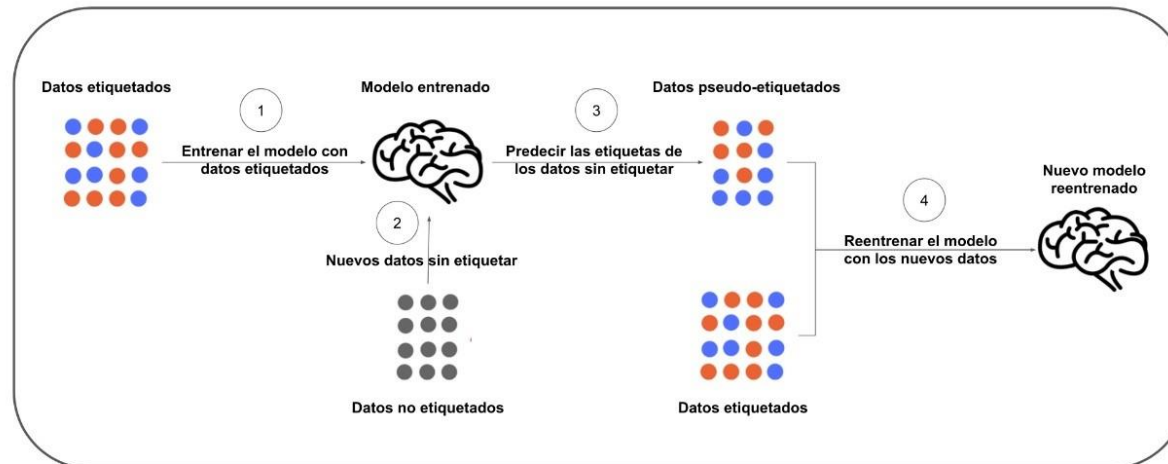
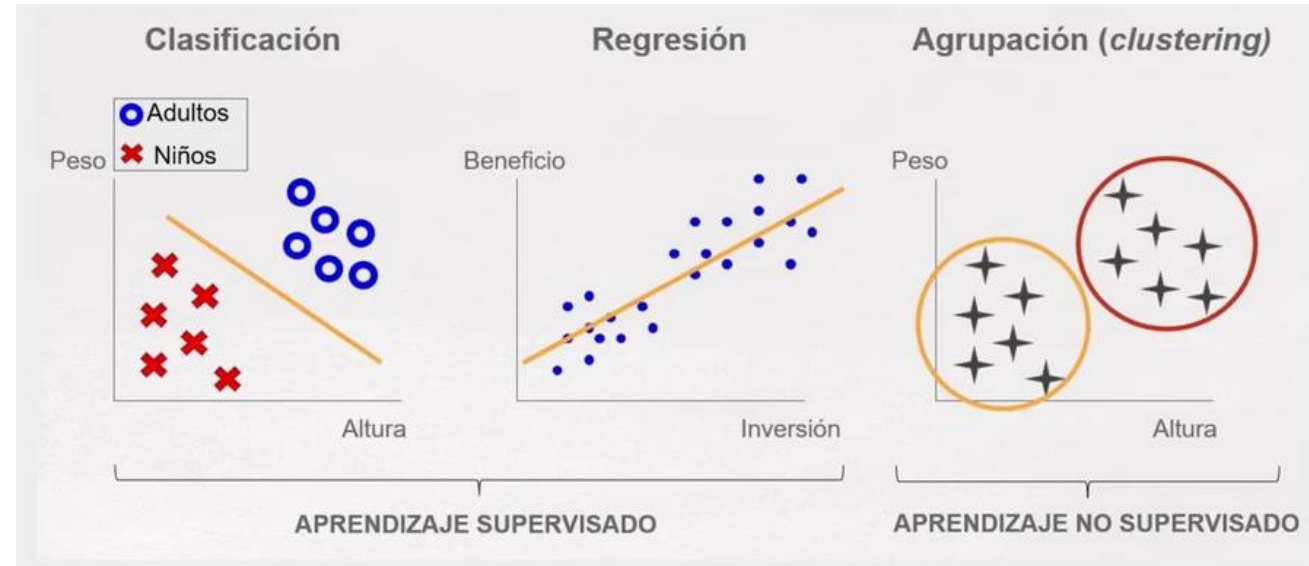
☐ Tengo tiempo?

☐ Será suficientemente grande y significativo? Estudiar sesgos y bias posibles

☐ Tengo la variable a predecir? U etiquetadores calificados (dinero para pagarlos, amigos? Etc)

DataSets Existentes

- Aprendizaje supervisado (todos los datos etiquetados)
- Aprendizaje no supervisado (ningún dato etiquetado)
- Aprendizaje semisupervisado (algunos datos etiquetados)



2. Similitudes entre atributos de entrada

- ☐ Estudiar cuales son las entradas/variables de los modelos hechos hasta ahora
- ☐ Voy a usar las mismas? Tengo acceso a ellas?
- ☐ Se me ocurren nuevas variables a estudiar? Introducir = Nuevo conocimiento? => estudiar por cada nueva aportación su impacto en el resultado? Mejora? Es significativo?
- ☐ Ejemplos:
 - ☐ En tweets, info del perfil? Incluir o no emojis? Incluir/separar info en #? Añadir información relevante al texto (pos tagging)
 - ☐ Limpieza tratamiento de los datos nueva
 - ☐ Incluir variables exógenas, datos de la población, lugar donde ha ocurrido algo?

Método de selección de variables

- ☐ ¿Qué métodos de selección de variables se han usado?
- ☐ ¿Se ha estudiado la relevancia de las variables?
- ☐ ¿Debo de usar yo un método de limpieza y selección?
 - ☐ Problemas con la explicabilidad de los modelos, algunos modelos son muy buenos, pero no se que hay por dentro. ¿Cómo puedo sacar insights?

3. Similitudes en los modelos/algoritmos utilizados

- ☐ ¿Qué técnicas se han usado?
 - ☐ Regresiones lineales, Lasso, Reglas, SVM, Random Forest, Naïve Bayes, Redes Neuronales
 - ☐ Métodos ensemble
 - ☐ ¿Qué resultados han obtenido?
 - ☐ Es reproducible para comparar?
 - ☐ Está disponible el código o se detalla la configuración de parámetros
 - ☐ El dataset utilizado está disponible
 - ☐ F1, AUC, precisión, recall, etc. ¿ Qué indicadores hay?
- ☐ Se me ocurre una nueva? Que no se haya probado?

Identificar la novedad/aportación de mi propuesta

- ☐ Necesito estudiar todos los puntos anteriores
- ☐ Para ello debo buscar en la bibliografía
 - ☐ Deben ser artículos científicos
 - ☐ No blogs (no han pasado proceso de validación, peer review), work in progress etc
 - ☐ Ej: <https://medium.com/@adriensieg/text-similarities-da019229c894>

Búsqueda de bibliografía

- ❑ Puedo empezar una búsqueda general con Google Scholar

- ❑ <https://scholar.google.es/>

- ❑ Palabras relacionadas con mi propuesta

- ❑ Empezar por el objetivo (keywords) y el dominio, que por técnicas a usar ya que estas saldrán en los otros

- ❑ Con esto tengo un barrido inicial que me puede servir para encontrar términos a buscar

- ❑ Puedo filtrar por años (seleccionar los 3 últimos por ejemplo)

- ❑ Leer abstract de los que parezcan más relevantes a mi tema e identificar keywords para buscar más en profundidad

Búsqueda de bibliografía

☐ Google Scholar

Google Scholar

text similarity



Artículos

Aproximadamente 43.100 resultados (0,07 s)

Cualquier momento

Desde 2020

Desde 2019

Desde 2016

Intervalo específico...

Ordenar por relevancia

Ordenar por fecha

Cualquier idioma

Buscar sólo páginas en español

☒ incluir patentes

☒ incluir citas

☒ Crear alerta

[HTML] [Learning short-text semantic similarity with word embeddings and external knowledge sources](#)

[HT Nguyen](#), [PH Duong](#), [E Cambria](#) - Knowledge-Based Systems, 2019 - Elsevier

We present a novel method based on interdependent representations of short texts for determining their degree of semantic **similarity**. The method represents each short **text** as two dense vectors: the former is built using the word-to-word **similarity** based on pre-trained ...

☆ Citado por 15 Artículos relacionados Las 2 versiones

[HTML] [sciencedirect.com](#)

Texto completo para UC3M

[Bridging the gap between relevance matching and semantic matching for short text similarity modeling](#)

[J Rao](#), [L Liu](#), [Y Tay](#), [W Yang](#), [P Shi](#), [J Lin](#) - Proceedings of the 2019 ..., 2019 - aclweb.org

A core problem of information retrieval (IR) is relevance matching, which is to rank documents by relevance to a user's query. On the other hand, many NLP problems, such as question answering and paraphrase identification, can be considered variants of semantic ...

☆ Citado por 12 Artículos relacionados Las 5 versiones

[PDF] [aclweb.org](#)

[Text as policy: Measuring policy similarity through bill text reuse](#)

[F Linder](#), [B Desmarais](#), [M Burgess](#)... - Policy Studies ..., 2020 - Wiley Online Library

The identification of substantively similar policy proposals in legislation is important to scholars of public policy and legislative politics. Manual approaches are prohibitively costly in constructing datasets that accurately represent policymaking across policy domains ...

☆ Citado por 29 Artículos relacionados Las 4 versiones

[PDF] [wiley.com](#)

Full View

Búsqueda de bibliografía

- ❑ Con los keywords identificados realizar una búsqueda en profundidad en webs científicas de recopilación de artículos como pueden ser
 - ❑ Scopus. Es una base de datos bibliográfica de resúmenes y citas de artículos de revistas científicas. Cubre aproximadamente 18.000 títulos de más de 5000 editores internacionales, incluyendo la cobertura de 16.500 revistas revisadas por pares de las áreas de ciencias, tecnología, medicina y ciencias sociales, incluyendo artes y humanidades. Está editada por Elsevier y es accesible en la Web para los subscriptores. Las búsquedas en Scopus incorporan búsquedas de páginas web científicas mediante Scirus, también de Elsevier, y bases de datos de patentes.
 - ❑ Web of Science (WOS). es un servicio en línea de información científica, suministrado por Clarivate Analytics, integrado en ISI Web of Knowledge, WoK. Facilita el acceso a un conjunto de bases de datos en las que aparecen citas de artículos de revistas científicas, libros y otros tipos de material impreso que abarcan todos los campos del conocimiento académico.

Búsqueda de bibliografía

- ❑ Un artículo se identifica por su
 - ❑ Título
 - ❑ Abstract
 - ❑ Keywords
- ❑ Luego haremos búsquedas en esas 3 características
- ❑ Filtramos años recientes
- ❑ Realizamos la intersección de ambos resultados
- ❑ Anotamos n° de resultados
- ❑ Prefiltramos leyendo título y abstract para comprobar son de nuestro interés
- ❑ Seleccionamos los que sí sean relevantes, anotamos el n° y procedemos a leer los artículos completos
- ❑ Realizamos informe numérico de todo
 - ❑ Inicial
 - ❑ Pre-filtrado
 - ❑ Pos-lectura

Búsqueda de bibliografía: Ejemplo

Scopus

```
(
TITLE-ABS-KEY("recommender*")
OR TITLE-ABS-KEY("recommendation system*")
OR TITLE-ABS-KEY("recommendation service*")
OR TITLE-ABS-KEY("recommendation approach*")
OR TITLE-ABS-KEY("recommendation model*")
OR TITLE-ABS-KEY("recommendation method*")
OR TITLE-ABS-KEY("recommendation algorithm*")
OR TITLE-ABS-KEY("recommendation application*")
OR TITLE-ABS-KEY("recommendation engine*")
OR TITLE-ABS-KEY("recommendation framework*")
OR TITLE-ABS-KEY("collaborative filtering")
)
AND
(
TITLE-ABS-KEY("smart cit*")
OR TITLE-ABS-KEY("smart building*")
OR TITLE-ABS-KEY("smart home*")
OR TITLE-ABS-KEY("internet of things")
OR TITLE-ABS-KEY("iot")
OR TITLE-ABS-KEY("smart governance")
OR TITLE-ABS-KEY("smart economy")
OR TITLE-ABS-KEY("smart people")
OR TITLE-ABS-KEY("smart living")
OR TITLE-ABS-KEY("smart mobility")
OR TITLE-ABS-KEY("smart environment")
OR TITLE-ABS-KEY("smart health*")
OR TITLE-ABS-KEY("smart education")
OR TITLE-ABS-KEY("smart tourism")
)
```

WOS

```
(
TI="recommender*" OR TS="recommender*"
OR TI="recommendation system*" OR TS="recommendation system*"
OR TI="recommendation service*" OR TS="recommendation service*"
OR TI="recommendation approach*" OR TS="recommendation approach*"
OR TI="recommendation model*" OR TS="recommendation model*"
OR TI="recommendation method*" OR TS="recommendation method*"
OR TI="recommendation algorithm*" OR TS="recommendation algorithm*"
OR TI="recommendation application*" OR TS="recommendation
application*"
OR TI="recommendation engine*" OR TS="recommendation engine*"
OR TI="recommendation framework*" OR TS="recommendation
framework*"
OR TI="collaborative filtering" OR TS="collaborative filtering"
)
AND
(
TS="smart cit*" OR TI="smart cit*"
OR TS="smart building*" OR TI="smart building*"
OR TS="smart home*" OR TI="smart home*"
OR TS="internet of things" OR TI="internet of things"
OR TS="iot" OR TI="iot"
OR TS="smart governance" OR TI="smart governance"
OR TS="smart economy" OR TI="smart economy"
OR TS="smart people" OR TI="smart people"
OR TS="smart living" OR TI="smart living"
OR TS="smart mobility" OR TI="smart mobility"
OR TS="smart environment" OR TI="smart environment"
OR TS="smart health*" OR TI="smart health*"
OR TS="smart education" OR TI="smart education"
OR TS="smart tourism" OR TI="smart tourism"
)
```

https://www.scopus.com/search/form.uri?display=advanced&origin=searchbasic&txGid=05784cb973e5b43fedd3b896e0b238b4



Brought to you by [Universidad Carlos III de Madrid - Biblioteca](#)



[Search](#) [Sources](#) [Lists](#) [SciVal](#)



Advanced search

[Compare sources](#)

☐ Documents ☐ Authors ☐ Affiliations [Advanced](#)

[Search tips](#)

Enter query string

[Outline query](#) [Add Author name / Affiliation](#)

[Search](#)

ALL("Cognitive architectures") AND AUTHOR-NAME(smith)
TITLE-ABS-KEY(*somatic complaint wom?n) AND PUBYEAR AFT 1993
SRCTITLE(*field ornith*) AND VOLUME(75) AND ISSUE(1) AND PAGES(53-66)




Operators

[AND](#) +
[OR](#) +
[AND NOT](#) +
[PRE/](#) +
[W/](#) +

Field codes

[Textual Content](#) ✓
[Affiliations](#) ✓
[Authors](#) ✓
[Biological Entities](#) ✓
[Chemical Entities](#)

<https://www.recursoescientificos.fecyt.es/>
https://apps.webofknowledge.com/UA_AdvancedSearch_input.do?SID=E5177I68aGrGWNwED8t&product=UA&search_mode=AdvancedSearch




FUNDACIÓN ESPAÑOLA
PARA LA CIENCIA
Y LA TECNOLOGÍA

Web of ScienceInCitesJournal Citation ReportsEssential Science IndicatorsEndNotePublonsKopernioMaster Journal List

Inicio sesión ▼Ayuda ▼Español ▼

Web of Science



Herramientas ▼Búsquedas y alertas ▼Historial de búsquedaLista de registros marcados

Web of Science realizará tareas de mantenimiento programadas del 1 de septiembre de 2020 a las 11:00 GMT al 1 de septiembre de 2020 a las 23:00 GMT.
Durante este período, el acceso puede ser intermitente. Disculpe las molestias.

Seleccionar una base de datos

Todas las bases de datos ▼

Búsqueda básicaBúsqueda de referencia citadaBúsqueda avanzada

Use etiquetas de campo, operadores booleanos, paréntesis y conjuntos de consultas para crear su consulta. Los resultados aparecerán en el historial de búsqueda situado en la parte inferior de la página. (Más información sobre la búsqueda avanzada)

Ejemplo: TS=(nanotub* AND carbon) NOT AU=Smalley RE
#1 NOT #2 [más ejemplos](#) | [ver el tutorial](#)

Buscar

Período de tiempo

Todos los años (1900 - 2020) ▼

Booleanos: AND, OR, NOT, SAME, NEAR

Etiquetas de campo:

TS= Tema

TI= Título

AU= Autor [\[Índice\]](#)


AI= Identificadores de autores

GP= Autoría conjunta [\[Índice\]](#)

ED= Editor

AB= Abstract

AK= Palabras clave de autor

KP= Keyword Plus 

SO= Nombre de publicación [\[Índice\]](#)

DO= DOI

PY= Año de publicación

AD= Dirección

SU= Área de investigación

IS= ISSN/ISBN

Tareas

- ❑ Vamos a reutilizar nuestros conocimientos de Ingeniería del Software
- ❑ Vamos a crear nuestro backlog de tareas a hacer
- ❑ Vamos a crear 1 sprint de 5 semanas
 - ❑ Para cubrir esta temática
 - ❑ Escribir un informe = Secciones 1 y 2 de un paper de investigación
 - ❑ Servirá también para el análogo en la redacción de un Proyecto
- ❑ Usaremos tecnologías recurrentes para
 - ❑ Visualizar y generar
 - ❑ Tarjetas de responsabilidad
 - ❑ Evolución
 - ❑ Estimación de tiempos
 - ❑ Informes

Tareas

1. Identificación de keywords

- ☐ Búsquedas auxiliares
 - ☐ Google scholar

2. Búsqueda bibliográfica en profundidad.

- ☐ Realización de queries en WOS y Scopus.
 - ☐ Ajustarse a últimos años
 - ☐ Prefiltrado leyendo títulos, abstracts de n° d papers a leer
 - ☐ Reparto de papers

☐ Lectura de informes/papers científicos relacionados.

- ☐ Divide y vencerás
 - ☐ Anotación de papers
 - ☐ Resumen
 - ☐ Incorporación a tabla de atributos a valorar en común

☐ Análisis ordenado de la bibliografía

- ☐ Lectura de resultados de compañeros
- ☐ Puesta en conjunto

3. Definición clara de la aportación del proyecto y de la novedad

- ☐ Motivación de necesidad
- ☐ Idea clara de pasos a seguir (metodología, datos, arquitectura básica)

4. Redacción

Semana del 8 al 14 Marzo

Semana del 8 al 14 Febrero

Semana del 15 al 28 Febrero

Semana del 8 al 14

Semana del 1 al 7 Marzo

Puede abrir campo de búsqueda. Requiere consulta con equipo

Estudio de la bibliografía

- ☐ Ejemplo de **excel común dinámico** que podemos rellenar, para no olvidar, releer, sirva para la redacción del estado del arte etc
- ☐ Subrayar en los papers las frases clave
- ☐ Al acabar de leer escribir resumen
 - ☐ Qué hacen?
 - ☐ Para qué necesidad?
 - ☐ Que datasets usan?
 - ☐ Que método usan?
 - ☐ Tienen experimentos?
 - ☐ Que resultado tienen? (escribir métrica)
 - ☐ Lista de papers que citan
 - ☐ Construir un árbol jerárquico
 - ☐ Si se menciona otro paper
 - ☐ Que puede tener mi compañero
 - ☐ Tener lista de qué títulos tiene cada compañero
 - ☐ Si en el mio dicen que es claramente mejor y explican porque, comentar con compi
 - ☐ Lectura más ligera

ID Paper	Año de publicación	Artículo survey/revisión de otros (S/N)	Propone n/usan dataset/corpus (S/N)	Origen del dataset (twitter, etc)	Es descargable (S/N)	link en caso de descargable	es aprendizaje supervisado/semi-supervisado?	Usan procesamiento del lenguaje natural (S/N) = preprocesado de textos	Enumerar técnicas NLP (bag of words, stemming, stopwords, tf-idf, etc)	Usan técnicas para reducir la dimensión de las features (S/N)	Enumerar técnicas (lasso, chisquare, las Vegas wrapper, etc)	Usan machine learning para clasificar? (S/N)	Enumerar técnicas (redes neuronales, tipo de red, svm, naive bayes, etc)	Detalles específicos de configuración (capas de red, espacio vectorial, etc)	Usan métricas para medir la calidad de sus resultados (S/N)	Enumerar técnicas (MAE, F1, precisión, recall, AUC, etc)	Resultado de su mejor método o predicción	Usan análisis de redes sociales o técnicas de grafos? (S/N)	Enumerar técnicas y objetivos	Otros, características especiales
----------	--------------------	---	-------------------------------------	-----------------------------------	----------------------	-----------------------------	--	--	--	---	--	--	--	--	---	--	---	---	-------------------------------	-----------------------------------

Realizaremos un estudio jerárquico

Redacción sección 2: Estado del arte

- ☐ Agruparemos por temática / dominio
 - ☐ Agruparemos artículos que hayan usado mismos datasets
 - ☐ Seleccionaremos de entre los diferentes datasets, métodos, etc los que tengan mejores resultados dentro de cada variación
 - ☐ Identificaremos datasets disponibles
 - ☐ Estudiaremos si es posible compararse con las metodologías que den resultados mejores ya que podamos implementar su solución o obtener sus datasets
 - ☐ Justificarlo
- ☐ Realizaremos tablas comparativas por artículos
 - ☐ Identificaremos atributos/inputs utilizados, listaremos cuales han reportado resultados mejores
 - ☐ Identificaremos métodos utilizados listaremos cuales han reportado resultados mejores
 - ☐ Así tendremos un valor claro de qué mejorar
 - ☐ qué métrica debemos superar para realizar una aportación?
- ☐ Incluiremos el resumen de los trabajos (versión abreviada)
 - ☐ Guardar para nosotros la versión larga

Al final: Seremos capaces de identificar dentro de nuestra propuesta **el objetivo y la aportación del proyecto**

Realización de un relato

Redacción sección 1: Introducción

☐ A día de hoy...

☐ Existe xxxx

☐ Ej. Sobrecarga de información, bulos, etc etc

☐ Motivados por la xxxxxx

☐ Que tiene estas carencias/ necesidades

☐ Hasta ahora se ha echo esto

☐ Resumen corto

☐ Que tiene estas carencias/ necesidades

☐ En este trabajo se pretende

☐ Xxx para suplir estas carencias

☐ Se realizará por medio de

☐ Idea general, arquitectura

☐ Por lo que las contribuciones serán

☐ Xxxxxx

☐ Que servirán para

☐ Retorno a la sociedad, a la empresa....

☐ Vender el Proyecto, porque va a ser útil

Vuestro Proyecto

Esquema sección 3: Metodología

☐ Debéis debatir esto para poder realizar:

- ☐ La propuesta de proyecto
- ☐ Siguiente sprint
- ☐ Organización del proyecto
- ☐ Del testeo
- ☐ De los resultados esperables

☐ Tras la búsqueda debéis ya tener claro:

☐ Estudio de datasets existentes

- ☐ Son reutilizables? Son gratis? Tienen los datos que necesitamos? hay que completarlos? que estudios los usan? que bias tienen? son suficientemente grandes? búsqueda además online y en bancos de datos.
- ☐ Conclusión: informe y justificación de qué datos se van a usar, ya existentes? reutilizar parte? crear nuevos?

☐ Estudio de atributos predictivos

- ☐ cuales se han usado? cuales son relevantes? que tratamientos se les hacen? se me ocurren nuevos? tengo/puedo conseguir datos para estudiar estos nuevos?
- ☐ Conclusión: informe y justificación de qué atributos se van a analizar y si es posible su obtención y estudio, identificación de posibles bias.

☐ Estudio de qué métodos y algoritmos predictivos que se han usado

- ☐ son eficientes? voy a usar alguno ya hecho? voy a proponer alguno nuevo? Cuales han reportado mejores resultados? Voy a combinar varios? Los papers aportan detalles de la configuración de sus experimentos? son replicables? aportan código? hay en github? puedo reutilizar parte?
- ☐ Conclusión: Informe sobre métodos usados hasta ahora, ordenados por eficiencia, y propuesta de método a estudiar e implementar.

Tecnologías



☐ Herramienta para el seguimiento colaborativo del proyecto y tareas:

☐ Ej. Jira, trello



☐ Gestor bibliográfico

☐ Ej. Mendeley

☐ Herramienta para editar el artículo, informes, proyectos de forma colaborativa

☐ Ej. Latex, overleaf



☐ Alojamiento del Proyecto, servidor común

☐ Ej. Bitbucket



☐ Gestor de versions

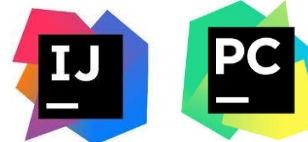
☐ Pushes y pulls organizados organizados etc

☐ Ej. Sourcetree, git, IntelliJ, Pycharm



☐ Plataforma de desarrollo (out of scope)

☐ Notebooks, IntelliJ, Pycharm, etc





←

→

🏠

🔒 https://laraqresearch.atlassian.net/secure/BrowseProjects.jspa

⋮ 📄 ☆ 🔍

🔍 Buscar

🔧

📌 Jira Software

Tu trabajo

Proyectos

Filtros

Paneles

Personas

Aplicaciones

Crear

🔍

Buscar

🔔

?

⚙️

👤

Proyectos

🔍

Todos los tipos



←

→

↺

🏠

🔒 https://laraqresearch.atlassian.net/secure/RapidBoard.jspa?rapidView=1&projectKey=TWEET&view=planning.nodetail&issueLimit=100 ... 🔔 ⭐

🔍 Buscar

Jira Software

Tu trabajo

Proyectos

Filtros

Paneles

Personas

Aplicaciones

Crear

🔍

Buscar

🔔 ? ⚙️ LS

TweetEvolutionAndImpact

Proyecto de software clásico

Tablero TWEET

Tablero

Backlog

Sprints activos

Informes

📄

Incidencias

📁

Componentes

🔗

Código

📁

Versiones

📄

Páginas de proyectos

📄

Añadir elemento

⚙️

Configuración del proy...

Proyectos / TweetEvolutionAndImpact / Tablero TWEET

Backlog

Compartir

🔍

LS

Solo Mis Incidencias

Recientemente Actualizadas

VERSIONES

EPICS

Inicio Proyecto: Parte 1

4 incidencias

Conseguir definir todo el trabajo a realizar e identificar la novedad del proyecto

01/feb/21 11:56 AM • 01/mar/21 11:56 AM

📄

Busqueda bibliografica

TWEET-2

↑

-

📄

Analisis de la bibliografia

TWEET-3

↑

-

📄

Identificar alcance y objetivos

TWEET-1

↑

-

📄

Redacción de informe

TWEET-9

↑

-

Backlog

5 incidencias

Crear sprint

📄

Generación/descarga y limpieza de dataset

TWEET-4

↑

-

📄

Generación de atributos

TWEET-5

↑

-

📄

Generación del modelo

TWEET-6

↑

-

📄

Pruebas y comparación

TWEET-7

↑

-

📄

Análisis y resultados / insights

TWEET-8

↑

-

+ Crear incidencia

Quickstart



TweetEvolutionAndImpact
Proyecto de software clásico

Tablero TWEET
Tablero

Backlog

Sprints activos

Informes

Incidencias

Componentes

Código

Versiones

Páginas de proyectos

Añadir elemento

Configuración del proy...

Proyectos / TweetEvolutionAndImpact / Tablero TWEET

Backlog

🔗 Compartir ...

🔍 **LS** Solo Mis Incidencias Recientemente Actualizadas

Inicio Proyecto: Parte 1 4 incidencias

Conseguir definir todo el trabajo a realizar e identificar la novedad del proyecto

01/feb/21 11:56 AM • 01/mar/21 11:56 AM

VERSIONES

EPICS

Busqueda bibliografica	TWEET-2	↑	-
Analisis de la bibliografía	TWEET-3	↑	-
Identificar alcance y objetivos	TWEET-1	↑	-
Redacción de informe	TWEET-9	↑	-

Backlog 5 incidencias

Crear sprint ...

Generación/descarga y limpieza de dataset	TWEET-4	↑	-
Generación de atributos	TWEET-5	↑	-
Generación del modelo	TWEET-6	↑	-
Pruebas y comparación	TWEET-7	↑	-
Análisis y resultados / insights	TWEET-8	↑	-

+ Crear incidencia

TWEET-2 1 🔗 ... ✕

Busqueda bibliografica

...

Tareas por hacer ▾

Descripción

Empezaremos buscando keywords relacionados en google scholar.
Diseñaremos un búsqueda en scopus y WOS basado en ello
Descargaremos los archivos relacionados

Subtareas

... +

0 % hecho

- TWEET-10 Reunión identificar ke... **EN CURSO**
- TWEET-11 Redactar quer... **TAREAS POR HACER**
- TWEET-12 Ejecutar la qu... **TAREAS POR HACER**
- TWEET-13 Identificar du... **TAREAS POR HACER**
- TWEET-14 leer titulo y a... **TAREAS POR HACER**

LS

Consejo de expertos: pulsa **M** para comentar



← → ↺ 🏠

🔒 https://laraqresearch.atlassian.net/secure/RapidBoard.jspa?rapidView=1&projectKey=TWEET&selectedIssue=TWEET-3

⋮ 🛡️ ☆ 🔍 Buscar

🔧 Jira Software

Tu trabajo

Proyectos ▾

Filtros ▾

Paneles ▾

Personas ▾

Aplicaciones ▾

Crear

🔍 Buscar

🔔 ? ⚙️ LS

TweetEvolutionAndImpact

Proyecto de software clásico

Tablero TWEET

Tablero

Backlog

Sprints activos

Informes

Incidencias

Componentes

Código

Versiones

Páginas de proyectos

Añadir elemento

Configuración del proy...

Proyectos / TweetEvolutionAndImpact / Tablero TWEET

Inicio Proyecto: Parte 1

Conseguir definir todo el trabajo a realizar e identificar la novedad del proyecto

🔍 LS

Solo Mis Incidencias

Recientemente Actualizadas

POR HACER

EN CURSO

LISTO

> 🟢 TWEET-2

TAREAS POR HACER

6 sub-tareas

Busqueda bibliografica

▼ Otras incidencias

3 incidencias

Identificar alcance y objetivos

🟢 ⬆️ -

TWEET-1

Redacción de informe

🟢 ⬆️ -

TWEET-9

Analisis de la bibliografia

🟢 ⬆️ -

TWEET-3

🔗 ☆ ⌚ 33 días restantes

Terminar sprint

🔗 ⋮

💡 Quickstart

The screenshot displays the Jira Software interface. The top navigation bar includes the Jira logo, 'Tu trabajo', and various project management tabs like 'Proyectos', 'Filtros', 'Paneles', 'Personas', and 'Aplicaciones'. A search bar is located on the right. The left sidebar shows a project overview for 'TweetEvolutionAndImpact' with sections for 'Tablero TWEET', 'Backlog', 'Sprints activos', 'Informes', 'Incidencias', 'Componentes', 'Código', 'Versiones', 'Páginas de proyectos', 'Añadir elemento', and 'Configuración del proy...'. The main content area shows the 'Inicio Proyecto' section with a 'POR HACER' list. A modal window is open for issue 'TWEET-3', titled 'Analisis de la bibliografía'. The modal contains a description, a list of activities, and a sidebar with fields like 'Responsable', 'Informador', 'Etiquetas', 'Story Points', 'Sprint', and 'Prioridad'. The bottom right corner features a 'Quickstart' button.

URL: <https://laraqresearch.atlassian.net/secure/RapidBoard.jspa?rapidView=1&projectKey=TWEET&modal=detail&selectedIssue=TWEET-3>

Buscar

Jira Software Tu trabajo Proyectos Filtros Paneles Personas Aplicaciones Crear

TweetEvolutionAndImpact Proyecto de software clásico

Tablero TWEET Tablero

Backlog

Sprints activos

Informes

Incidencias

Componentes

Código

Versiones

Páginas de proyectos

Añadir elemento

Configuración del proy...

Proyectos / TweetEvolutionAndImpact

Inicio Proyecto: Analisis de la bibliografía

Conseguir definir todo el proyecto

POR HACER

TWEET-2 TAREAS POR HACER

Otras incidencias 3 incidencias

Identificar alcance y objetivos

Redacción de informe

TWEET-3

Analisis de la bibliografía

Adjuntar Crear subarea Vincular incidencia

Descripción

Identificaremos: -datasets

- atributos de entrada
- limpieza
- modelos usados
- resultados
- orden cronológico y de impacto de resultados

Actividad

Mostrar: Comentarios Historial Registro de trabajo

Añadir un comentario...

Consejo de expertos: pulsa para comentar

Enviar comentarios 1

En curso

Responsable Sin asignar

Informador LARA QUIJANO SANCHEZ

Etiquetas Ninguno

Story Points Ninguno

Sprint Inicio Proyecto: Parte 1

Prioridad Medium

Mostrar 5 campos más

Estimación original, Seguimiento de tiempo, Epic Link, Compon...

Fecha de creación 4 de noviembre de 2020 11:49

Fecha de actualización hace 1 minuto

Configurar

Quickstart



← → ↺ 🏠 <https://laraqresearch.atlassian.net/secure/RapidBoard.jspa?rapidView=1&projectKey=TWEET&modal=detail&selectedIssue=TWEET-3> 🔍 Buscar

Jira Software Tu trabajo Proyectos Filtros Paneles Personas Aplicaciones Crear

TweetEvolutionAndImpact Proyecto de software clásico

Tablero TWEET Tablero

Backlog

Sprints activos

Informes

Incidencias

Componentes

Código

Versiones

Páginas de proyectos

Añadir elemento

Configuración del proy...

Proyectos / TweetEvolutionAndImpact

Inicio Proyecto: Consegir definir todo el t

POR HACER

TWEET-2 TAREAS PO

Otras incidencias 3 inci

Identificar alcance y ob

Redacción de informe

TWEET-3

Analisis de la bibliografía

Adjuntar Crear subtask Vincular incidencia

Descripción

Identificaremos:

- datasets
- atributos de entrada
- limpieza
- modelos usados
- resultados
- orden cronológico y de impacto de resultados

Actividad

Mostrar: Comentarios Historial Registro de trabajo

Añadir un comentario...

Consejo de expertos: pulsa M para comentar

Enviar comentarios 1

En curso

Responsable

Informador

Etiquetas

Story Points

Sprint

Prioridad

Mostrar 5 cam

Estimación origin

Fecha de creación 4 c

Fecha de actualizació

Registrar trabajo

Añadir marca

Convertir en subtask

Mover

Clonar

Eliminar

Imprimir

Exportar XML

Exportar Word

NUEVA VISTA DE INCIDENCIAS DE JIRA

Ver los aspectos destacados

Obtener más información

Desactivar por ahora

Abrir incidencias en la barra lateral

33 días restantes Terminar sprint

Quickstart



← → ↻ 🏠 <https://laraqresearch.atlassian.net/secure/RapidBoard.jspa?rapidView=1&projectKey=TWEET&modal=detail&selectedIssue=TWEET-3> 🔍 Buscar

Jira Software Tu trabajo Proyectos Filtros Paneles Personas Aplicaciones Crear

TweetEvolutionAndImpact Proyecto de software clásico

Tablero TWEET Tablero

Backlog

Sprints activos

Informes

Incidencias

Componentes

Código

Versiones

Páginas de proyectos

Añadir elemento

Configuración del proy...

Proyectos / TweetEvolu

Inicio Proyecto:

Conseguir definir todo el t

POR HACER

TWEET-2 TAREAS PO

Otras incidencias 3 inci

Identificar alcance y ob

Redacción de informe

TWEET-3

Analisis de la bibliografía

Adjuntar Crear subtask Vincular incidencia

Descripción

Identificaremos:

- datasets
- atributos de entrada
- limpieza
- modelos usados
- resultados
- orden cronológico y de impacto de resultados

Subtareas

¿Qué hay que hacer?

Crear Cancelar

Actividad

Mostrar: Comentarios Historial Registro de trabajo

LS Añadir un comentario...

Consejo de expertos: pulsa M para comentar

Enviar comentarios 1

En curso

Responsable Sin asignar

Informador LS LARA QUIJANO SANCHEZ

Etiquetas Ninguno

Story Points Ninguno

Sprint Inicio Proyecto: Parte 1

Prioridad Medium

Mostrar 5 campos más

Estimación original, Seguimiento de tiempo, Epic Link, Compon...

Fecha de creación 4 de noviembre de 2020 11:49

Fecha de actualización hace 2 minutos

33 días restantes Terminar sprint

Quickstart

Todos los informes

Agile



Haz un seguimiento del alcance total, independientemente de todo el trabajo realizado. Esto permite a tu equipo administrar el progreso y comprender mejor el efecto de los cambios de alcance.



Haz un seguimiento del trabajo total restante y pronostica la probabilidad de alcanzar el objetivo del sprint. Esto ayuda al equipo a gestionar sus propios avances y responder en consecuencia.



Comprende qué trabajo se ha terminado o devuelto al backlog en cada sprint. Esto te ayudará a determinar si el equipo se está comprometiendo a hacer demasiado o si hay demasiada corrupción del alcance.



Haz un seguimiento de la cantidad de trabajo realizado de sprint a sprint. Esto te ayudará a determinar la velocidad del equipo y a estimar de forma realista el trabajo que este puede hacer en los sprints futuros.



Muestra el estado de las incidencias en el transcurso del tiempo. Esto lo ayudará a identificar posibles cuellos de botella que es necesario investigar.



Haz un seguimiento de la fecha de entrega prevista para una versión. Esto te ayudará a controlar si la versión se entregará a tiempo, para que puedas tomar medidas si el trabajo se está atrasando.



Comprende cómo se ha avanzado hacia la conclusión de un epic en el transcurso del tiempo. Esto te ayudará a administrar los avances del equipo haciendo un seguimiento del trabajo restante no estimado o sin terminar.



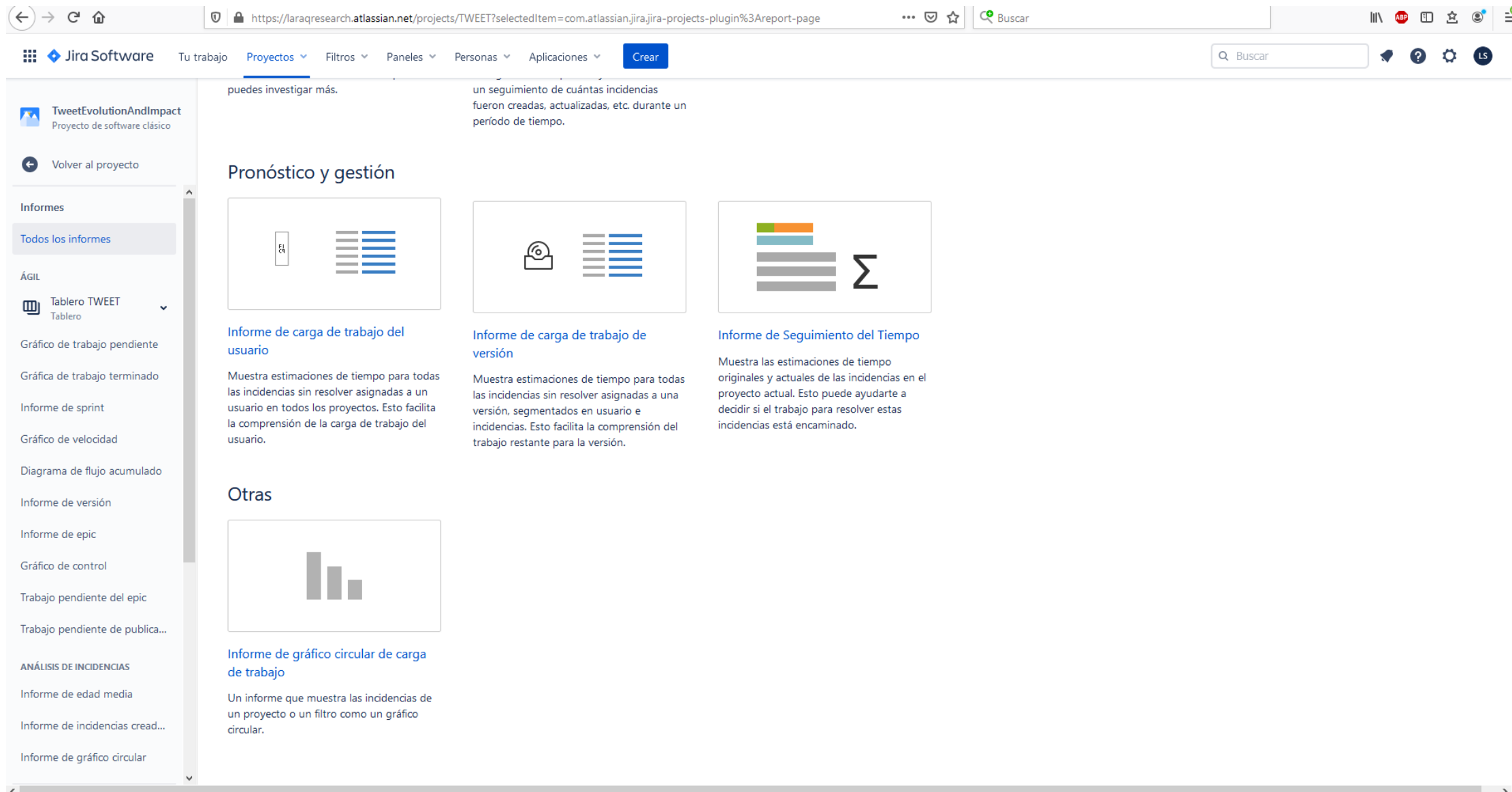
Muestra el tiempo de ciclo del producto, versión o sprint. Esto permite determinar si los datos del proceso actual pueden usarse para prever el rendimiento futuro.



Haz un seguimiento de la cantidad prevista de sprints necesarios para terminar el epic (optimizado para Scrum). Esto te ayudará a controlar si el epic se entregará a tiempo, para que puedas tomar medidas si el trabajo se está atrasando.



Haz un seguimiento de la fecha de entrega prevista para una versión (optimizado para Scrum). Esto te ayudará a controlar si la versión se entregará a tiempo, para que puedas tomar medidas si el trabajo se está atrasando.



Tablero

Planificación de tareas

Equipo TFG

Deberes

Haciendo

Descargar intelliJ;
https://www.jetbrains.com/idea/

Crear una cuenta en bitbucket -
https://bitbucket.org/

prueba

Añade otra tarjeta

Descargar git
0/3

Añade otra tarjeta

Descargar git

en la lista Haciendo

ETIQUETAS

Descripción

Añadir una descripción más detallada...

Pasos

0%

☐ https://git-scm.com/download/win

☐ Instalar con opciones por defecto pero haciendo que git se meta en el path

☐ en el paso Adjusting your PATH environment usar 2ª opción

Añade un elemento

Actividad

Escriba un comentario...

LARA QUIJANO SANCHEZ ha quitado a Checklist de esta tarjeta
30 de oct. de 2020 a las 14:26

LARA QUIJANO SANCHEZ ha añadido Checklist a esta tarjeta
30 de oct. de 2020 a las 14:26

LARA QUIJANO SANCHEZ ha movido esta tarjeta de Deberes a Haciendo
28 de oct. de 2020 a las 12:47

LARA QUIJANO SANCHEZ ha añadido Pasos a esta tarjeta
16 de oct. de 2015 a las 14:46

AÑADIR A LA TARJETA

Miembros

Etiquetas

Checklist

Vencimiento

Adjunto

Portada

POWER-UPS

Añadir Power-Ups

Consiga Power-Ups ilimitados y mucho más.

Actualizar el equipo

BUTLER NUEVO

Añadir botón en t...

ACCIONES

Mover

Copiar

Convertir en planti...

Seguir

Archivar

Compartir

Bitbucket

← → ↺ 🏠

🔒 <https://bitbucket.org/dashboard/projects>

⋮ 🛡️ ☆

🔍 Buscar

☰ 🔴 ABP 📄 👤 🌐 📢

Bitbucket

🔍 +

📄 Your work

📁 Repositories

📁 Projects

🔗 Pull requests

✂ Snippets

📱 ? L

Projects

Create project

Project	Key	Description	
📁 BigDataTurismApp	BDTA		🔒
📁 DataCleaningPeña	DAT		🔒
📁 DenunciasFalsas	DEN		
📁 domesticViolence	DOM		🔒
📁 HappyMovie	PROJ	Project created by Bitbucket for HappyMovie	
📁 Haternet	HAT		
📁 Ladorian Cofares	LC		🔒
📁 proyectoconversacional	PROVEC		
📁 Recommenders+SmartCities+EGovernance	REV		
📁 ShortestPathPaper	SHOR		🔒
📁 Tourism Paper	TP		
📁 Untitled project	PROJ	Project created by Bitbucket for LaraAndFederico	
📁 Untitled project	PROJ	Project created by Bitbucket for Lara	
📁 Untitled project	PROJ	Project created by Bitbucket for BigDataTourismApp	

<https://bitbucket.org/dashboard/projects>

Bitbucket

← → ↻ 🏠

🔒 https://bitbucket.org/dashboard/repositories

⋮ 🛡️ ☆ 🔍 Buscar

📁 Bitbucket

🔍 +

📄 Your work

📁 Repositories

📁 Projects

🔗 Pull requests

✂️ Snippets

📱

?

L

Create repository

Search repositories 🔍

Workspace ▾

Project ▾

Watching

Summary

Description

Updated ▾

Builds

</>

haternet paper

LaraAndFederico / Haternet

</>

Egovernance-Conversacional

UAM Recommenders / Recommenders+SmartCities+EGovernance

</>

proyectoconversacionalserver

ConLara / proyectoconversacional

</>

proyectoconversacionalweb

ConLara / proyectoconversacional

</>

domesticViolenceDataAnalysis

FLiberatore / domesticViolence

</>

Paper-Review: SmartCities+Recommenders

UAM Recommenders / Recommenders+SmartCities+EGovernance

</>

Policia-DenunciasFalsas

LaraAndFederico / DenunciasFalsas

</>

domesticViolenceRiskModel

FLiberatore / domesticViolence

📄

TieStrengthPaper

LaraAndFederico / Untitled project

📄

Report

LaraAndFederico / Ladorian Cofares

📄

ShortestPathPaper

LaraAndFederico / ShortestPathPaper

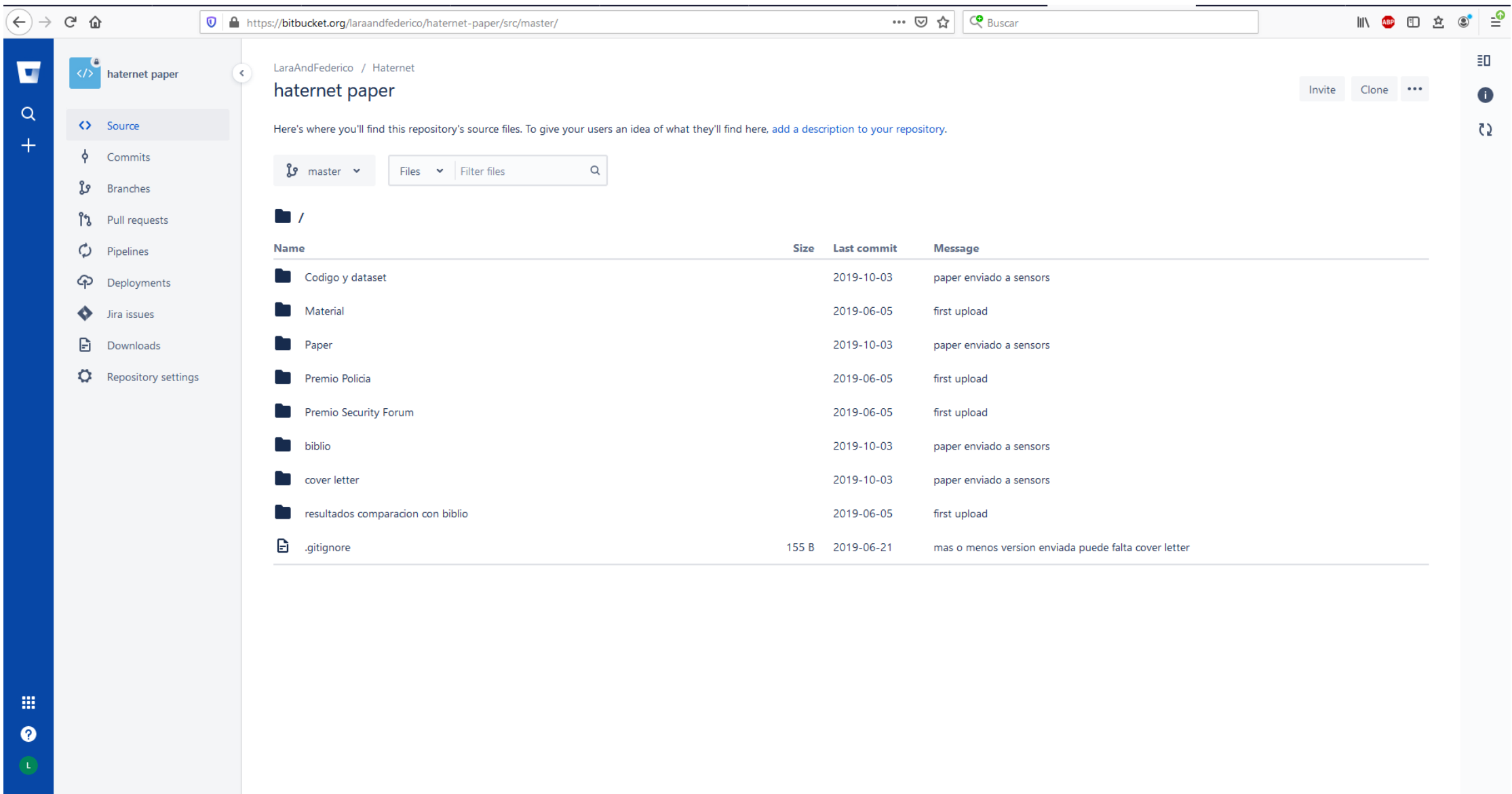
📄

Tourism Paper

LaraAndFederico / Tourism Paper

📄

Controller



Sourcetree

Re

Clone / New...
Ctrl+N

Open...
Ctrl+O

Exit Sourcetree
Alt+F4

Branch

Merge

Stash

Discard

Tag

ver - proyectoConversac...

Web - proyectoConversacio...

HaterNet

Egovernance_Conversacional

Git-flow

Remote

Terminal

Explorer

Settings

FILE STATUS

Working Copy

BRANCHES

master

TAGS

REMOTES

STASHES

All Branches

Show Remote Branches

Date Order

Graph

Uncommitted changes

master origin/master paper enviado a sensors

mas o menos version enviada puede falta cover letter

first upload

Jump to:

Description	Date	Author	Commit
paper enviado a sensors	12 ene. 2021 14:44	*	*
mas o menos version enviada puede falta cover letter	3 oct. 2019 12:47	LaraQ <lara.quijan	9668d9e
first upload	21 jun. 2019 12:48	LaraQ <lara.quijan	976b563
	5 jun. 2019 14:52	LaraQ <lara.quijan	0da6ae2

Pending files, sorted by file status

Staged files

Unstage All

Unstage Selected

Unstaged files

Stage All

Stage Selected

cover letter/coverletter.pdf

Paper/biblio.bib

Paper/exampleUseLara.png

Paper/HaterNet__Sensors_Special_Issue.pdf

Paper/HaterNetPaper.tex

Paper/usementionedLara.png

Paper/userInterfacelara.png

File Status

Log / History

Search

Select a file to view the diff

Overleaf

Menu

Source Rich Text

Recompile

images

ACM-Reference-Format.bbb

ACM-Reference-Format.bst

ACM-Reference-Format.cbx

ACM-Reference-Format.dbx

ACM_SAC.tex

ACM_SAC_short.tex

acmart.bib

acmart.cls

acmart.dtx

acmart.ins

biblio.bib

Makefile

File outline

Introduction

Case study: Decide Madrid Platform

Electronic participatory budgeting

The Decide Madrid platform

Mining topics of interest

Text processing

Document similarity

Document clustering

Citizen participation analysis tool

Data analysis functionalities

Analysis insights

Conclusions

ACM SAC 2021

transport, the most affected is the university district, Moncloa-Aravaca, which should have enough resources to cover the great transport demand that exists and that despite the large number of stations on the main metro lines, it should be bare in mind that the Madrid Metro does not have night service, with all the burden falling (taking into account what this means in a university neighborhood) on night buses. %The district only has four night city buses, all originating from Plaza de Cibeles, and also taking into account the large time lapses between buses, it is logical that Moncloa-Aravaca calls for an improvement in its night transport services.

357 %This fact is highlighted for example in in Figure \ref{fig:dashboard4} left.

358

359 %REPETIDO While on the right of the Figure we can check the impact the Tourism category, where the bulk of proposals come from the central region, where the number of places of interest is far higher.

360

361 \iffalse

362 \begin{figure}[t]

363 \centering

364 \includegraphics[width=0.6\linewidth]{images/d4}

365 \caption{Heat map showing the relevance of proposal topics (communities) and categories on each district}

366 \label{fig:dashboard4}

367 \end{figure}

368 \fi

369

370 \vspace{-0.3cm}

371

372 %%%

373 \section{Conclusions}

374 \label{sec:conclusions}

375 %%%

376

377 Nowadays, cities are implementing online platforms for citizen participation where inhabitants participate in municipal decisions and actions by means of proposals and debates. To date, these platforms are suffering problems related to the citizens' frustration for the lack of visibility and impact of their proposals, and to a low participation due to information overload and exploration difficulties. To address these problems, researchers have proposed the development of technological solutions to summarize and contextualize citizen feedback, and visualize individual and community needs and concerns \cite{marzouki2017relevance, cantador2020exploiting}. Following this direction, we have presented a flexible interactive tool that provides a variety of visualization and analysis functionalities for citizen generated content, and facilitates to both citizens and local governments the understanding of the underlying problems in a city and proposed solutions.

378

379 \vspace{-0.3cm}

380

381 %%%

382 %\section{Acknowledgements}

383 %\label{sec:acks}

384 %%%

A flexible and lightweight interactive data mining tool to visualize and analyze digital citizen participation content

Sergio Bachiller

s.bachillerrubia@gmail.com

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Madrid, Spain

Lara Quijano-Sánchez*

lara.quijano@uam.es

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Madrid, Spain

Iván Cantador

ivan.cantador@uam.es

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Madrid, Spain

ABSTRACT

Citizen collaboration through current digital participation platforms can entail the generation of large amounts of complex content, which may hide relevant citizens' concerns, requests and initiatives, diluted in isolated individual proposals. Addressing this problem, in this paper, we present an interactive data mining tool for citizen participation data visualization and analysis. The tool implements natural language processing, text similarity, and graph clustering techniques to group proposals with common objectives, identify trends and recurrent topics of interest, and filter and present information according to several criteria. The tool is flexible, able to process different sources of data, and lightweight as it uses simple data structures and dynamic HTML-based visualization and interaction. As a case study, the tool has been instantiated with a dataset obtained from the Decide Madrid e-participatory budgeting platform.

CCS CONCEPTS

• Applied computing → Document management and text processing; • Human-centered computing → Visualization; • Information systems → Data mining.

KEYWORDS

citizen participation, data visualization, data clustering, text similarity, civic technologies

ACM Reference Format:

Sergio Bachiller, Lara Quijano-Sánchez, and Iván Cantador. 2021. A flexible and lightweight interactive data mining tool to visualize and analyze digital citizen participation content. In *Proceedings of ACM SAC Conference (SAC'21)*. ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/xx.xxx/xxxx_x

1 INTRODUCTION

With the advent of social media and mobile computing, nowadays there is a plethora of digital citizen participation channels, ranging from general purpose online social networks to ad hoc e-participation, e-consultation and e-voting platforms.

The huge, ever increasing citizen generated content leads to an information overload problem for both citizens and government stakeholders in decision and policy making tasks. In particular, they may feel overwhelmed by the large amount of data, whose exploration and understanding could result challenging and frustrating. Also, citizens may feel thwarted if their proposals do not reach sufficient visibility and impact. In this sense, we may find several proposals by different authors that address the same problem, but in different ways, for distinct city locations, or entailing distinct initiatives and potential solutions.

To address the above problems, there is a need for information systems capable to process and mine citizen generated content, as well as to summarize, visualize and analyze relevant extracted information overtime. Motivated by this need, in this paper we present a flexible, lightweight data mining tool that helps unraveling public deliberation contents to both governments and citizens. On the one hand, local governments are constantly evolving organizations, and when neither long-term growth plans nor daily operations of critical city services are stable, the need for robust, accessible and meaningful community engagement is crucial. The goal here is to develop systems and strategies that enable people to form policies with an impact on their communities and own lives [18]. In this context, our tool makes these activities simpler by interactively analyzing knowledge generated from public initiatives, e.g., detecting trends in the concerns posed to policymakers, persistent demands or particular seasonal problems. On the other hand, considering the citizens' contributions is an important objective for governments, scientists and companies, since it can have a great impact on issues of common interest. Existing civic technologies focus mainly on rewards, only dealing with values such as power, achievement, security and encouragement, and leaving a gap with respect to other values [26]. The presented tool allows unifying objectives by grouping and visually monitoring proposals (that could have been accepted or remain unanswered), thus serving as a stimulus for citizens to increase quality (i.e., more informed and argued proposals) and quantity (more proposals due to ease of use) on their generated content.

In summary, we present an interactive tool for citizen participation data visualization and analysis, which is built upon the Tableau interactive data visualization software, and which is lightweight and easy to configure, as well as generic, since it is reusable for other related domains and different languages as it uses data from

Review

Share

Submit

History

Chat

Overleaf

https://www.overleaf.com/project/5f60a319d6d76300015286b6

ACM SAC 2021

Menu

Source Rich Text

images

ACM-Reference-Format.bbx

ACM-Reference-Format.bst

ACM-Reference-Format.cbx

ACM-Reference-Format.dbx

ACM_SAC.tex

ACM_SAC_short.tex

acmart.bib

acmart.cls

acmart.dtx

acmart.ins

biblio.bib

Makefile

File outline

Introduction

Case study: Decide Madrid Platform

Electronic participatory budgeting

The Decide Madrid platform

Mining topics of interest

Text processing

Document similarity

Document clustering

Citizen participation analysis tool

Data analysis functionalities

Analysis insights

Conclusions

transport, the most common resources to cover night service, with stations on the neighborhood) on Plaza de Cibeles, that Moncloa-Aravaca. This fact is higher. While on the bulk of proposals higher.

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

Nowadays, cities are implementing online platforms for citizen participation where inhabitants participate in municipal decisions and actions by means of proposals and debates. To date, these platforms are suffering problems related to the citizens' frustration for the lack of visibility and impact of their proposals, and to a low participation due to information overload and exploration difficulties. To address these problems, researchers have proposed the development of technological solutions to summarize and contextualize citizen feedback, and visualize individual and community needs and concerns [cite{marzouki2017relevance, cantador2020exploiting}]. Following this direction, we have presented a flexible interactive tool that provides a variety of visualization and analysis functionalities for citizen generated content, and facilitates to both citizens and local governments the understanding of the underlying problems in a city and proposed solutions.

378

379

380

381

382

383

384

With Mendeley integration you can import your references from Mendeley into your Overleaf projects.

Link to Mendeley

Cancel Create

Lightweight interactive data mining tool to analyze digital citizen participation content

Lara Quijano-Sánchez
lara.quijano@uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Iván Cantador
ivan.cantador@uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

participation platforms of complex content, requests and initiatives, addressing this problem, mining tool for citizen participation data visualization and analysis. The tool implements natural language processing, text similarity, and graph clustering techniques to group proposals with common objectives, identify trends and recurrent topics of interest, and filter and present information according to several criteria. The tool is flexible, able to process different sources of data, and lightweight as it uses simple data structures and dynamic HTML-based visualization and interaction. As a case study, the tool has been instantiated with a dataset obtained from the Decide Madrid e-participatory budgeting platform.

CCS CONCEPTS

- Applied computing → Document management and text processing;
- Human-centered computing → Visualization;
- Information systems → Data mining.

KEYWORDS

citizen participation, data visualization, data clustering, text similarity, civic technologies

ACM Reference Format:

Sergio Bachiller, Lara Quijano-Sánchez, and Iván Cantador. 2021. A flexible and lightweight interactive data mining tool to visualize and analyze digital citizen participation content. In *Proceedings of ACM SAC Conference (SAC'21)*. ACM, New York, NY, USA, Article 4, 9 pages. <https://doi.org/10.1145/3456789>

1 INTRODUCTION

With the advent of social media and mobile computing, nowadays there is a plethora of digital citizen participation channels.

*Corresponding author.

© Copyright 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions.acm.org.

SAC'21, March 22-March 26, 2021, Gwangju, South Korea

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8104-8/21/03...\$15.00

<https://doi.org/10.1145/3456789>

range from general-purpose online social networks to ad hoc e-participation, e-consultation and e-voting platforms.

The huge, ever increasing citizen generated content leads to an information overload problem for both citizens and government stakeholders in decision and policy making tasks. In particular, they may feel overwhelmed by the large amount of data, whose exploration and understanding could result challenging and frustrating. Also, citizens may feel thwarted if their proposals do not reach sufficient visibility and impact. In this sense, we may find several proposals by different authors that address the same problem, but in different ways, for distinct city locations, or entailing distinct initiatives and potential solutions.

To address the above problems, there is a need for information systems capable to process and mine citizen generated content, as well as to summarize, visualize and analyze relevant extracted information overtime. Motivated by this need, in this paper we present a flexible, lightweight data mining tool that helps unraveling public deliberation contents to both governments and citizens. On the one hand, local governments are constantly evolving organizations, and when neither long-term growth plans nor daily operations of critical city services are stable, the need for robust, accessible and meaningful community engagement is crucial. The goal here is to develop systems and strategies that enable people to form policies with an impact on their communities and own lives [18]. In this context, our tool makes these activities simpler by interactively analyzing knowledge generated from public initiatives, e.g., detecting trends in the concerns posed to policymakers, persistent demands or particular seasonal problems. On the other hand, considering the citizens' contributions is an important objective for governments, scientists and companies, since it can have a great impact on issues of common interest. Existing civic technologies focus mainly on rewards, only dealing with values such as power, achievement, security and encouragement, and leaving a gap with respect to other values [26]. The presented tool allows unifying objectives by grouping and visually monitoring proposals (that could have been accepted or remain unanswered), thus serving as a stimulus for citizens to increase quality (i.e., more informed and argued proposals) and quantity (more proposals due to ease of use) on their generated content.

In summary, we present an interactive tool for citizen participation data visualization and analysis, which is built upon the *Tableau* interactive data visualization software, and which is lightweight and easy to configure, as well as generic, since it is reusable for other related domains and different languages as it uses data from

<https://www.tableau.com>

Gestor bibliográfico

❑ Mendely

❑ <https://www.youtube.com/watch?v=5Uc4e6ULzI4>

❑ https://biblioguias.unex.es/ld.php?content_id=30163509

❑ Se sincroniza con Overleaf

