

Desde la teoría y la práctica tradicionales hacia la archivística computacional (CAS): nuevos conceptos y enormes posibilidades para el avance de nuestra disciplina

Maria Esteva

3a sesión del Seminario *Permanente de Investigación en Archivística y
Gestión de Documentos*

Departamento de Historia Moderna y Contemporánea

Universidad de Barcelona

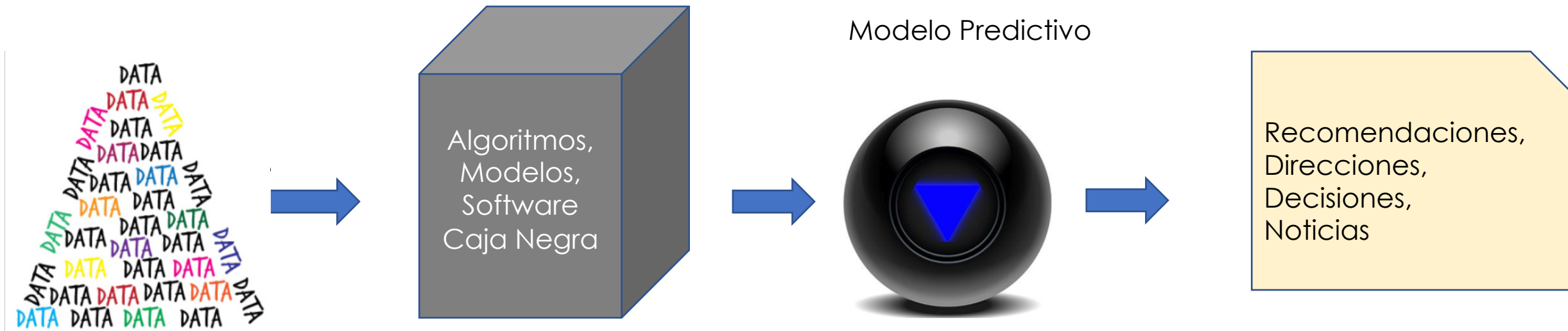
10 de Febrero de 2021

Problemas viejos, nuevos interrogantes

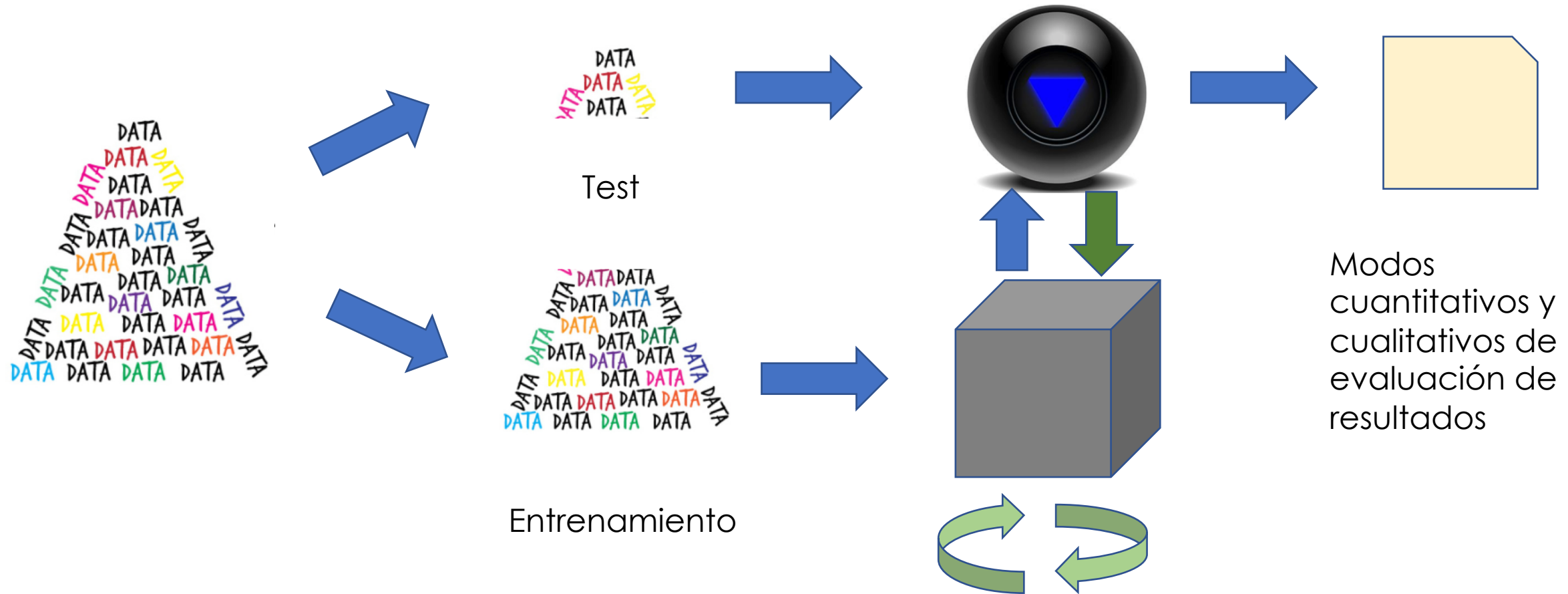
- Nuevos métodos para procurar acceso a archivos digitales de gran escala.
- El aprendizaje automático puede acelerar y mejorar el procesamiento de registros digitales.
- Con quiénes vamos a trabajar?
- Cómo integrar los valores y la teoría archivística en los procesos computacionales?
- Qué conocimientos debemos adquirir para adoptar y adaptar estos métodos?
- Cuáles serán nuestras contribuciones?

Veamos el proceso desde un caso de estudio.

Qué es el aprendizaje automático?



Funcionamiento y evaluación



Descargos

- No somos los únicos.
- Proceso interdisciplinario y experimental.
- Conocimientos funcionales de estadísticas y análisis de datos.
- Modelar procesos a técnicas computacionales:
 - No significa automatizar pasos.
 - Definir y deconstruir = abstraer = modelar.
 - Para crear soluciones.

Formular y alinear objetivos

Archiveros

- Generar metadatos/etiquetas para registros orales.
- Se puede emular el trabajo descriptivo de los archiveros?
- Se pueden utilizar materias (aplicadas manualmente) para entrenar modelos de aprendizaje automático?
- Cómo integramos valores y prácticas profesionales en los procesos computacionales?

Científicos computación

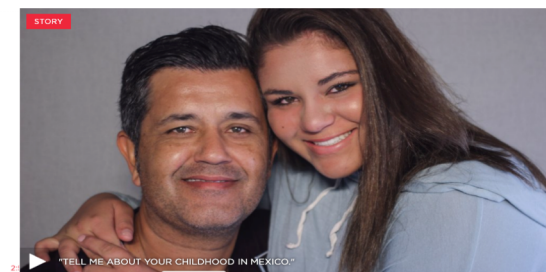
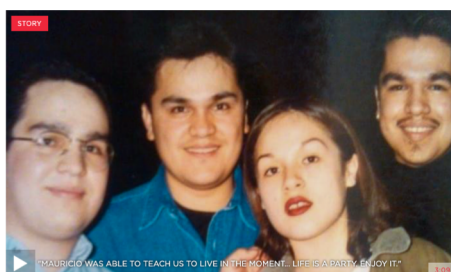
- Cuáles son los modelos mas efectivos para predecir metadatos/etiquetas?
- Cómo evaluar y seleccionar los datos para entrenar a los modelos?
- Hasta cuántas etiquetas podremos predecir con confianza?
- Qué interfaces son necesarias para bajar las barreras de adopción para los archiveros?

Los objetivos se revisan a medida que el trabajo progresa y se alinean las perspectivas profesionales.

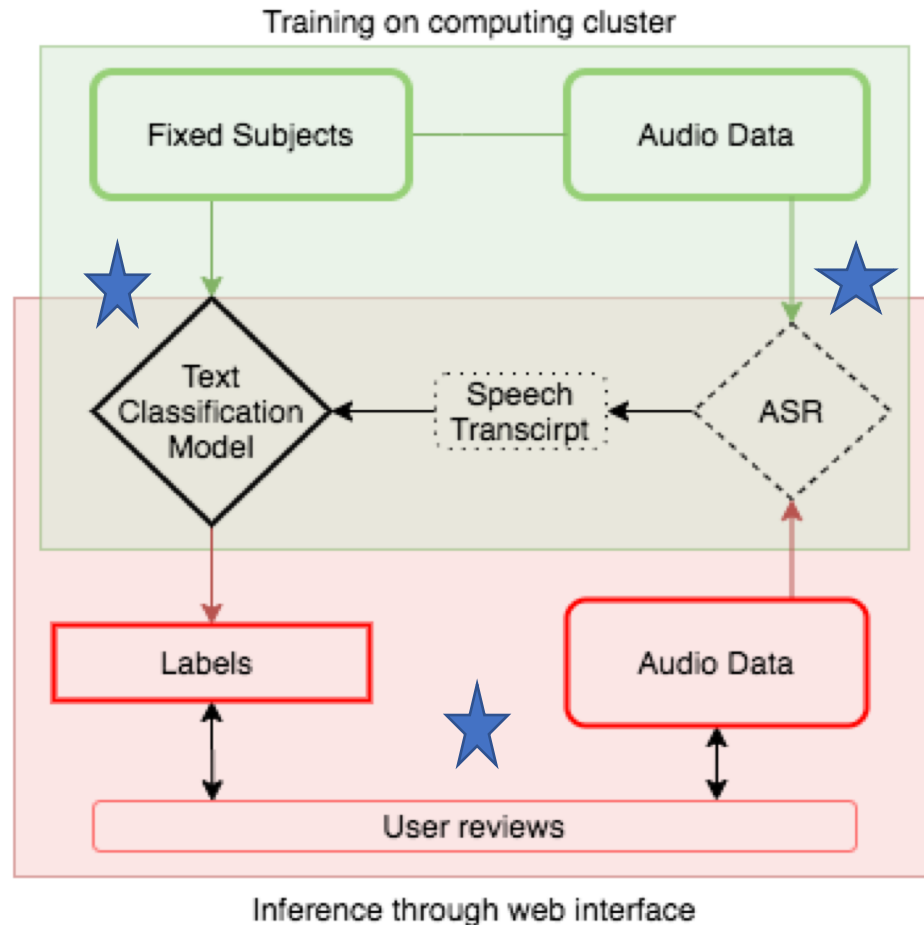
Historias: una colección de [StoryCorps](#)

- Programa de la radio pública.
- Captura experiencias de vida de la comunidad Latina en USA.
- Entrevistas entre dos participantes con la presencia de un facilitador.
- En inglés, con salpicado de castellano.
- 815 archivos de audio de hasta una hora de duración.
- Los facilitadores generan metadatos de cada entrevista en el momento.
- Aplican entre 5 y 15 **materias fijas de un vocabulario controlado y** palabras claves.
- Identificadas al comienzo, mitad, y final de la entrevista.
- Énfasis en descripción general para acceso.
- La aplicación de materias puede ser inconsistente entre los facilitadores.

Entrevistamos a las archiveras de StoryCorps para conocer la proveniencia de las entrevistas y su procesamiento.



Metodología



- Discurso-a-texto.
 - Experimentos
- Texto-a-etiquetas.
 - Experimentos
- Evaluación cuantitativa y cualitativa.
- Flujo de trabajo interactivo mediante la aplicación web IDOLS.
 - Usabilidad
 - Transparencia
 - Reproducibilidad

Trabajos relacionados

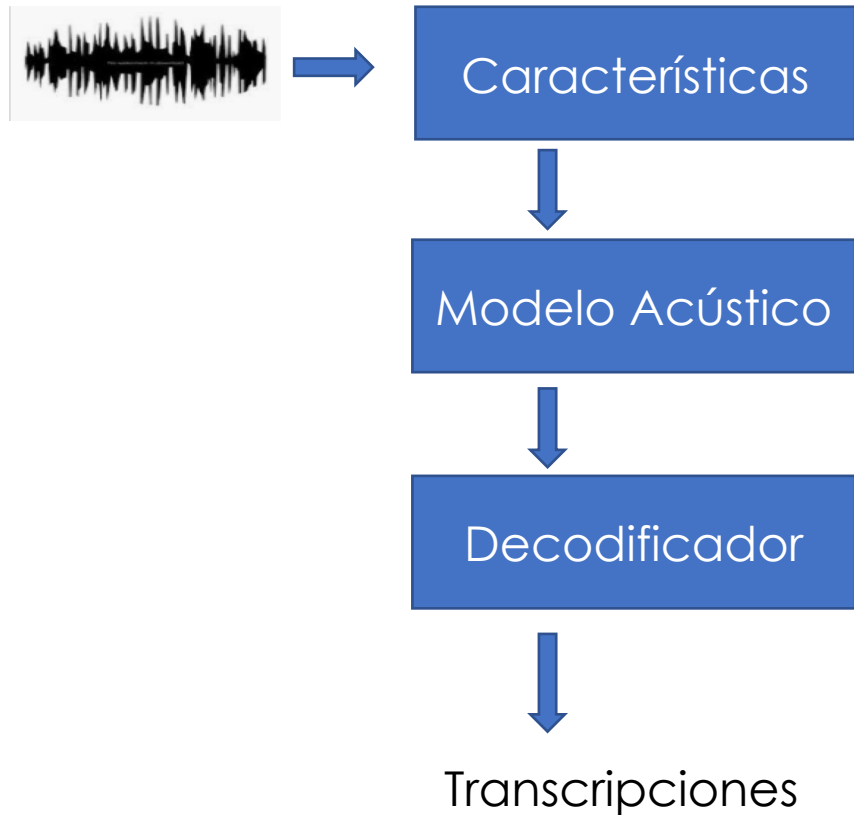
Archiveros

- Aumentan las iniciativas [y fóruns](#) de AI/ML para archivos y [bibliotecas digitales](#).
- Discurso-a-texto:
 - Transcripciones que representen al registro con precisión.
- Texto-a-etiquetas:
 - Uso de metadatos como datos de entrenamiento.
 - No son perfectos para entrenar modelos de descripción [de imágenes](#).
- Adopción
 - Principios de [Santa Clara](#).
 - Interfaces para interactuar con los modelos de aprendizaje de máquina.

Científicos de computación

- Espacio de investigación activo desde hace mucho tiempo.
- Discurso-a-texto
 - Reconocimiento automatizado de discurso (Automated Speech Recognition ASR).
- Texto-a-etiquetas:
 - Problema de clasificación.
- Adopción
 - Portal a tecnología implementados en sistemas computacionales compartidos que pueden ser accedidos de modo remoto.

Discurso-a-texto



- Google Speech to Text (GST)
 - <https://cloud.google.com/speech-to-text>
 - ~ 2.4 cent por minuto.
 - Soporta modelos acústicos complejos y múltiples idiomas.
- DeepSpeech (DS)
 - <https://github.com/mozilla/DeepSpeech>
 - Código abierto.
 - Reporta buena performance.
 - Modelos acústicos y de idiomas limitados.
- Transcripción Humana (HT)
 - Servicios profesionales de la Biblioteca de la Universidad de Texas.

Discurso-a-texto: resultados

- Google Speech to Text (GST)

- 311 transcripciones
- Implementación mas difícil de lo promocionado.
- Post procesamiento (NLP) para mejorar los resultados.

“well it's a great opportunity to talk to my mom and it's I'm here now for an interview but just to have a conversation with my mom about her life and and how wonderful it has been so this far so I'm going to have some questions we're going to talk about a lot of things”

- DeepSpeech (DS)

- 815 transcripciones.
- Modelos pre-entrenados con grabaciones de estudio.
- Mapeo directo de sonido a palabras.

“weathered oportuna talk to my maman um it's oh i'm here now for nine view but do just have a conversation with my mam about her life and um and how wonderful has been so this fort so amantha some questions when i talk about it not a thing”

- Transcripción Humana (HT)

- 81 transcripciones.
- Standard de Golden standard.
- Incluye puntuación y separación de participantes.

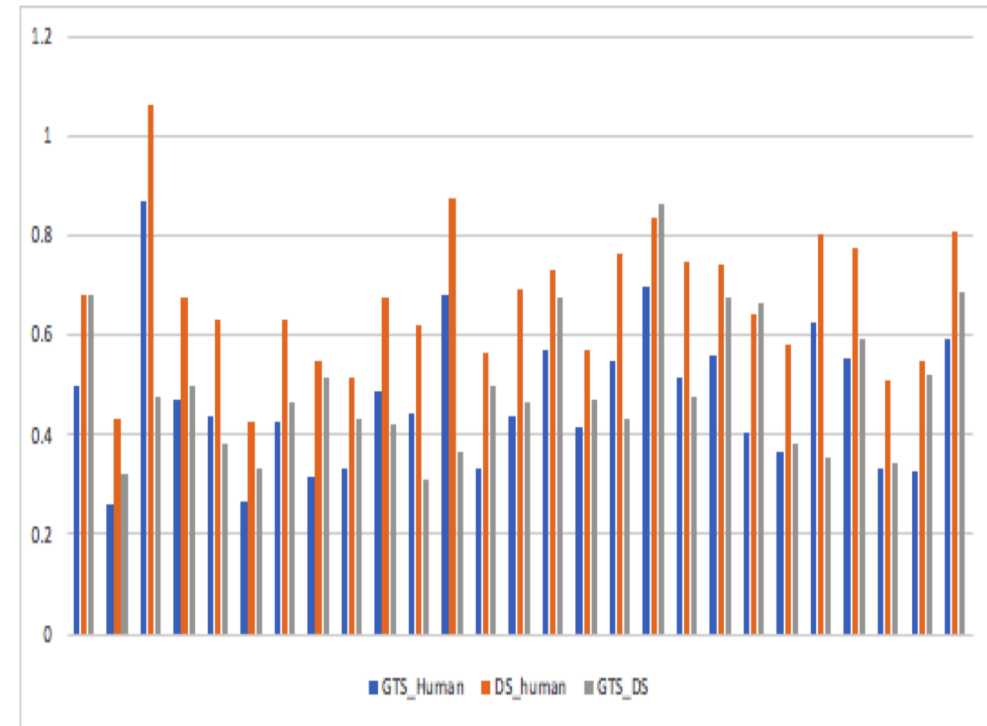
[Interviewee:] Well, this is a great opportunity to talk to my mom, and I'm here now for an interview, but just to have a conversation with my mom about her life and how wonderful it has been this far. So I'm going to have some questions, and we're going to talk about a lot of things.

Discurso-a-texto: comparación de resultados

- Índice de error de palabra (Word Error Rate)

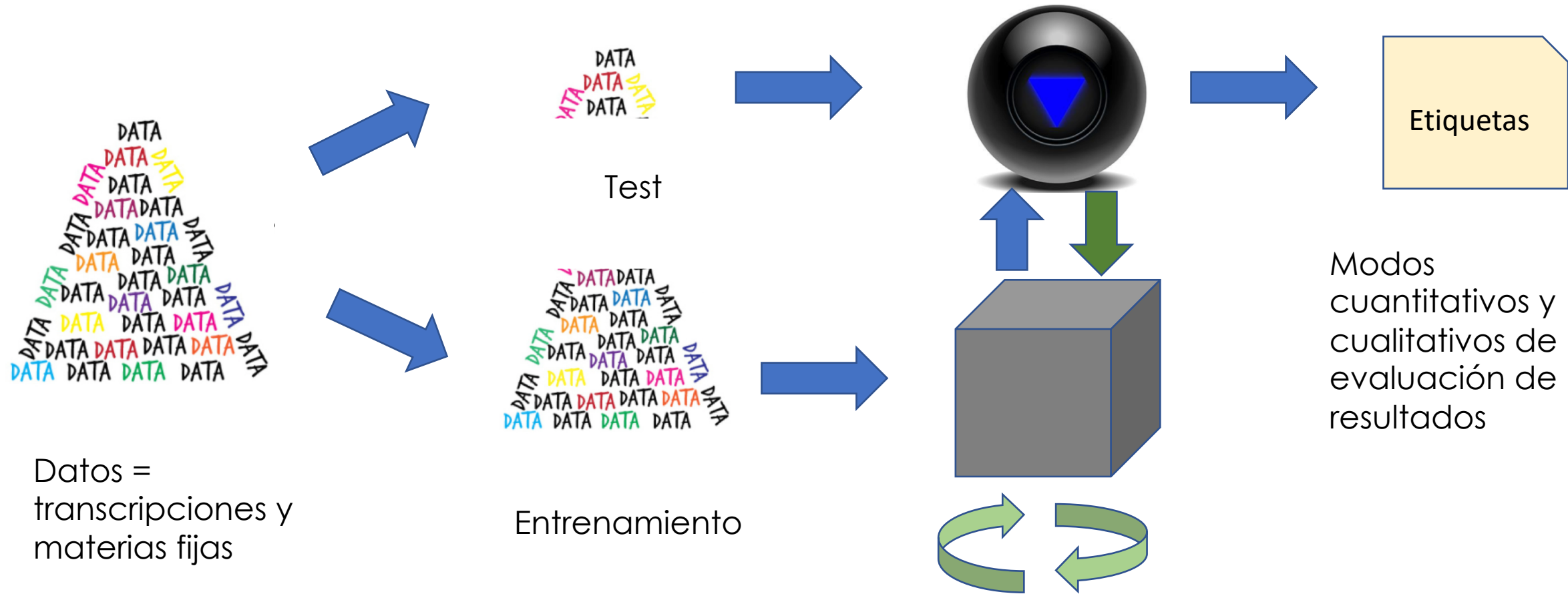
$$\text{WER} = \frac{\text{number of total edits required}}{\text{total length of the reference sequence}}$$

Los mejores resultados (menor índice de error) corresponden al modelo de Google (azul)



Comparación de resultados WER entre transcripciones DS, GST y HT

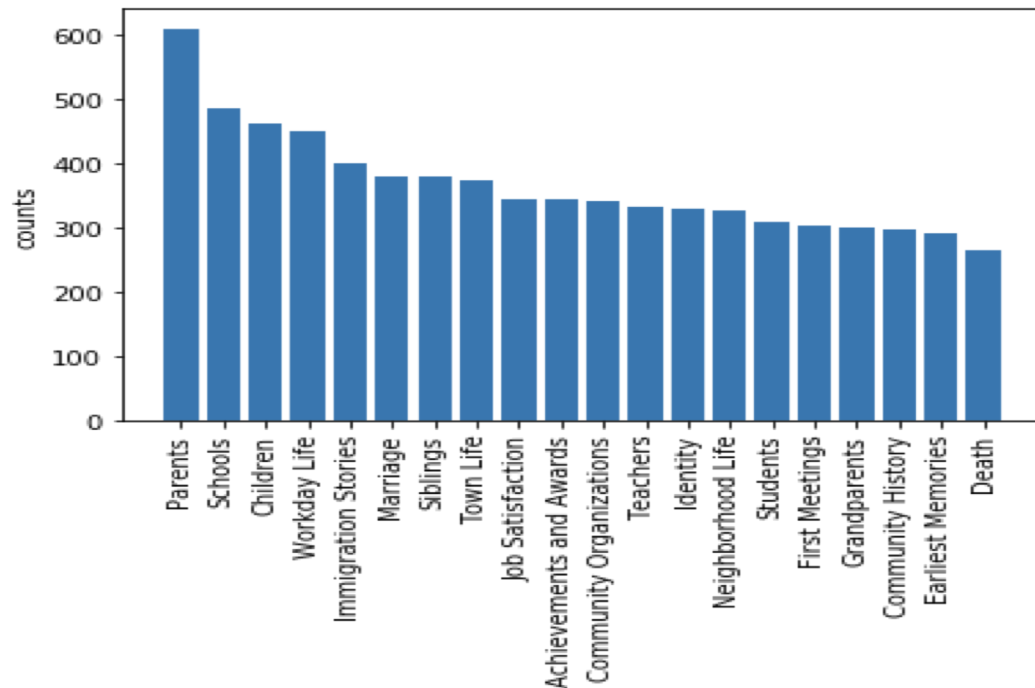
Texto-a-etiquetas



Problema de clasificación

Atención: Weijia Xu

Texto-a-etiquetas: materias para entrenamiento



- Para entrenar el modelo necesitamos materias que se encuentren en la mayoría de las entrevistas.
- Utilizamos las 20 materias asignadas con mayor frecuencia a través del corpus (Top 20)

Distribución de las top 20 materias fijas sobre numero de entrevistas.

Podemos predecir al menos tres etiquetas por entrevista con confianza?

Texto-a-etiquetas: modelos de aprendizaje supervisados

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait réuni les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) deux divinités, groupées deux à deux, s'ordonnaient en six couples : dans une dièdre. Au centre de la frise, en sarronbre, les deux divinités (féminine et masculine) qui présidaient aux unions : Aphrodite et Hésios. Dans cette série de huit couples divins, il en est un qui fait problème. Hermès-Hestia. Pourquoi les appeler ? Rien dans leur génétique ni dans leur légende qui puisse justifier cette association. Ils ne sont ni mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héstios-Héraclès), ni frère et sœur (comme Apollon-Artemis, Hélios-Séléné), ni ne et fille (comme Aphrodite-Eros), ni protectrice et protégé (comme Héra-Héraclès). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent égarés l'un à l'autre ? On ne saurait leur assigner une fantaisie personnelle du sculpteur. Quand il exécutait une œuvre sacrée, l'artiste ancien en tenait de se conformer à certains modèles : à l'initiative s'exerce dans le cadre des schémas imposés par la tradition. Hestia - non propre d'une déesse mais aussi non comme déesse le foyer - se prêtait moins que les autres dieux grecs à la réification anthropomorphe. On la voit rarement figurée. Quand elle est, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Zeus. De Zeus dans l'art religieux, l'association Hermès-Hestia.



Palabras



Representación

Materias fijas



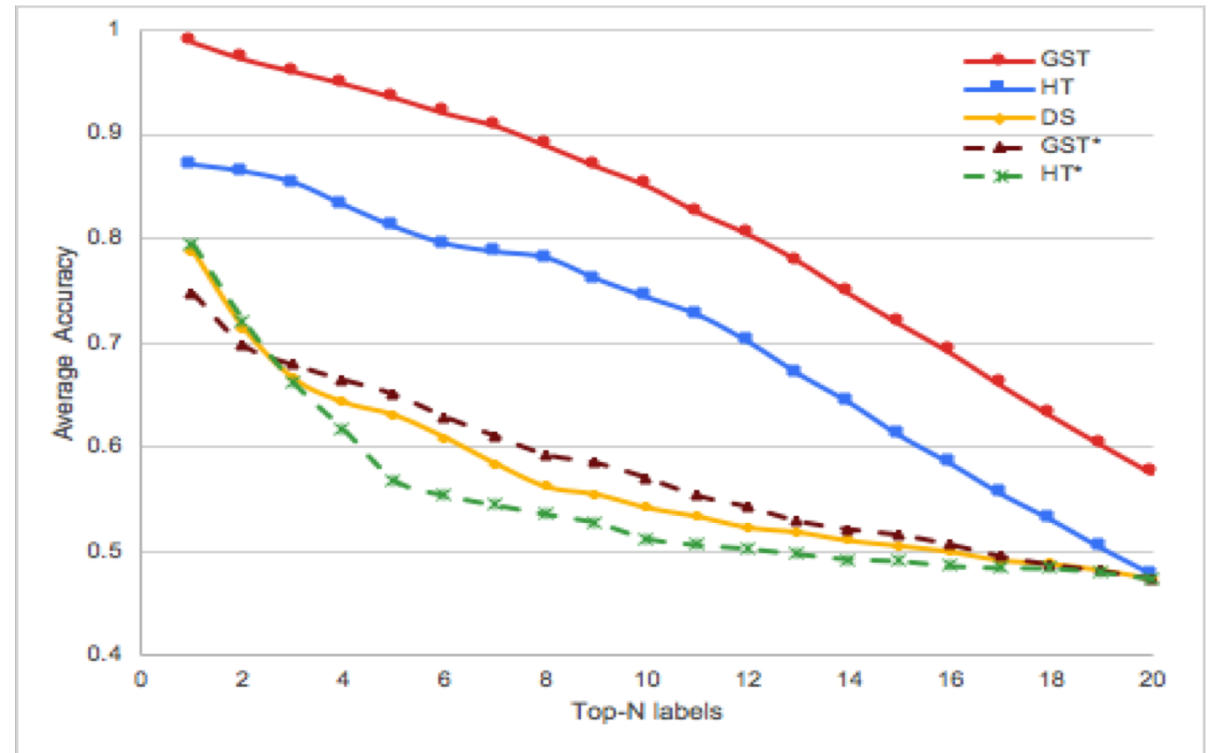
Modelo de clasificación

Etiqueta simple
vs múltiples
etiquetas

Modelos:
Random Forest,
SVM
Naïve Bayes
Deep Learning
...

Texto-a-etiquetas: evaluación cuantitativa

- Probamos diferentes combinaciones de datos y modelos:
 - Modelos de clasificación de etiquetas simples y múltiples.
 - Entrenamos el modelo usando 80% de las transcripciones.
 - Para el testeo utilizamos 20% de las transcripciones.
- Las combinaciones se evalúan mediante el promedio de exactitud (average accuracy)
- Mejores resultados con un modelo de clasificación para múltiples etiquetas y transcripciones de GTS.
- Los resultados cuantitativos no descartan totalmente el uso de DS.



Comparación de promedios de exactitud de etiquetas múltiples con diferentes sets de datos.

Texto-a-etiquetas: evaluación cualitativa (un ejemplo)

Materias asignadas por los facilitadores (24):

Achievements and Awards, Architecture, Children, Christmas, Community Businesses, Community History, Corner Stores, Death, Earliest Memories, Engagements, Farm Life, First Job, First Meetings, Marriage, Neighborhood Life, Parents, Schools, Siblings, Town Life, Weddings, Workday Life

Top 10 predicciones con DS – **iguales en todas las entrevistas!!**

Parents, Schools, Workday Life, Children, Siblings, Immigration Stories, Marriage, Town Life, Achievements and Awards, Identity

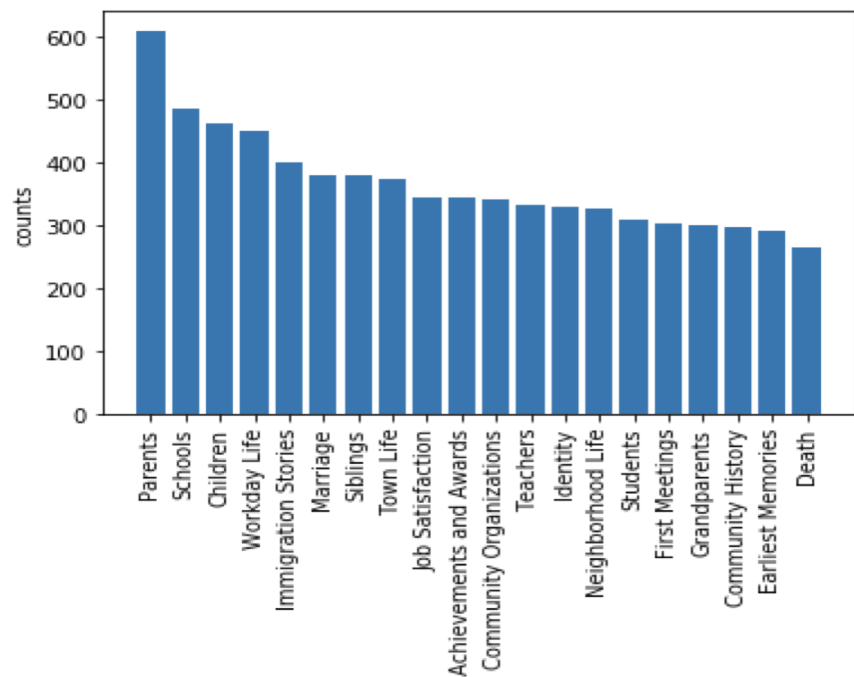
Top 10 predicciones con GST

*Parents, Children, Marriage, Siblings, Workday Life, Schools, Town Life, **Achievements and Awards**, Earliest Memories, Death*

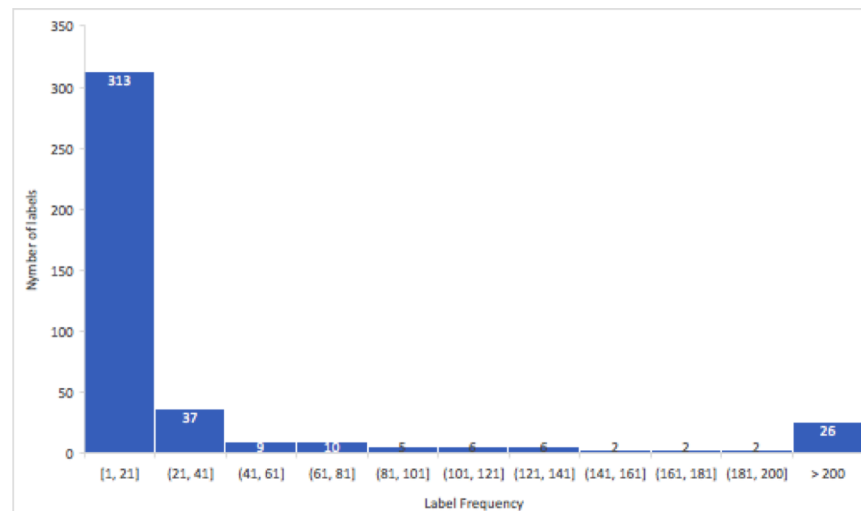
Es posible predecir etiquetas de transcripciones automáticas y de materias existentes (metadatos), pero estas serán tan buenas o malas como las seleccionadas.

Evaluación archivística

- Contenido y distribución de las top 20 materias.
- La mayoría de las materias se aplicaron a pocas entrevistas.
- Distribución inconsistente a través del corpus.
- Superposición y redundancia.
- Entender como se asignan materias para mejorar el diseño de los datos de entrenamiento.



Representación
de contenido



Cobertura

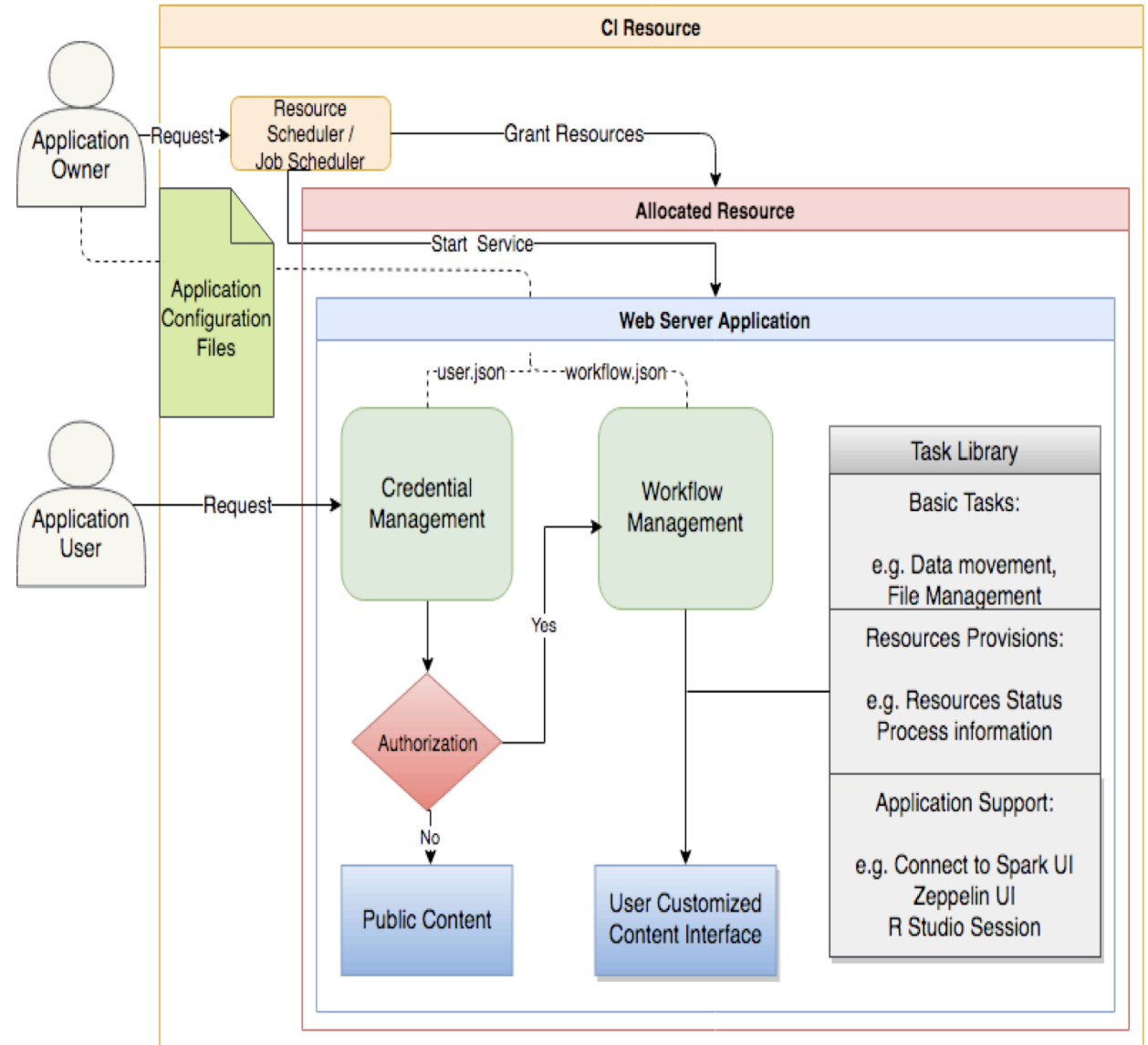
Las estadísticas nos ayudan a entender nuestras prácticas.

Evaluación archivística

- Dos ejercicios para comprender:
 - Consistencia en la asignación de materias:
 - Profesionales que seleccionan materias para un mismo contenido difieren en criterio.
 - Diversidad de materias para representar un corpus:
 - Listas de materias curadas y consensuadas.
 - Balance entre descripción general y detallada.
- Diseñar materias para entrenamiento de modelos:
 - Curar vocabularios controlados.
 - Uso de relaciones jerárquicas entre términos amplios y detallados.
 - Mejorar la distribución de materias implementando la estrategia de comienzo, medio y final de la entrevista.
 - Entrelazar estrategias con evaluación.

IDOLS

- Aplicación web reconfigurable a demanda.
- Interfaz entre los profesionales de archivo y los recursos computacionales remotos.
- Pueden ver el demo cuando se registren [aquí](#)



Interfaces para transparencia y reproducibilidad

Workflow Management

Define Task Predecessor

Task 1

None

Collapse Step 1

Task 2

Task 3

Collapse Step 2

Task 3

None

Collapse Step 3

Task 4

None

Collapse Step 4

Task 5

None

Collapse Step 5

Task 6

None

Collapse Step 6

Run All Tasks

Step 1: Show Audio

Description: Select audio to listen to

Define root directory to start exploring

/home/ldols/resources/audio_data/human_transcribed_audio_0

Expand Directory

Expand JSON as Directory

Refresh Directory

File Path

/home/ldols/resources/audio_data/human_transcribed_audio/0

Group 1

Group 2

Group 3

Group 4

Group 5

Group 6

Group 7

Group 8

Group 9

mbyo06228_mono_16khz.WAV

MBY006244_mono_16khz.WAV

MBY006255_mono_16khz.WAV

MBY006272_mono_16khz.WAV

MBY006277_mono_16khz.WAV

MBY006309_mono_16khz.WAV

MBY006347_mono_16khz.WAV

MBY006380_mono_16khz.WAV

MBY006388_mono_16khz.WAV

Go to parent directory

Expand chosen directory

Show

0:00 / 0:00

Download

Step 2: Run Speech Recognition Script

Description: Driving script of the speech recognition. Edit audio_name to the name of the audio file you would like to recognize

audio_name

MBY006309_mono_16khz.WAV

Executable

/home/ldols/resources/run3_idols.sh

Show contents

Show

```
#!/bin/sh

# mi python3 cuda9.0 cudnn7.0

# if [[ ! "$PYTHONPATH" =~ (^|:)/work/03076/huang/maverick2/.local/lib/python3.6/site-packages/ ]]; then

# export PYTHONPATH="/work/03076/huang/maverick2/.local/lib/python3.6/site-packages:$PYTHONPATH"

# fi

# while [ -f "/tmp/myscript.lock" ]; do
# sleep `awk -v min=10 -v max=20 'BEGIN{srand(); print int(min+rand()*(max-min+1))'`
# done
# touch /tmp/myscript.lock
FILENAME=MBY006309_mono_16khz.WAV
FILEPATH="/home/ldols/resources/audio_data/MBY006309_mono_16khz.WAV"
FILESIZE=$(stat -c%s "$FILEPATH")
resultfile="/home/ldols/resources/audio_transcription/MBY006309_mono_16khz.WAV.txt"
if [ -f "$resultfile" ]; then
```

Save edits to file and/or Run script in the text area above

Save file

Run script

MBY006558_idols.txt

MBY006603_idols.txt

MBY006615_idols.txt

MBY006645_idols.txt

MBY006648_idols.txt

MBY006659_idols.txt

Go to parent directory

Expand chosen directory

Hadoop File System

Top

10 rows

Show Contents

Show

Download

Results:

well greet londr ye bicamer viceroy month shilpit five year old lay golden twenty first twenty ten abaturino mexico right year paniculata centr two conduct interview dear old dad rose marr loedoes labor rosemarin ramiro born nverton niteen twenty eight make sixti two year plu breadlin let across son victoria love much well glad make jon father day lot time togeth last coupl day chanc compar note want go histori book away lover histori red marward think record librori card cite wonder thing love event better well hope get weath hear whole bunch thing top bottom know spirit particular niseit bob histori project thought would help listen friend famli folk studi futur get call bollerpi histori stuff belt impatcular like hear talk litti bit self identifi mean coverag erat okay culican excetti gener ever like nation patricio familiar hottest come accept call day like come south west panic came oppos anti barrn oh see neither historiian additilan tell ator want go term later haard comment use one pelican litti right paper big differ betun panic set questi panic heast north east panic know get favorti topic know mexico long talk better belong tool place spanish connecten cour mexican tricki let mexican american new mexican come talk albertopoli chang buri letter cornetico mexico dominica becam part countri union telesteal macavoy gener away strang spoke differ differ fort problem rate small bare popul still pay identifi ross wonder use wonder dustranivel come peopl think bit deliver understand take fred mexican tacitu ri know werow done mexican read love love speak spanish write spanish first languag spoke mother would unthink bespeak master brother would unthink speak father older peopl andi anoth bodi neighborhood grew overbult school liner engienderinnen lee read hallow lot read mostti daylight get spanish read pick pick read much read read boii usg made one keep know hope disappoint perfectti bilingu brought us appreci love spanish ah enough get hereto understand complet also know andi dynam ingent languag ah reason ate kept leppert big deat profit car tell truth reader like read english like read classic also like read much aval delecti bring home eve get hold bit use bring spanish languag comeboo turandott great number ego ape pickup copli rit abi first keep rise isolanth know hear talk engouement favoriti topic volitar rol quantin right ye kin tell bistori relay hear betray born metronym an old famli came on true att ishmailt walk away ye ye ye left tabareau mexico bite cold minion begin depress like work hand ochet father sat live half dozen differ grace autopetrip read find settlement also close zoo great glass cold first school capicco pransconen call acnapura spanish languag none sea good word never knew much school learn benefactor first year batonica school went north public school finish ah earli memori come big tabl order brother sister ah capac pleas beneficiari go giver time mother father nephew idea live away fate one first come crook use drop everi week everi week hat chanc good memori come three o'clock bonaday mother better happi see im drossast particulati good cousin male father love sing still rememb arm long paharia go tatatatatatatatata rotherwood saw look ariano forth booi dotoobuu nevi nemoth differenti sank talaiporon rode better yet cano comiak ravish use learn good got dose well know reali tell makin sun pedro bit continu neighborhood grew know critia call dorothi let hope one last night becam poor editori peopl litti well seringueira ensue poorer peopl set peopl knew use refer self live come flat spoke shapley wear totlia call delawar call tortilia movi dunderbunk sunbreak rot prestrat usm boy othello requir know talk littest thu opinion totin appoint explain desardrou came mangat mother roar born rear sober innel south side river across sedentario sudden thought claim far conrad halt extraordinari hotel man born rale anatolia father earli age ten erecidoun mind two mile incarco maint around tnoaset himinbrot tydon erybodi went work nineteen twenti five might gave mother father go druellett day year reason know day reach read newspap abbey would came altuanguero hanothan memori dread notori next thing know depress child yether ressorti thing art first nicolin benight thirti three probat still one gener kept eye ground found found empti pan tabul bipedular got idea bodi said boot leverag retaliatori feverheat nickel would reprob nice took oultit booster saw come redli show five cant also said relaton money make newspap tuse feast found way go town sell newspap easi us grow andistene dlogtoun go readi be day ceterera way sold paper through noon addit onarti good sailor start sell cizate line gaid start real way order everybod genial way year reason know day reach read newspap abbey work spade civil war mainwar later clicket cousin read fi berstain e good abistatt guid rufir write retoul boss man gave scosigo week springhale leadin job pretti seriou deavast pay mother found good excus scotamen two week rill father power run natset never nouth colder nasett weak pension sultan lee hundberasont hors use drive rill still cold took rubin school feast buriu cold back use take boat frepki kennebacke senur forth one

original_label_fs	original_label_g	predictions: Fixed Subjects label	predictions: score	predictions: Fixed Subjects label	predictions: score	predictions: Fixed Subjects label
Community,Achievements and wards,Community History,Community Organizations,Discrimination,Racism,Urban Life,Workday Life,Identity	community worthies,craft, skills, and procedures,culture,Hispanic Chamber of Commerce,historical events/people,history,Houston, TX,influential people,Latino,Latino Learning Center,League of United American Citizens,LULAC,MALDF,memories of former times,mentor,Mexican American Legal Defense Fund,personal experiences,scholarship,social beliefs and practices,sterotypes	Workday Life	0.58	ts and Awards	0.6	Workday Life

Showing 1 to 1 of 1 entries

submit form

Previous

1

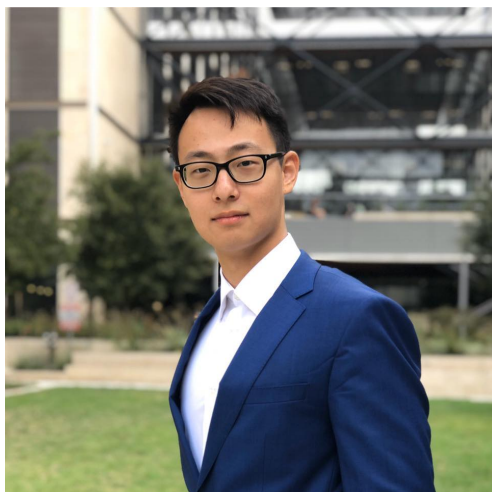
Next

Conclusiones

- Nuestra investigación continua.
- Los registros digitales necesitan ser abordados como tales.
 - Los usuarios demandan otras soluciones para los archivos.
- Nuevas posibilidades y roles para desarrollar trabajos interdisciplinarios.
- Rol de investigación, basado en hipótesis basadas en teorías de archivo que pueden ser comprobadas.
 - Queda mucho por explorar.
 - Trabajo continuo.
- Incorporar el método de evaluar nuestros procesos.
- Haciendo y aprendiendo.
- Esta presentación se basa en:

Xu, Esteva, Cui, Castillo, Want, Hopkins, Clement, Choate & Huang (2020) A Study of Spoken Audio Processing Using Machine Learning for Libraries, Archives and Museums. CAS Workshop, IEEE Big Data Conference 2020. In print

Gracias, comentarios y preguntas



Peter Cui
Hanna-Robbins Hopkins
EJ. Castillo
Aaron Choate
Weijia Xu
Tanya Clements



Agradecemos a StoryCorps y a la Benson Latin American Collection por proveer los datos utilizados en este estudio. Notamos los recursos de computación provistos por TACC. Este estudio fue financiado por UT Austin Good Systems en UT Austin, y por la Fundacion Nacional de Ciencias de USA (# 1726816).