machinalis

Information Extraction: herramientas y visualización, todo con IEPY

Javier Mansilla

Machinalis

jmansilla@machinalis.com

21 de Mayo, 2015

Quien soy

machinalis

- Javier Mansilla.
- Analista en Computación.
- Programador desde hace varios años, y aprendiz en Machine Learning y NLP.
- Co-Fundador de Machinalis.
- Participante activo en el desarrollo de IEPY.

Sumario

machinalis

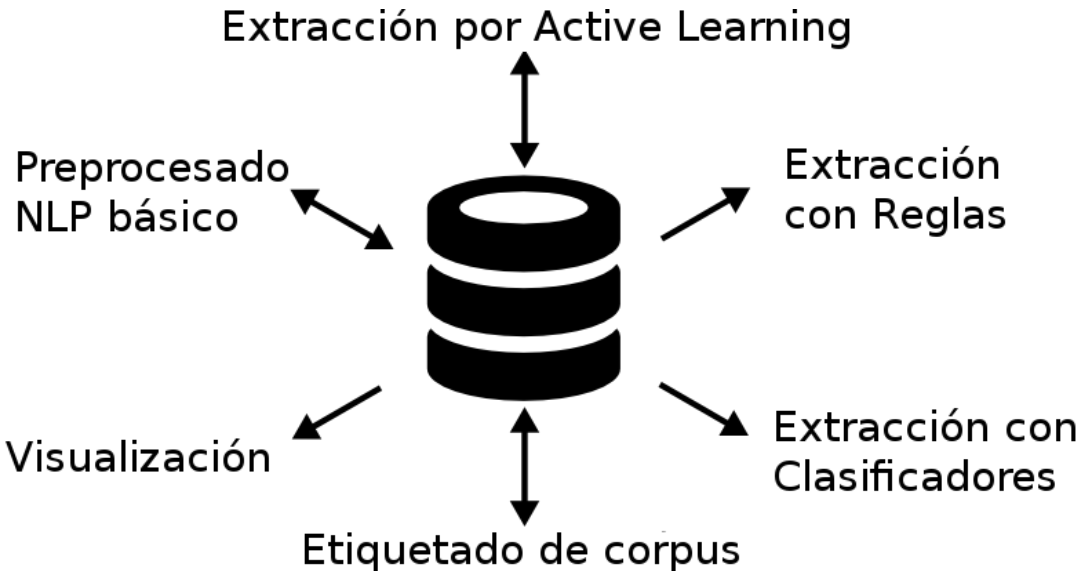
¿Qué es IEPY?

machinalis

- Una herramienta para *extracción de información*, y NLP en general
- Funciona out-of-the-box
- Provee una plataforma para prototipado rápido en extracciones:
  - Basado en Reglas.
  - Basado en Clasificadores.
  - Active Learning.
- Software Libre (BSD).
- Escrito en Python 3.
- Testing integrado al desarrollo
- Pasen y vean: github, issue tracker, readthedocs, etc.
- `https://github.com/machinalis/iepy`

Arquitectura

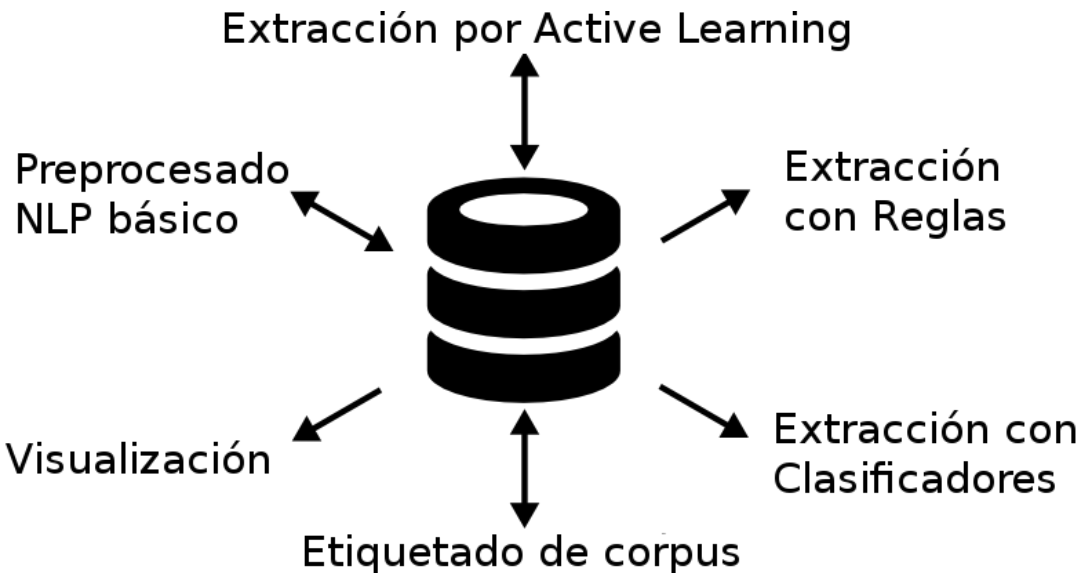
machinalis



- Diseño en estrella rodeando una BD relacional, lo que permite
  - Modularización
  - Paralelización de tareas
- Django ORM para manipulación de los datos
- Django Framework para la interfaz visual (web UI).

Arquitectura

machinalis



- Se ingresan los documentos a trabajar como texto plano
- Procesamiento NLP con correr 1 comando

Ejemplo pre procesamiento

machinalis

Lionel Andrés Messi Cuccittini (born 24 June 1987) is a professional footballer who plays for FC Barcelona and the Argentina national team. He is a forward and serves as captain for Argentina.

Ingreso de documentos

Ejemplo pre procesamiento

machinalis

Lionel Andrés Messi Cuccittini (born 24 June 1987) is a professional footballer who plays for FC Barcelona and the Argentina national team.

He is a forward and serves as captain for Argentina.

Separado en oraciones

Ejemplo pre procesamiento

machinalis

Lionel	Andrés	Messi	Cuccittini	(	born	24	June	1987	)	is	a

professional	footballer	who	plays	for	FC	Barcelona	and	the	Argentina	national	team	.

He	is	a	forward	and	serves	as	captain	for	Argentina	.

Tokenizado

Ejemplo pre procesamiento

machinalis

Lionel	Andrés	Messi	Cuccittini	(	born	24	June	1987	)	is	a
NNP	NNP	NNP	NNP	LRB	VBN	CD	NNP	CD	RRB	VBZ	DT

professional	footballer	who	plays	for	FC	Barcelona	and	the	Argentina	national	team	.
JJ	NN	WP	VBZ	IN	NNP	NNP	CC	DT	NNP	JJ	NN	.

He	is	a	forward	and	serves	as	captain	for	Argentina	.
PRP	VBZ	DT	RB	CC	VBZ	IN	NN	IN	NNP	.

Etiquetado gramatical / POS Tagging

Ejemplo pre procesamiento

machinalis

Lionel	Andrés	Messi	Cuccittini	(	born	24	June	1987	)	is	a
NNP	NNP	NNP	NNP	LRB	VBN	CD	NNP	CD	RRB	VBZ	DT
Lionel	Andrés	Messi	Cuccittini	-lrb-	be	24	June	1987	-rrb-	be	a

professional	footballer	who	plays	for	FC	Barcelona	and	the	Argentina	national	team	.
JJ	NN	WP	VBZ	IN	NNP	NNP	CC	DT	NNP	JJ	NN	.
professional	footballer	who	play	for	FC	Barcelona	and	the	Argentina	national	team	.

He	is	a	forward	and	serves	as	captain	for	Argentina	.
PRP	VBZ	DT	RB	CC	VBZ	IN	NN	IN	NNP	.
he	be	a	forward	and	serve	as	captain	for	Argentina	.

Lematización

Ejemplo pre procesamiento

machinalis

Lionel	Andrés	Messi	Cuccittini	(	born	24	June	1987	)	is	a
NNP	NNP	NNP	NNP	LRB	VBN	CD	NNP	CD	RRB	VBZ	DT
Lionel	Andrés	Messi	Cuccittini	-lrb-	be	24	June	1987	-rrb-	be	a
P	P	P	P			D	D	D			

professional	footballer	who	plays	for	FC	Barcelona	and	the	Argentina	national	team	.
JJ	NN	WP	VBZ	IN	NNP	NNP	CC	DT	NNP	JJ	NN	.
professional	footballer	who	play	for	FC	Barcelona	and	the	Argentina	national	team	.
	P				O	O			L			

He	is	a	forward	and	serves	as	captain	for	Argentina	.
PRP	VBZ	DT	RB	CC	VBZ	IN	NN	IN	NNP	.
he	be	a	forward	and	serve	as	captain	for	Argentina	.
P									L	

Detección de Entidades y resolución co-referencias

Ejemplo pre procesamiento

machinalis

Demo

Etiquetando Relaciones

machinalis

- IEPY tiene integrada una herramienta para etiquetar corpus.
- Interfaz Web.
- Permite múltiples *jueces* simultaneos.

Etiquetando Relaciones

machinalis

- IEPY tiene integrada una herramienta para etiquetar corpus.
- Interfaz Web.
- Permite múltiples *jueces* simultaneos.

Demo!

Etiquetando Relaciones

machinalis

- IEPY tiene integrada una herramienta para etiquetar corpus.
- Interfaz Web.
- Permite múltiples *jueces* simultaneos.

Demo!

Productividad de 50 documentos por hora-hombre

Reglas

machinalis

```
Subject + Token(", born") + Object + anything
```

- Motor de extracción basado en reglas
- Permite reglas positivas y también negativas
- Expresiones regulares enriquecidas
- Tarea: Dado un contexto, decidir si se encuentra la relación o no

Reglas

machinalis

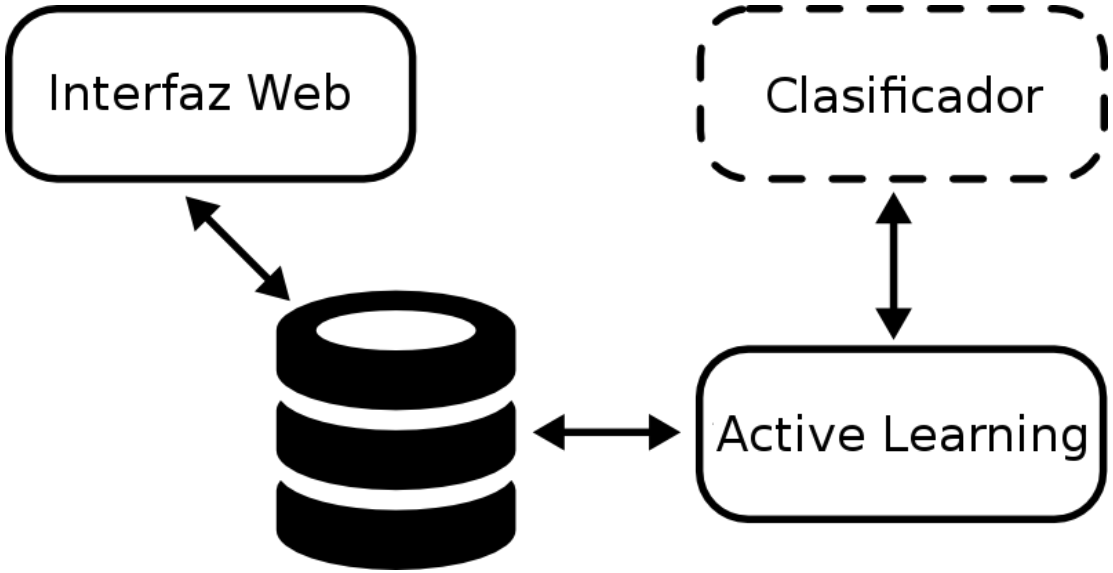
```
Subject + Token(", born") + Object + anything
```

- Motor de extracción basado en reglas
- Permite reglas positivas y también negativas
- Expresiones regulares enriquecidas
- Tarea: Dado un contexto, decidir si se encuentra la relación o no

Ver un contexto

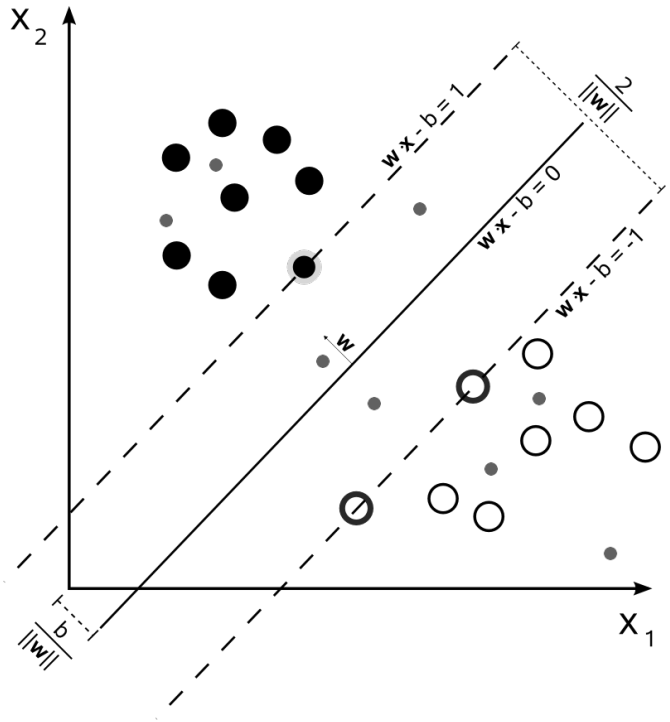
Extraccion de Información con Clasificadores

machinalis



- Tarea del clasificador: dado un contexto, decidir si o no
- Personalizar y probar: Un clasificador básico tiene 13 lineas (`sklearn`)
- Un clasificador es sólo un módulo Python. La única conexión con IEPY es el tipo de datos a clasificar
- El clasificador por defecto es el resultado de 1 año de trabajo.
- Opcionalmente, puede trabajarse guiado por *active learning*

Active learning

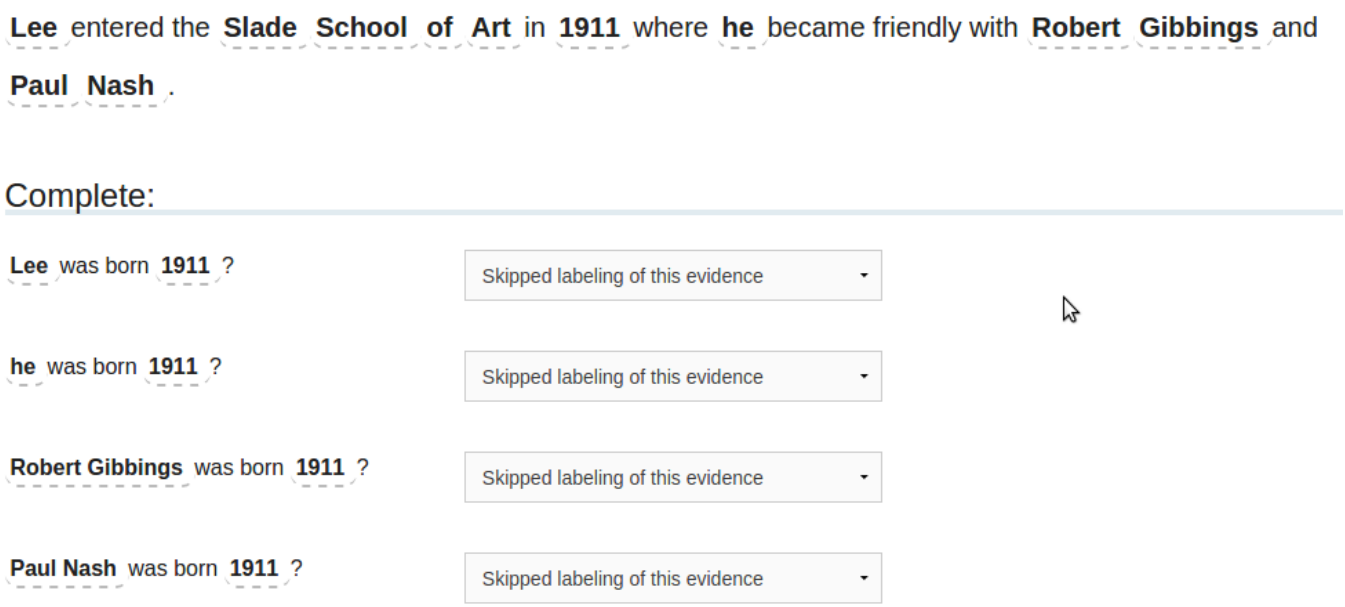


Enfoque simple que permite trabajar con un esquema de *active learning*

Pedirle al clasificador los puntos más dudosos de clasificar.

Active learning

machinalis



Interfaz para Active Learning

Múltiples colaboradores pueden contestar en simultáneo.

Repaso final

machinalis

- Herramienta open source para IE
- Construcción de corpus colaborativa
- Motor de IE basado en reglas
- Motor de IE basado en clasificadores completamente *hackable*
- Out-of-the-box active learning

Ejemplos reales

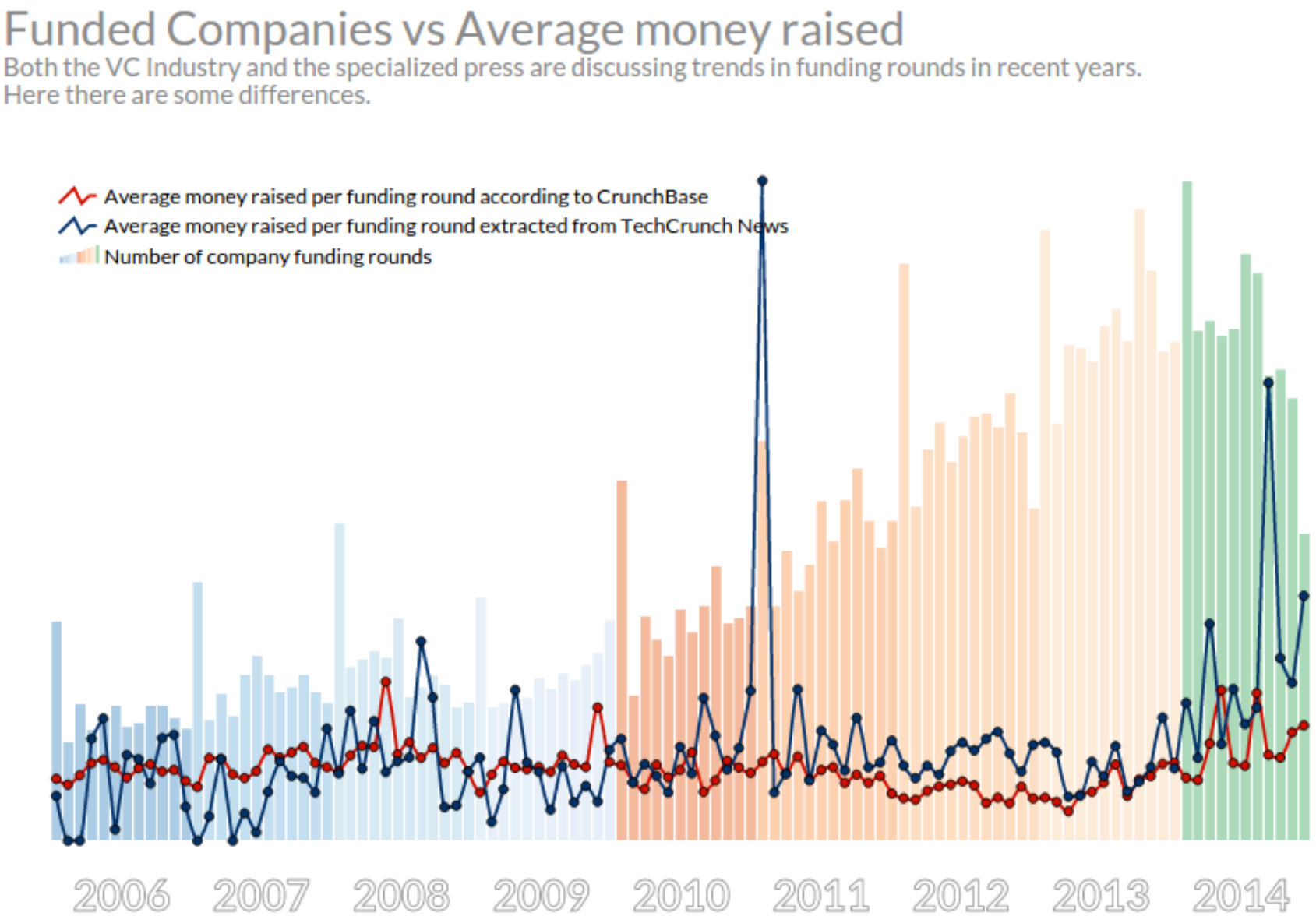
machinalis

Software cordobés para procesar documentos de la dictadura

Investigadores de Famaf elaboran novedosas herramientas informáticas que facilitan el trabajo del Archivo Provincial de la Memoria.

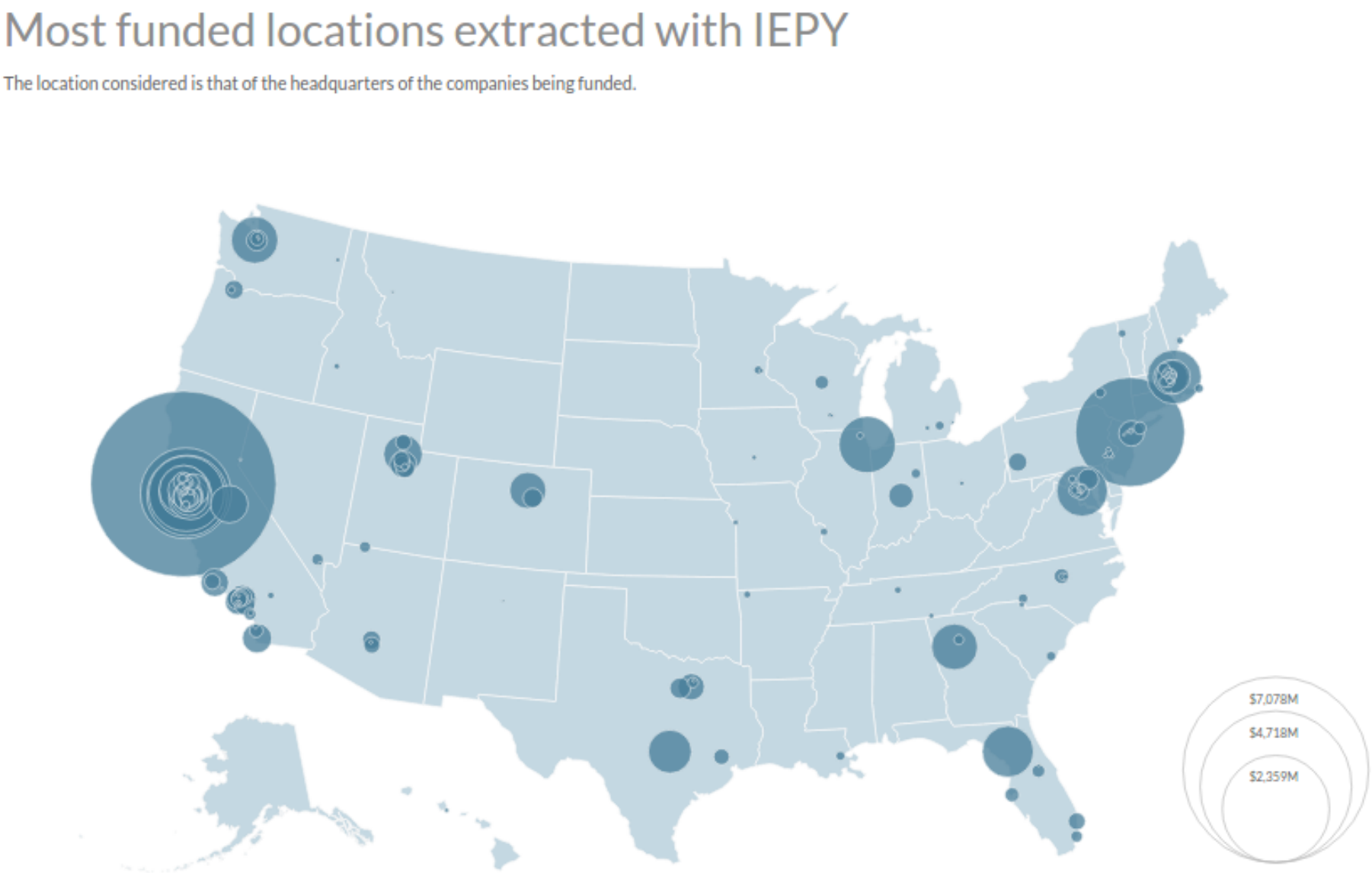
Ejemplos reales

machinalis



Ejemplos reales

machinalis



Fin

machinalis

¡Muchas gracias!

(¿preguntas, comentarios, sugerencias...?)

Custom features

machinalis

Información disponible en un contexto

- Tokens
- POS tags
- Todo lo referido a entidades presentes
- Parse tree

Custom features

machinalis

Información disponible en un contexto

- Tokens
- POS tags
- Todo lo referido a entidades presentes
- Parse tree

Más...

- Actualmente usa Stanford's CoreNLP.
- Los Features y el Clasificador viven como módulos Python independientes.
- Pueden usarse reglas como nuevos features