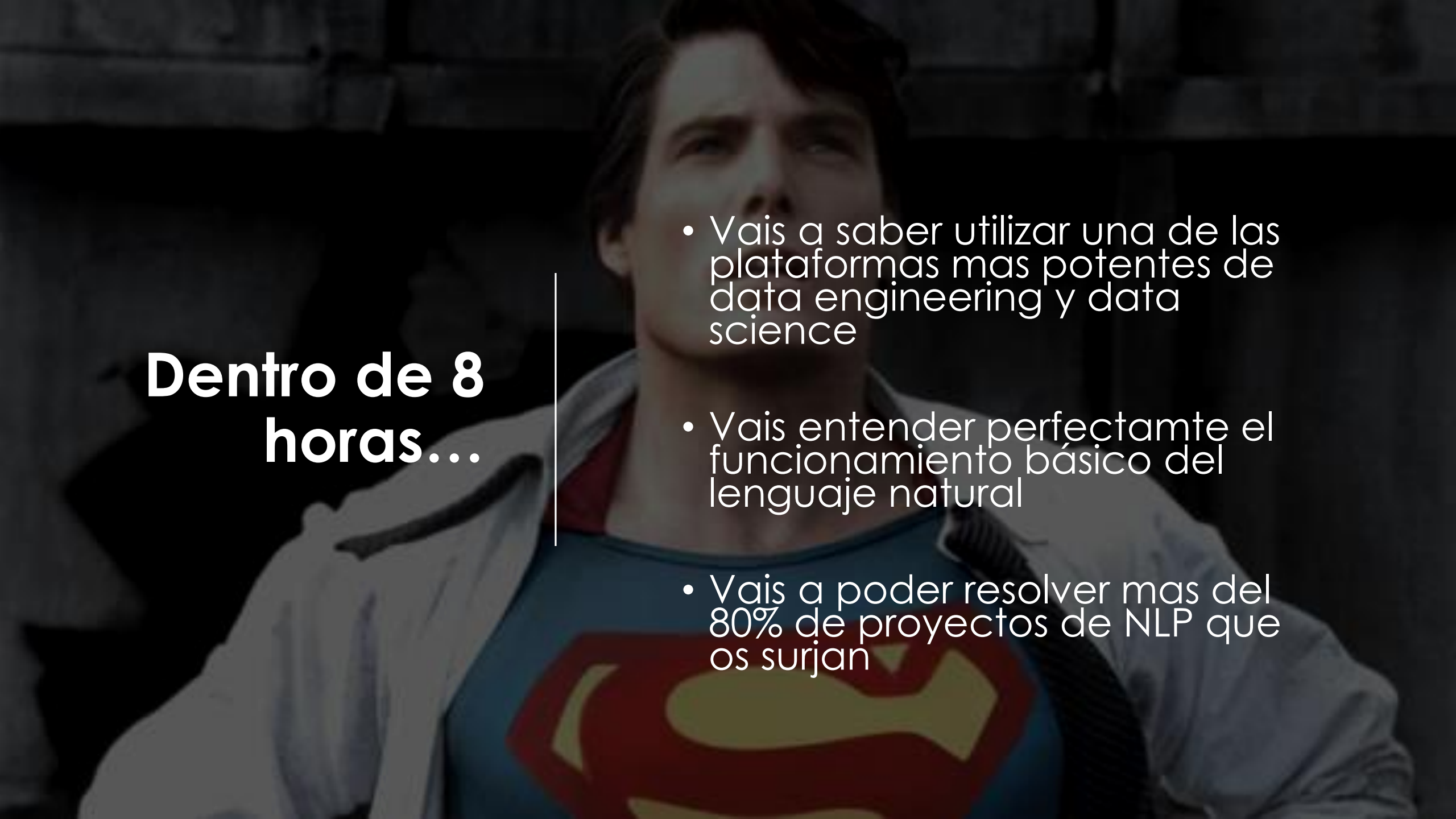
plain concepts

Rediscover
the meaning of technology

Day 3 – Databricks & NLP



Dentro de 8 horas…

- Vais a saber utilizar una de las plataformas mas potentes de data engineering y data science
- Vais entender perfectamte el funcionamiento básico del lenguaje natural
- Vais a poder resolver mas del 80% de proyectos de NLP que os surjan

Agenda

- Databricks
  - Clusters
  - Notebooks
  - Datasources
  - Jobs
- Introducción NLP
  - De palabras a vectores
  - Word Embeddings: Librerías y arquitecturas
    - Word2Vec
    - FastText
  - Hiperparametros
  - Preprocessing
- NLP: Real Cases
  - Topic Modelling
  - Text Classificatiopn
  - Information Retrieval
  - Natural Text Generation
- Real Project

plain concepts

Training

- 7 Labs:
  - 1 Lab Databricks
  - 2 Labs NLP
  - 3 Demo proyectos
  - 1 Demo Final
- 2 + 1 Formatos:
  - Intermedio
  - Avanzado 🌶️
  - Resuelto

Azure Databricks

- Que es?
- Que características tiene?
- Donde nos aporta un valor diferencial?

Azure Databricks ¿Qué es?

- Workspace
- Diseñado para data engineering y data scientist
- Computación distribuida

Azure Databricks Características

- Fuentes de datos
- Notebooks
- Clústers
- Jobs

Azure Databricks
Valor diferencial

- Fuentes de datos:

Conexión con distintas fuentes de datos distribuidas

- Notebooks

Versionado de código y conexión con git

- Clústers

Configuración Spark y Autoescalado de clúster muy sencillo

Los clusters se pueden apagar y encender

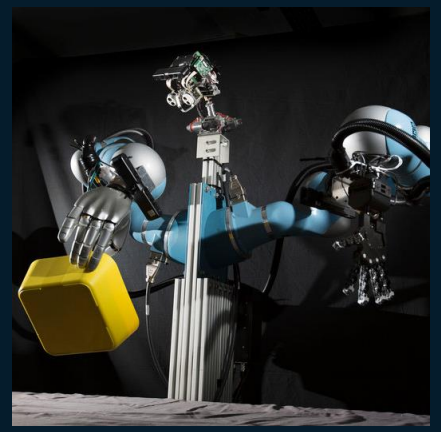
- Jobs

Muy fácil automatización de tareas. Serverless

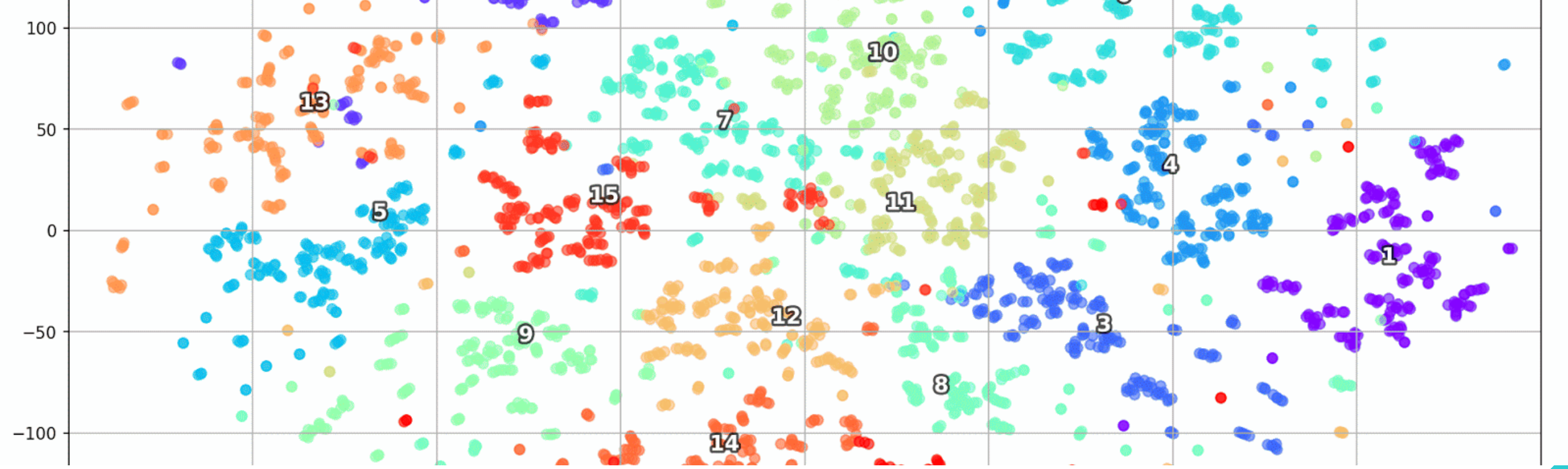
Databricks Lab

ETL

Vamos a automatizar una ETL donde leeremos de un blob, transformaremos los datos para calcular el gasto semestral y por último guardaremos en otro blob. Y automatizarlo para que se ejecute durante el desayuno ☺



plain concepts

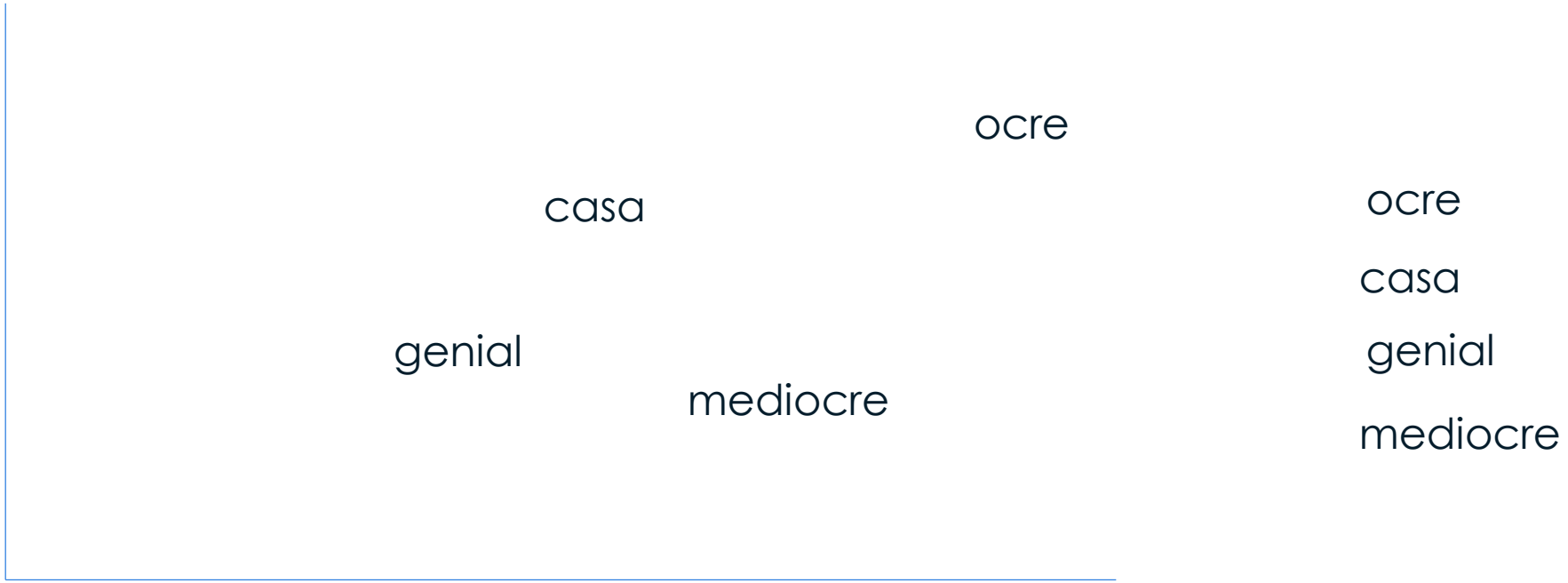


NLP

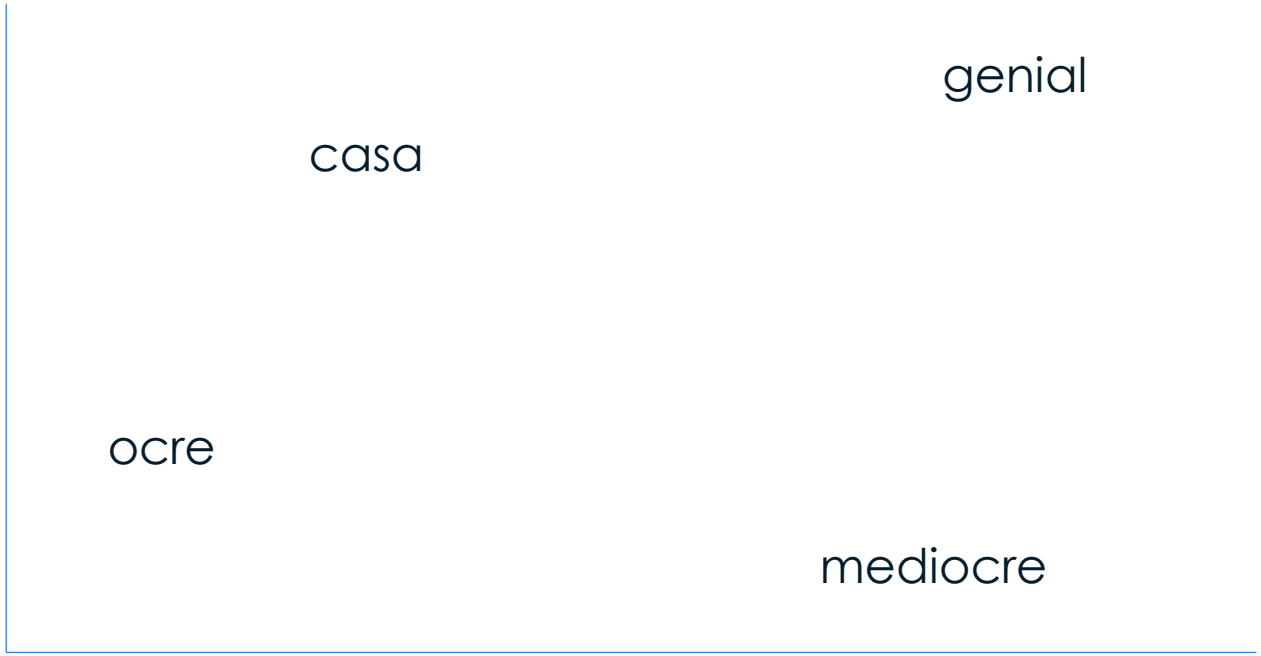
From Words to Vectors

plain concepts

From words to vectors



From words to vectors



Word embeddings

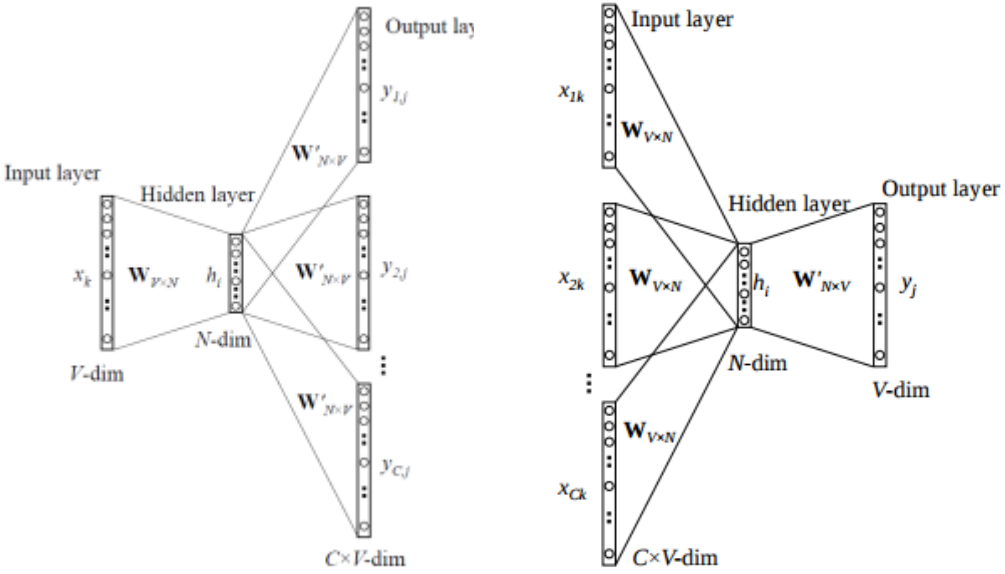
- ¿Cómo lo haces? Llevas años escuchando palabras. Los modelos igual.
- Atienden a como se relacionan las palabras entre ellas para ajustar el vector de cada una.
- Hands on practice:
  - Word2Vec
  - Fasttext

Word Embeddings

Word2Vec & Fasttext

- Redes neuronales simples. ¿DL or not DL?
- Word2Vec
  - CBOW
  - Skip-gram
- Fasttext
  - N-grams

gensim
topic modelling for humans



fastText

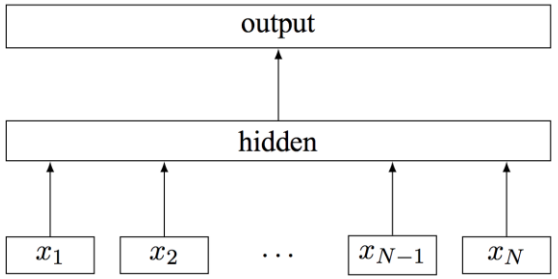


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

plain concepts

NLP Lab1

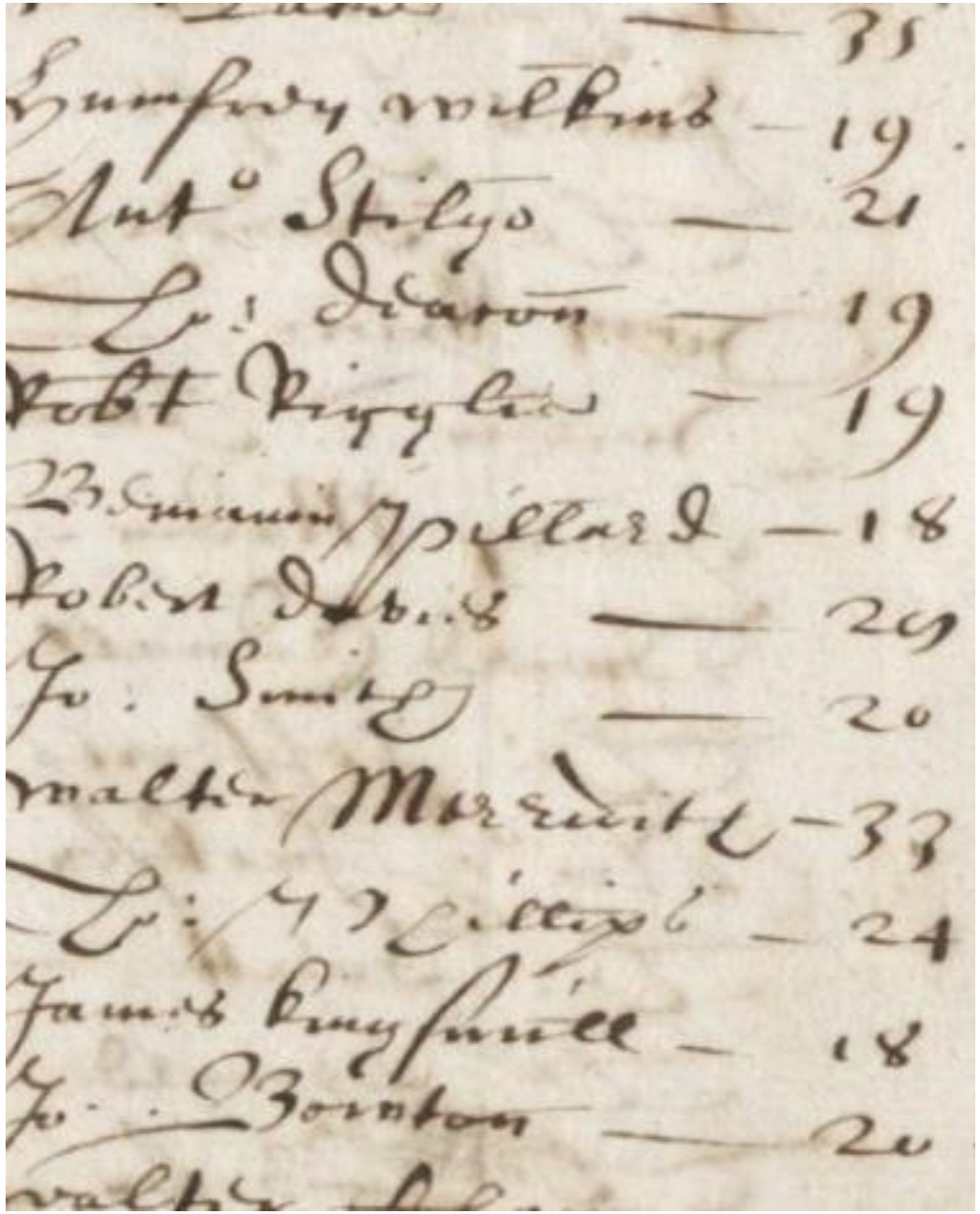
Preprocessing y Hiperparametros

Entrenar un modelo word2vec y un modelo fasttext sobre un texto y ver el vector de algunas palabras. Encontrar las palabras mas similares a una dada.

plain concepts

Hiperparámetros

- N-grams
- Vector size
- Window size
- Min count/ Max vocab/ Max count
- Parametros de la red neuronal:
  - Learninig rate
  - Decay..



Preprocessing

- Tokenize
- Standarize
- Punkt
- Stop Words
- Lemmatization
- Stemming

NLTK

spaCy

Preprocessing

Frase Original

'HEMOS RECIBIDO PETICIONES DE CLIENTES QUE NOS PIDEN PANTALONES ANCHOS CON ATADURAS EN TOBILLO.UN SALUDO .JAVIER .CABALLERO ‘

Estandarización y Eliminación signos de puntuación

'hemos recibido peticiones de clientes que nos piden pantalones anchos con ataduras en tobillo un saludo javier caballero’

Eliminación stop words

'recibido peticiones clientes piden pantalones anchos ataduras tobillo saludo javier caballero’

Stemming

'recib peticion client pid pantalon anchos atadur tobill salud javi caballer'

Vocabulary Expansion

Utilizar modelos preentrenados como base y reentrenar sobre mi vocabulario específico.

Inglés:

Wikipedia

Google News 300

Español:

Wikipedia 100

NLP Lab2

Preprocessing y Hiperparametros

Conseguir preprocesar el texto y ajustar los hiperparametros para obtener un buen resultado con el fragmento del Quijote.

Vocabulary Expansion

Repetir el proceso expandiendo el vocabulario con la Wikipedia en español y realizar las siguientes operaciones con vectores:

Caballero – Hidalgo + locura

Rocin + Quijote

plain concepts

NLP Recap

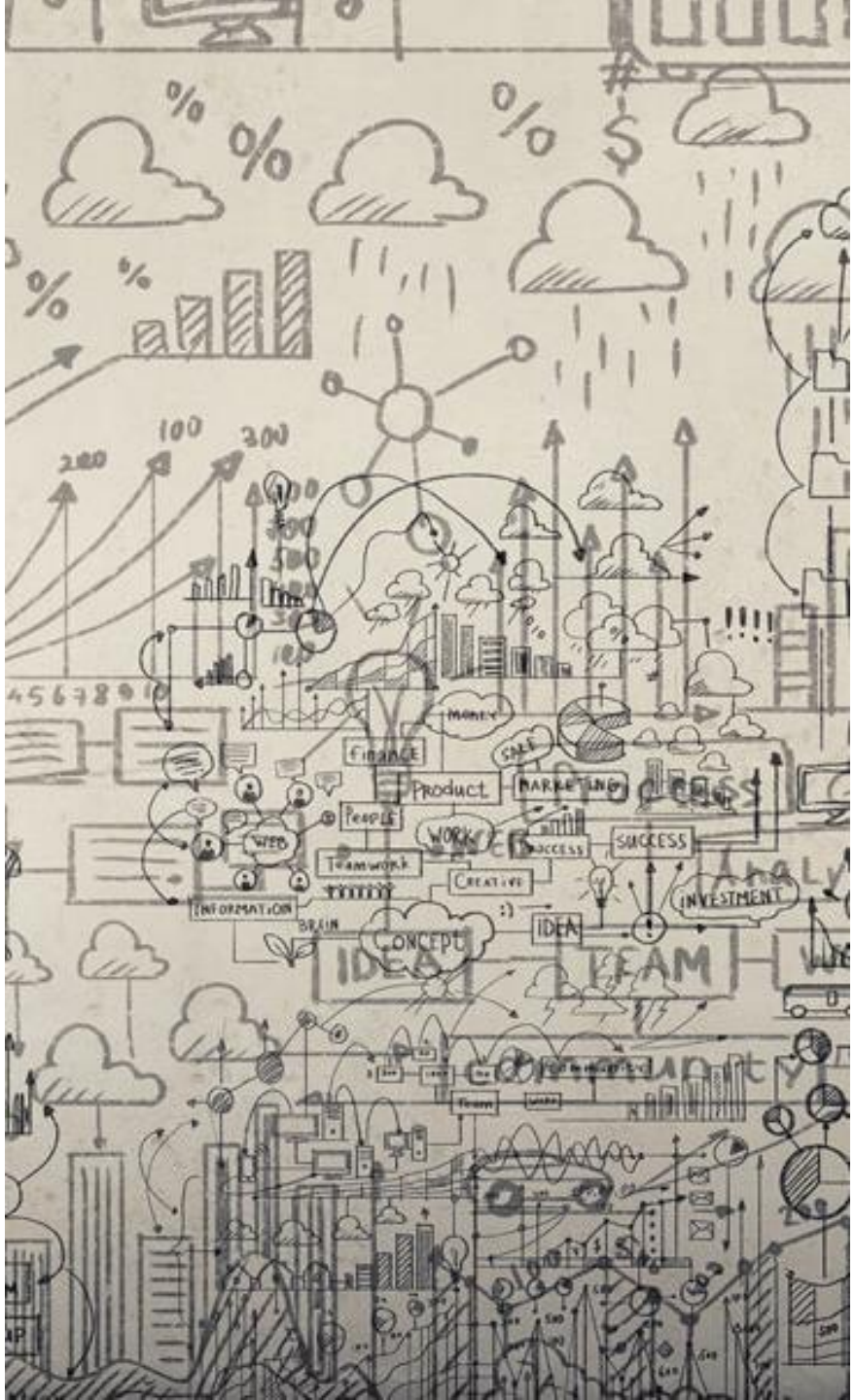
Word embeddings

Preprocessing

Vocabulary expansión

Que podemos hacer con ello?

plain concepts



What's next

- Textos: Secuencias de palabras
- Problemas más típicos con textos
  - Clasificación de texto
  - Topic Modelling
  - Information Retrieval
  - Generación de texto

plain concepts

BBC News

- 2225 Noticias
- 5 Posibles categorías:
  Sports, Tech, Politics, Business, Entertainment

plain concepts

Text classification

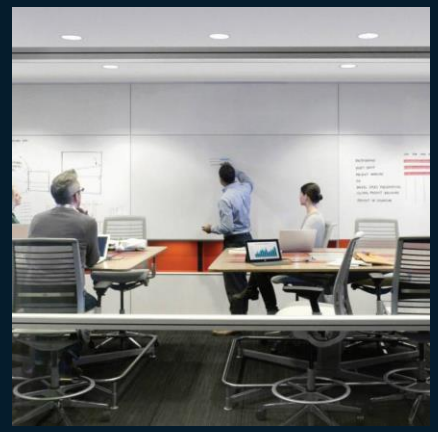
- Aprendizaje: Supervisado
- Modelo: Capa clasificación al final de la red, debe decidir entre las posibles categorías
- Preprocesado:
  - Estandarización
  - Stopwords
  - Puntuación
  - Tokenización
  - Stemming/Lemma
- Librerías:
  - Fasttext

plain concepts

Text classification Lab

BBC News

Vamos a clasificar titulares de noticias de BBC. Para ello encontraremos un dataset en blob donde están cada noticia y su categoría: Sports, Bussines, Politics, Entertainment or tech.



plain concepts

Topic Modelling

- Aprendizaje: No supervisado
- Modelo: Asume que un topico está compuesto por un subconjunto de palabras y cada documento se puede descomponer en probabilidades de cada tópico.
- Preprocesado:
  - Estandarización
  - Stopwords
  - Puntuación
  - Tokenización
  - Stemming/Lemma
- Librerías:
  - Gensim/Scikit
  - pyLDAvis

plain concepts

Topic Modelling Lab

BBC News

Vamos a clasificar noticias de BBC. Para ello encontraremos un dataset en blob donde están cada noticia y su categoría: Sports, Bussines, Politics, Entertainment or tech.

plain concepts

Information Retrieval

- Aprendizaje: No supervisado
- Modelo: Encontrar los documentos mas cercanos en base a las distancias entre los vectores de cada documento.
- Preprocesado:
    - Estandarización
    - Stopwords
    - Puntuación
    - Tokenización
    - Stemming/Lemma
- Librerías:
    - Gensim/Fasttext
    - Tf-idf (scikit)

Information Retrieval Lab

BBC News

Vamos a encontrar los titulares de noticias de BBC mas cercanos a una query que le pasemos. Para ello analizaremos los textos de BBC entrenando el modelo sobre ellos y procesaremos la query para encontrar los que mas se acerquen. Lo haremos de dos formas por coincidencia de términos o por significado semántico.

plain concepts

Text Generation

- Aprendizaje: No supervisado
- Modelo: Encoder-Decoder. RNNs, Transformers. El encoder aprende a hablar y el decoder aprende a transformar la salida del encoder a lo que queramos.
- Preprocesado:
  - Tokenización
- Librerías (DL):
  - Tensorflow
  - PyTorch

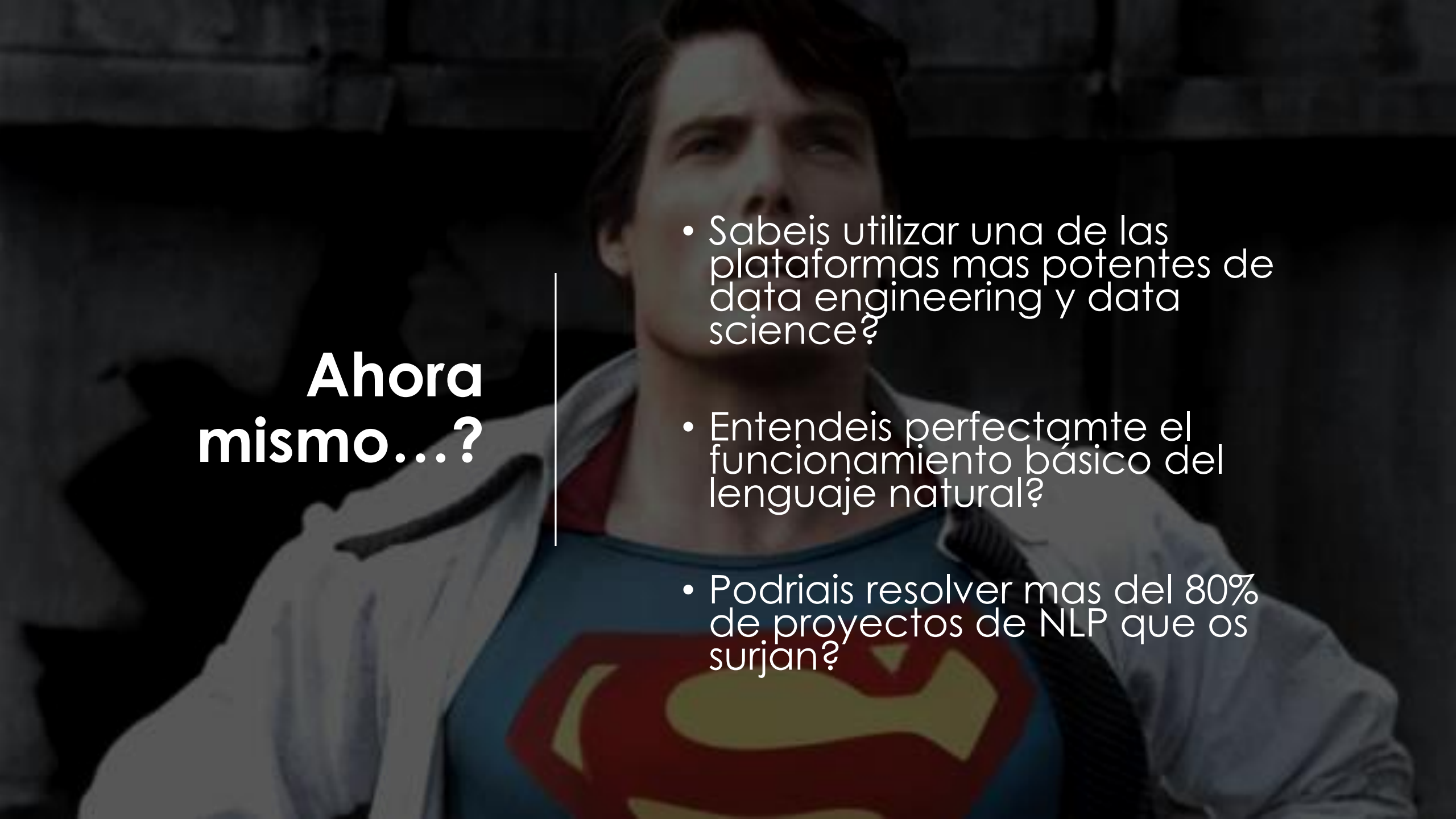
plain concepts

Final Lab end2end

BBC News

Vamos a desarrollar un clasificador de texto en tensorflow y keras sobre el dataset de noticias. El entrenamiento lo haremos en Azure ML worksace, registraremos el modelo y generaremos un pipeline de predicción con Databricks y Azure ML Services.

plain concepts



Ahora mismo…?

- Sabeis utilizar una de las plataformas mas potentes de data engineering y data science?
- Entendeis perfectamte el funcionamiento básico del lenguaje natural?
- Podriais resolver mas del 80% de proyectos de NLP que os surjan?

plain concepts

Rediscover
the meaning of technology

MADRID

BARCELONA

BILBAO

SEVILLA

LEÓN

CORUÑA

LONDON

FRANKFURT

www.plainconcepts.com

For further information

info@plainconcepts.com

AMSTERDAM

SEATTLE

DUBAI