ACÁMICA

TEMA DEL DÍA

Preprocesamiento del lenguaje natural

Relativamente fácil para los humanos, no tanto para las computadoras. Aunque tal vez eso esté cerca de cambiar.

Hoy, te mostraremos algunas herramientas clásicas y algunas de las que están por venir.

Agenda

Daily

Explicación: TF-IDF.

Break

Hands-on training

Proyecto

Cierre

Daily

Daily

Sincronizando...

Bitácora

¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

Challenge

¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

Term Frequency - Inverse Document Frequency

Vectorizar • Bag of words con N-gramas

Problema: hay palabras que cobran sentido cuando se las agrupa con otras, ejemplos: “Plaza Italia” y “Control Remoto” .

Vectorizar • Bag of words con N-gramas

Solución

Además de cada palabra por separado, agregamos los grupos de 2 (ó N) palabras contiguas a nuestro vector de Features.

Para implementar esto usando CountVectorizer:

```
CountVectorizer(max_features=max_features, stop_words="english" , ngram_range=(1, 2))
```

Vectorizar • Bag of words con N-gramas

Solución

Además de cada palabra por separado, agregamos los grupos de 2 (ó N) palabras contiguas a nuestro vector de Features.

Para implementar esto usando CountVectorizer:

```
CountVectorizer(max_features=max_features, stop_words="english" , ngram_range=(1, 2))
```

Ojo con la cantidad de Features

Vectorizar • TF - IDF

Observación: si buscamos diferenciar cada documento por las palabras que lo componen, las palabras que están en todos ellos no aportan información.

Vectorizar • TF - IDF

Observación: si buscamos diferenciar cada documento por las palabras que lo componen, las palabras que están en todos ellos no aportan información.

Idea: hay que medir no sólo cuanto aparece una palabra en una instancia (documento), sino también qué tan frecuente es esa palabra en todo el corpus.

Term Frequency - Inverse Document Frequency

TF - IDF

Vectorizar • TF - IDF

Term Frequency

Frecuencia de una palabra (*term*) en una instancia o documento (*doc*).

$$\textbf{TF}(term, doc) = \frac{\#\text{ de veces que el } term \text{ aparece en el } doc}{\#\text{ de } terms \text{ diferentes en el } doc}$$

Vectorizar • TF - IDF

Term Frequency

Frecuencia de una palabra (*term*) en una instancia o documento (*doc*).

$$\textbf{TF}(term, doc) = \frac{\#\text{ de veces que el } term \text{ aparece en el } doc}{\#\text{ de } terms \text{ diferentes en el } doc}$$

Ejemplo en un documento:

$$\overbrace{\text{Hello,}}^{0.125}\ \text{my}\ \overbrace{\text{name}}^{0.125}\ \text{is}\ \overbrace{\text{Brandon.}}^{0.375}\ \text{Brandon Brandon.}\ \text{The}\ \overbrace{\text{elephant}}^{0.125}\ \overbrace{\text{jumps}}^{0.125}\ \text{over the}\ \overbrace{\text{moon.}}^{0.125}$$

Vectorizar • TF - IDF

Document Frequency

Fracción de todos los documentos en nuestro corpus que contienen el término.

$$\textbf{DF}(\textit{term},\textit{corpus}) = \frac{\text{\# de } \textit{docs} \text{ que contienen } \textit{term}}{\text{\# total de } \textit{docs}}$$

Ejemplo en un documento:

$$\overbrace{\text{Hello,}}^{0.125}\ \text{my}\ \overbrace{\text{name}}^{0.125}\ \text{is}\ \overbrace{\text{Brandon.}}^{0.375}\ \text{Brandon Brandon.}\ \text{The}\ \overbrace{\text{elephant}}^{0.125}\ \overbrace{\text{jumps}}^{0.125}\ \text{over the}\ \overbrace{\text{moon.}}^{0.125}$$

Vectorizar • TF - IDF

Inverse Document Frequency

Logaritmo inversa de DF.

$$\textbf{DF}(\textit{term},\textit{corpus}) = \text{Log}\left(\frac{\#\text{ total de } \textit{docs}}{\#\text{ de } \textit{docs} \text{ que contienen } \textit{term}}\right)$$

Ejemplo: si está en todos los docs log(N/N) = log(1) = 0

Vectorizar • TF - IDF

Inverse Document Frequency

Producto del valor de TF por el de IDF.

TF-IDF(*term*,*corpus*,*doc*) = **TF**(*term*,*doc*) X **IDF**(*term*,*corpus*)

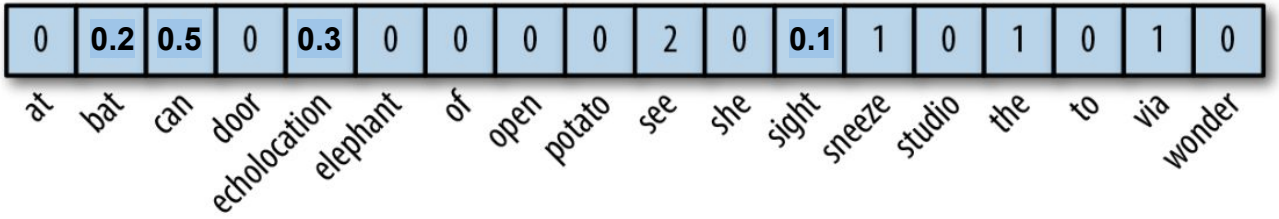
Vectorizar • TF - IDF

Inverse Document Frequency

Producto del valor de TF por el de IDF.

TF-IDF(*term*,*corpus*,*doc*) = **TF**(*term*,*doc*) X **IDF**(*term*,*corpus*)

Cada palabra tiene un valor asociado en cada documento, con esto formamos nuestro vector (no necesariamente serán valores enteros):



GPT-2

Parte 1

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was [illegible] today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which [illegible] Ashland stations.

In an email to Ohio news outlets, the U.S. Department of [illegible] working with the Federal Railroad Administration to find [illegible]

"The theft of this nuclear material will have significant [illegible] on public and environmental health, our workforce and [illegible] nation," said Tom Hicks, the U.S. Energy Secretary, [illegible] priority is to secure the theft and ensure it doesn't [illegible]

The stolen material was taken from the University of [illegible] Triangle Park nuclear research site, according to a news [illegible] Department officials.

The Nuclear Regulatory Commission did not immediately [illegible]

According to the release, the U.S. Department of Energy's [illegible]

TWO MINUTE PAPERS

WITH KÁROLY ZSOLNAI-FEHÉR

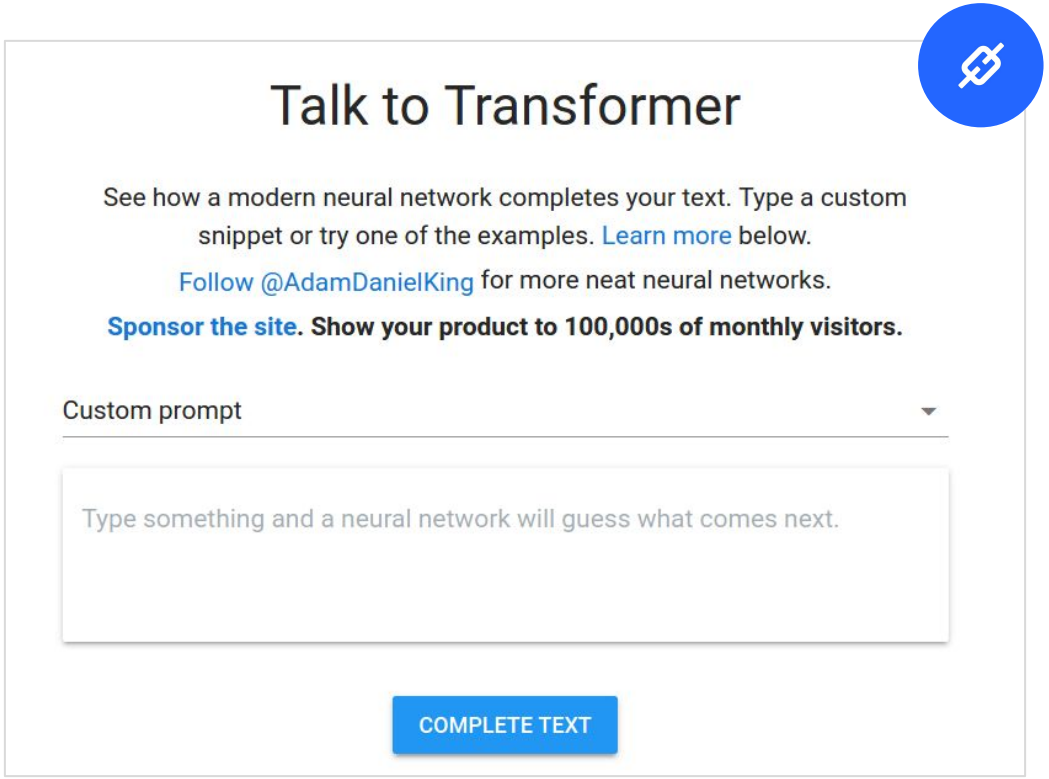
Hands-on training

Hands-on training

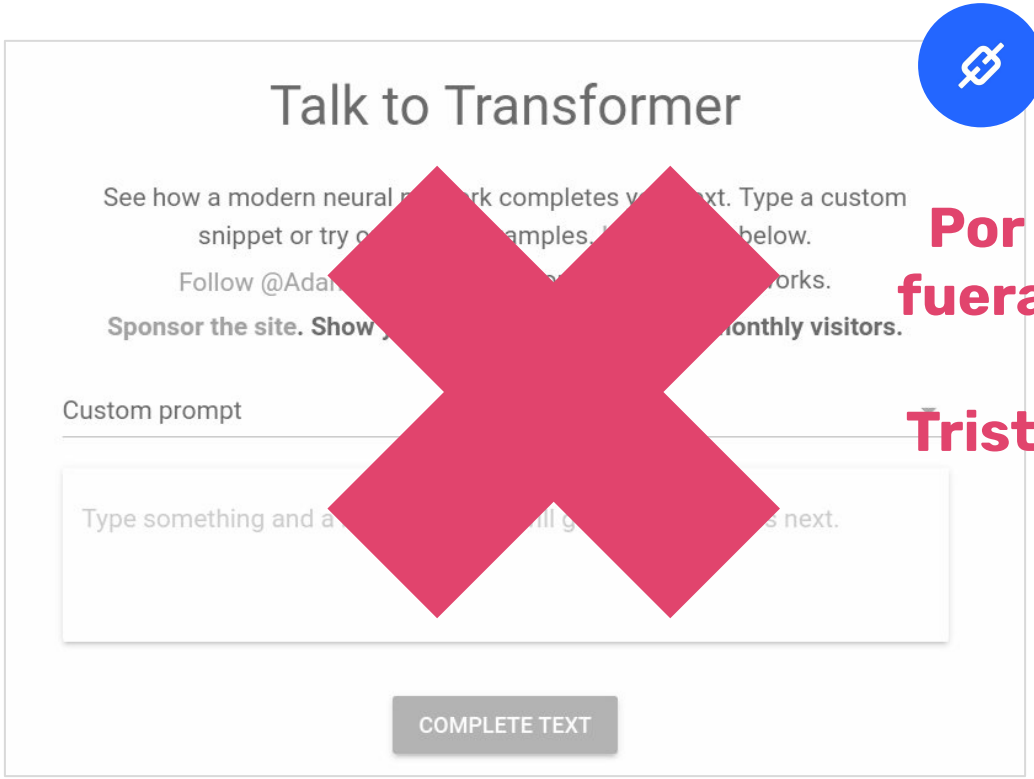
DS_Bitácora_39_y_40_NLP.ipynb

Sección 4 en adelante

Para jugar


Talk to Transformer
See how a modern neural network completes your text. Type a custom snippet or try one of the examples. Learn more below.
Follow @AdamDanielKing for more neat neural networks.
Sponsor the site. Show your product to 100,000s of monthly visitors.
Custom prompt
Type something and a neural network will guess what comes next.
COMPLETE TEXT

Para jugar



Por ahora, está fuera de servicio.

Tristeza nao tem fim.

GPT-2

Parte 2


TWO MINUTE
PAPERS
WITH KÁROLY ZSOLNAI-FEHÉR

Proyecto 03

El Proyecto 3 de tu carrera tiene como objetivo que apliques las herramientas aprendidas dentro de tres dominios específicos sobre los que aprenderás en el Bloque 3:

- Sistemas de Recomendación
- Procesamiento de lenguaje natural
- Series de tiempo

El Proyecto 3 de tu carrera tiene como objetivo que apliques las herramientas aprendidas dentro de tres dominios específicos sobre los que aprenderás en el Bloque 3:

- Sistemas de Recomendación
- Procesamiento de lenguaje natural
- Series de tiempo

Implementa un modelo para reconocer el puntaje asignado a un ítem de Amazon a partir de la crítica que hace un/a usuario/a.

- **Dataset.** Aquí puedes encontrar la descarga del dataset y aquí información sobre el mismo. Es importante que tengas en cuenta la licencia de este dataset.
- **Notebook.** Aquí te dejamos un notebook con algunas recomendaciones para tu trabajo.

amazon

Para la próxima

- Avanza con el notebook de hoy.
- Lee la bitácora 41 y carga las dudas que tengas al Trello.

En el encuentro que viene uno/a de ustedes será seleccionado/a para mostrar cómo resolvió el challenge de la bitácora. De esta manera, ¡aprendemos todos/as de (y con) todas/as, así que vengan preparados/as.

ACÁMICA