Sistema de Preguntas y Respuestas

Juan Pablo Herrera Musi & Daniela Pinto Veizaga

Profesor: Felipe González
Maestría en Ciencia de Datos, ITAM
Ciudad de México, 26 de mayo de 2020

Introducción

Objetivo: Emplear métodos de nlp y deep learning para desarrollar un sistema de preguntas y respuestas que, dada una pregunta, sea capaz de predecir la posición de la oración (dentro de un corpus o texto) que contiene la respuesta.

Contexto:

[0] Architecturally, the school has a Catholic Character. [1] Atop the main building's gold dome is a golden statue of the Virgin Mary. [2] Immediately in front of the main building and facing it, is a copper statue of Christ with arms upraised with the legend "venite ad me omnes". [3] Next to the main building is the Basilica of the Sacred Heart. [4] Immediately behind the Basilica is the Grotto, a Marian place of prayer and reflection. [5] It is a replica of the Grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. [6] At the end of the main drive (and in a direct line that connects through 3 statues and the gold dome), is a simple, modern stone statue of Mary. [7] As at most other universities, Notre Dame's students run a number of news media outlets. [8] The nine students run outlets include three newspapers, both a radio and television station, and several magazines and journals. [9] Begun as a one-page journal in September 1876, the scholastic magazine is issued twice monthly and claims to be the oldest continuous collegiate publication in the United States.

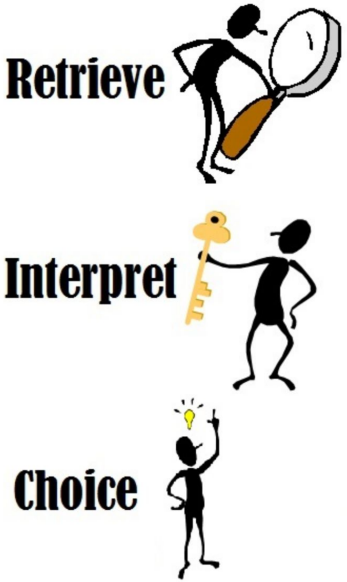
Pregunta:

¿To whom did Virgin Mary allegedly appear in 1858 in Lourdes France?

Respuesta:

To Saint Bernadette Soubirous in 1858.

Introducción



Tipos de Sistemas de QA:

- Dominio abierto
- Dominio cerrado

Tipos de preguntas:

- Factoid questions
- Cloze questions
- Sí/No
- Quién? Qué? Dónde? Cuándo?...
- Choice questions

Dataset: SQuAD V1.0

Estructura BD

Imperialism

The Stanford Question Answering Dataset

CONTEXTO

Imperialism is a type of advocacy of empire. Its name originated from the Latin word "imperium", which means to rule over large territories. Imperialism is "a policy of extending a country's power and influence through colonization, use of military force, or other means". Imperialism has greatly shaped the contemporary world. It has also allowed for the rapid spread of technologies and ideas. The term imperialism has been applied to Western (and Japanese) political and economic dominance especially in Asia and Africa in the 19th and 20th centuries. Its precise meaning continues to be debated by scholars. Some writers, such as Edward Said, use the term more broadly to describe any system of domination and subordination organised with an imperial center and a periphery.

PREGUNTAS/RESPUESTAS

The word imperialism has it's origins in which ancient language?
Ground Truth Answers: Latin | Latin | Latin | Latin | Latin

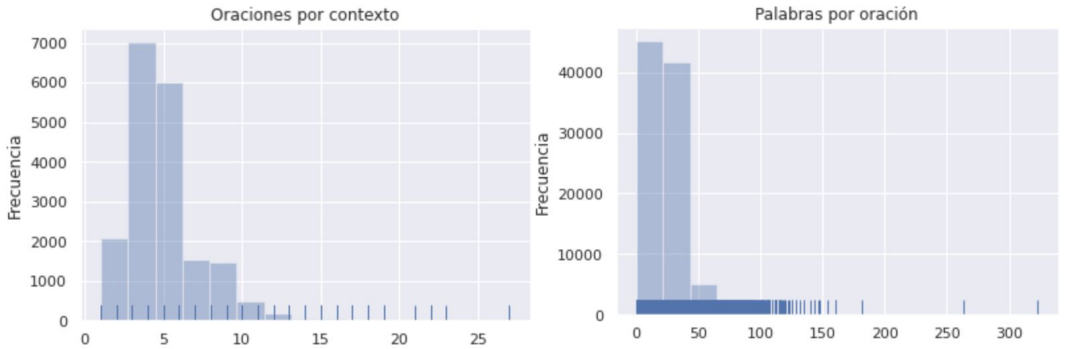
By what means is imperialism usually administered?
Ground Truth Answers: military force | colonization, use of military force, or other means | colonization, use of military force, or other means | colonization, use of military force, or other | colonization

The term imperialism has been applied to western countries, and which eastern county?
Ground Truth Answers: Japan | Japan | Japanese | Japanese | Japan

Imperialism is responsible for the rapid spread of what?
Ground Truth Answers: technologies and ideas | technologies and ideas | technologies and ideas | technologies and ideas | technologies and ideas.

Tipo de Respuestas

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

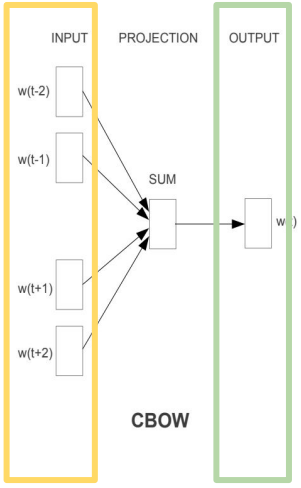


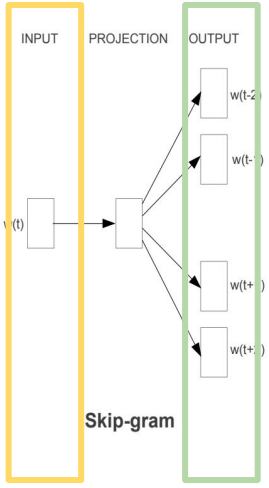
Conteos Básicos

IMPLEMENTACIÓN

Procesamieno y Feature engineering

Word Embeddings

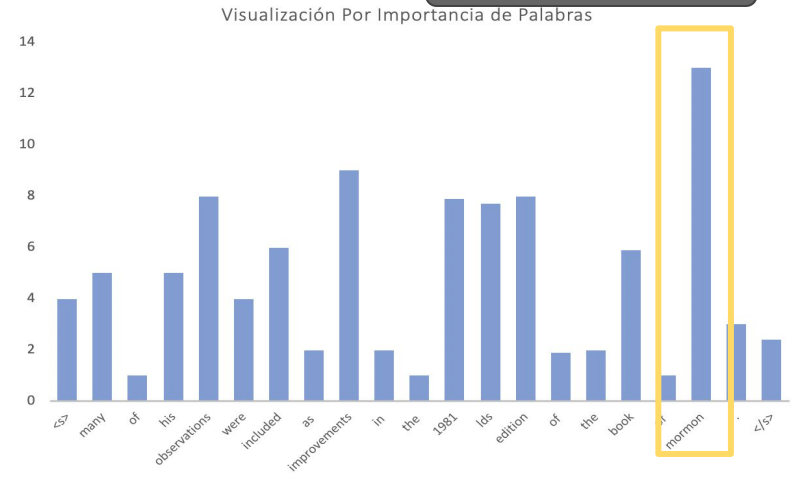




Word2Vec

Skipgram		CBOW	
Palabra	Similitud	Palabra	Similitud
construction	0.820	centre	0.855
constructing	0.799	campus	0.820
constructed	0.794	station	0.812
buildings	0.792	site	0.809
housed	0.790	shopping	0.804

Sentence Embeddings



Cosine Distance

Contexto:

[0] Architecturally, the school has a Catholic Character. [1] Atop the main building's gold dome is a golden statue of the Virgin Mary. [2] Immediately in front of the main building and facing it, is a copper statue of Christ with arms upraised with the legend "venite ad me omnes". [3] Next to the main building is the Basilica of the Sacred Heart. [4] Immediately behind the Basilica is the Grotto, a Marian place of prayer and reflection. [5] It is a replica of the Grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. [6] At the end of the main drive (and in a direct line that connects through 3 statues and the gold dome), is a simple, modern stone statue of Mary. [7] As at most other universities, Notre Dame's students run a number of news media outlets. [8] The nine students run outlets include three newspapers, both a radio and television station, and several magazines and journals. [9] Begun as a one-page journal in September 1876, the scholastic magazine is issued twice monthly and claims to be the oldest continuous collegiate publication in the United States.

Pregunta:

¿To whom did Virgin Mary allegedly appear in 1858 in Lourdes France?

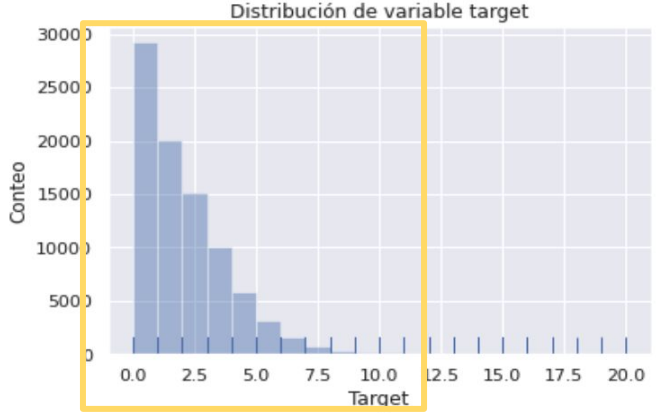
Respuesta:

To Saint Bernadette Soubirous in 1858.

Distancia Coseno

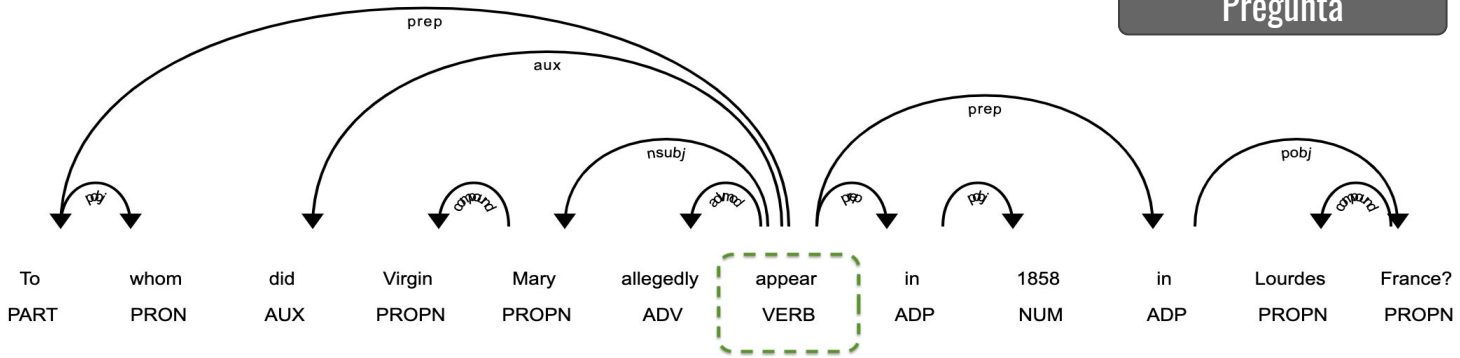
Oración	Distancia
0	0.515
1	0.560
2	0.491
3	0.539
4	0.511
5	0.388
6	0.515
7	0.506
8	0.525
9	0.382

Target Variable

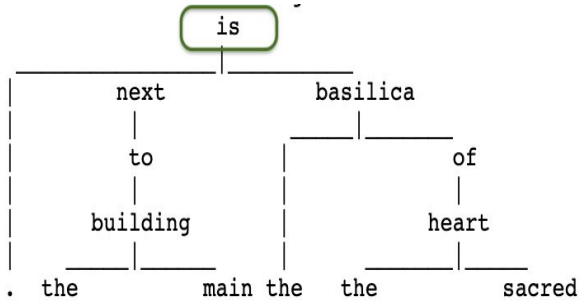


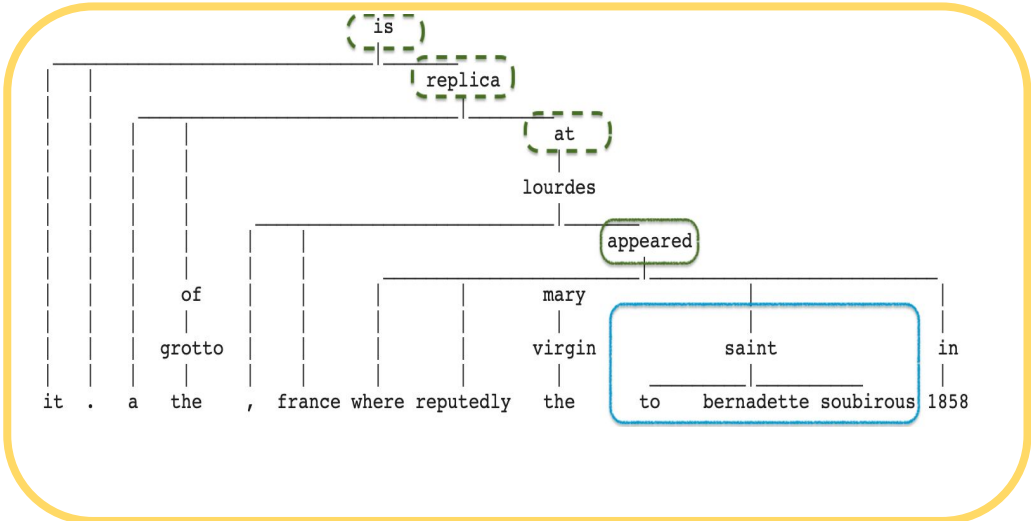
Root Analysis

Pregunta

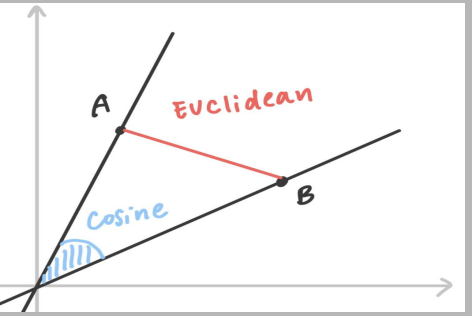


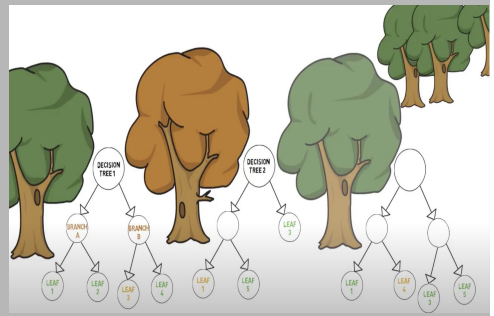
Oraciones





Modelos y Resultados





Modelos y Resultados

No Supervisados			Supervisados		
Modelo	Acc.-Training	Acc.-Test	Modelo	Acc.-Training	Acc.-Test
Distancia Coseno	0.5996	0.600	Random Forest	0.628	0.599
Distancia Euclidiana	0.36966	0.370	Adaboost	0.337	—
			Decision Tree	0.524	—

Modelos Supervisados: Tiempos	
Modelo	Tiempo-Training
Random Forest	804 s.
Adaboost	3 s.
Decision Tree	114 s.

Conclusiones

Para conocer más de la implementación, revisar el siguiente link: QuestionAnswer System