Proyecto final Recuperación de Información Multimedia CC5213 Primavera 2020

AUTOR: ERWIN PAILLACÁN

PROFESOR: JUAN MANUEL BARRIOS

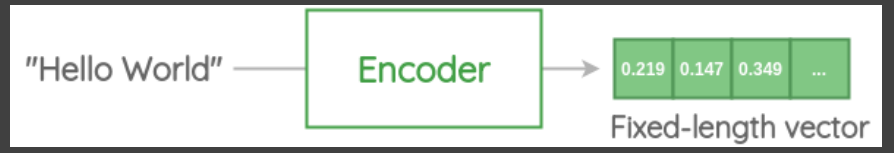
Motivación y objetivos

Motivación: Buscar frases que son usadas como citas atribuidas a algún autor, pero no se sabe en qué libro y por lo tanto no se sabe si es cierto.

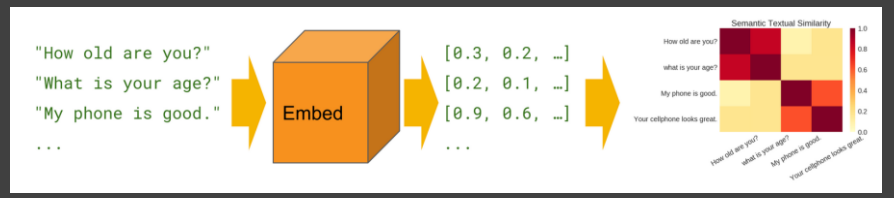
Objetivo: Dada una cita exacta o parafraseada encontrar el libro donde esta escrita, si es que lo está.

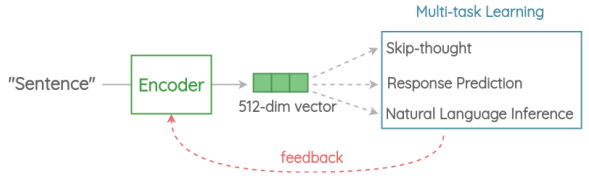
Multilingual Universal Sentence Encoder for Semantic Retrieval

- Se desea transformar cualquier palabra, frase o párrafo en un vector de dimensión fija.
- Si bien un enfoque muy usado es tomar el promedio de las embeddings de tokens y promediar. El orden si importa.
- 2 Modelos:
  - Transformers
  - CNN

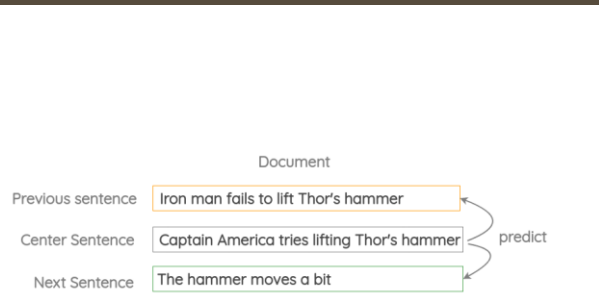


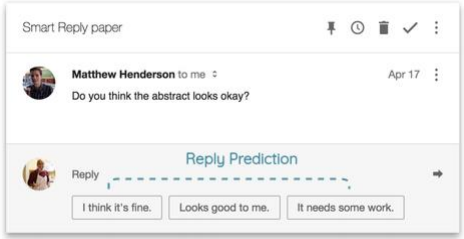

```
>>> nlp('this is cool').similarity(nlp('is this cool'))
1.0
```





Premise	Hypothesis	Judgement
A soccer game with multiple males playing	Some men are playing a sport	entailment
I love Marvel movies	I hate Marvel movies	contradiction
I love Marvel movies	A ship arrived	neutral





Entrenamiento multitarea

El modelo es entrenado resolviendo múltiples tareas:

1. Modified Skip-thought: Usar sentencia actual para predecir sentencia anterior y posterior.
2. Response prediction: Predecir respuesta correcta dada una entrada y una lista de múltiples posibles respuestas.
3. Natural language inference: Predecir el juicio (vinculación – neutral – contradicción) dada una premisa y una hipotesis

Buscador de citas

1) Extraer texto de libros EPUB.
2) Separar el texto en frases, enfoque actual separa frases por “.”
3) Normalizar y limpiar texto.
4) Inferir el vector embedding para cada frase y guardar una matriz de vectores para cada libro. La matriz tiene dimensiones [n x 512], donde n es el numero de frases.
5) Dada un cita, pre-procesar el texto e inferir su vector embedding.
6) Encontrar la frase más similar para cada frase de un libro de acuerdo con similitud coseno. Elegir el libro que tenga la similitud coseno más alta.

Resultados sobre 28 citas

ESCENARIO	PRECISIÓN
Con StopWords	78%
Sin StopWords	82%

Ejemplo:

Cita a buscar: ***“Aquel incapaz de deshacerse del tesoro en necesidad se parece a un esclavo”***

Cita encontrada*:* ***“Quien no es capaz de desprenderse de un tesoro en un momento de necesidad es como un esclavo encadenado”***

Libro: *Las dos torres. Tolkien*

Similitud: 0,88.

Demo

Conclusiones

- Difícil de medir eficacia.
- Búsqueda limitada a los libros pre-procesados
- Mejora futura: Mejor conjunto de pruebas, para medir eficacia.