Tipo : Guía de Laboratorio
Capítulo : NLP en Python
Duración : 30 minutos

I. OBJETIVO

Demostrar competencias básicas en uso de un framework NLP..

II. REQUISITOS

Los siguientes elementos de software son necesarios para la realización del laboratorio:

- Instalar Anaconda en Windows
- Navegador web

III. EJECUCIÓN DEL LABORATORIO

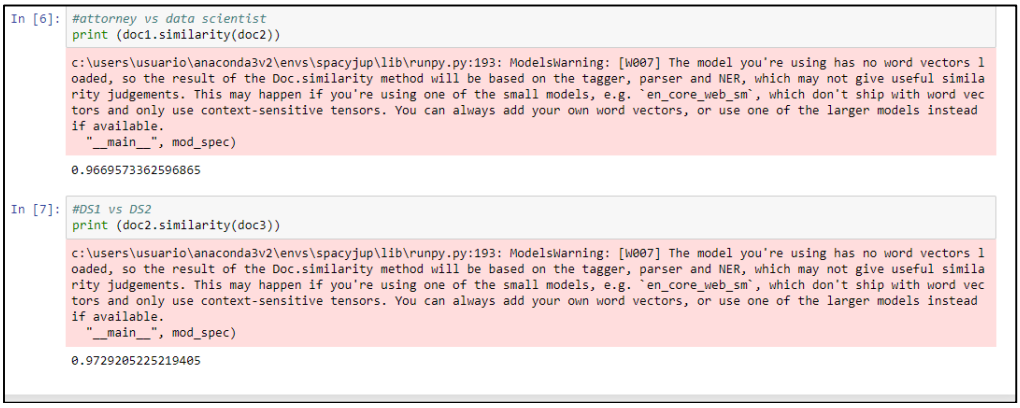
- Ejercicio 3.1: spaCy

  - Crear un entorno virtual
    a. conda create --name spacyjup python=3.5
    b. activate spacyjup
    c. pip install
       i. jupyter
       ii. spacy
    d. python -m spacy download en
  - Activar jupyter en la línea de comandos, con jupyter notebook
  - Abrir spacyjup.ipynb en el browser
  - Ejecutar el código y consultar

1. **Carga**

```
In [1]: import spacy

In [2]: nlp = spacy.load('en_core_web_sm')

In [3]: def get_file_contents(filename):
            with open(filename, 'r') as filehandle:
                filecontent = filehandle.read()
                return (filecontent)

In [4]: fn1_doc=get_file_contents('attorney_profile.txt')
        fn2_doc=get_file_contents('data_scientist_profile.txt')
        fn3_doc=get_file_contents('data_scientist_profile_2.txt')

In [5]: doc1 = nlp(fn1_doc)
        doc2 = nlp(fn2_doc)
        doc3 = nlp(fn3_doc)
```

2. Similitud de documentos



```
In [6]: #attorney vs data scientist
        print (doc1.similarity(doc2))

        c:\users\usuario\anaconda3v2\envs\spacyjup\lib\runpy.py:193: ModelsWarning: [W007] The model you're using has no word vectors l
        oaded, so the result of the Doc.similarity method will be based on the tagger, parser and NER, which may not give useful simila
        rity judgements. This may happen if you're using one of the small models, e.g. `en_core_web_sm`, which don't ship with word vec
        tors and only use context-sensitive tensors. You can always add your own word vectors, or use one of the larger models instead
        if available.
          "__main__", mod_spec)

        0.9669573362596865

In [7]: #DS1 vs DS2
        print (doc2.similarity(doc3))

        c:\users\usuario\anaconda3v2\envs\spacyjup\lib\runpy.py:193: ModelsWarning: [W007] The model you're using has no word vectors l
        oaded, so the result of the Doc.similarity method will be based on the tagger, parser and NER, which may not give useful simila
        rity judgements. This may happen if you're using one of the small models, e.g. `en_core_web_sm`, which don't ship with word vec
        tors and only use context-sensitive tensors. You can always add your own word vectors, or use one of the larger models instead
        if available.
          "__main__", mod_spec)

        0.9729205225219405
```

IV. EVALUACIÓN

1. ¿Cuál es el inconveniente de usar un framework con un modelo predefinido?

    a. **Respuesta:** En primer lugar, está el tema del idioma, puesto que spaCy es un buen framework para tareas de NL; pero, particularmente, para trabajar en inglés. En segundo lugar, a pesar que el resultado mostrado parece tener sentido, se apreciarán problemas en otras comparaciones y es porque el modelo de palabras cargado no está optimizado para la comparación de perfiles de trabajo.