CoGrammar

Natural Language Processing

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

Data Science Lecture Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(FBV: Mutual Respect.)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Open Classes. You can submit these questions here: **Open Class Questions**

Data Science Lecture Housekeeping cont.

- For all **non-academic questions**, please submit a query: www.hyperiondev.com/support

- Report a **safeguarding** incident: www.hyperiondev.com/safeguardreporting

- We would love your **feedback** on lectures: Feedback on Lectures

Lecture Objectives

- Employ NLP techniques using **spaCy** for tasks such as **tokenization, named entity recognition, and semantic similarity assessments,** culminating in the creation of a **basic film recommendation engine using word vectors.**

101010
100101
NPL
Hello!
Hi!

NLP Introduction

- ★ **Natural Language Processing (NLP)** is a crucial field in AI that focuses on the **interaction between computers and humans using natural language.** The ultimate objective of NLP is to **read, decipher, understand, and make sense of human languages in a manner that is valuable.**

- ★ **Importance of NLP:** It's essential for various applications, including **sentiment analysis, language translation, and information extraction**, enabling computers to **process and analyze vast amounts of natural language data.**

spaCy

★ spaCy is **a leading NLP library in Python**, known for its efficiency and ease of use. Unlike other libraries, **spaCy is designed specifically for production use, offering fast and accurate linguistic annotations.**

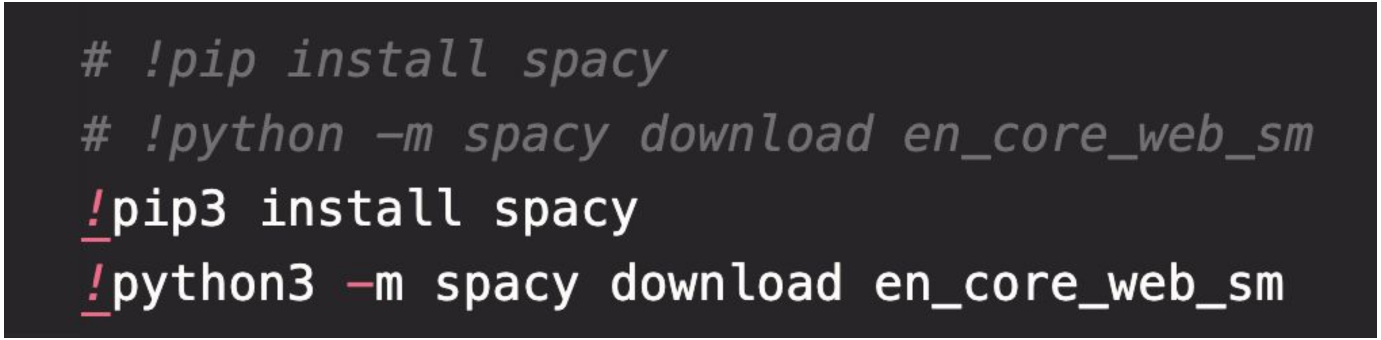
★ For the adventurous amongst you, search for HuggingFace - a platform that also hosts amazing NLP models used in production applications.

spaCy

- ★ spaCy provides **out-of-the-box support for many NLP tasks**, such as tokenization, POS tagging, and NER, making it more robust and faster than libraries like NLTK or TextBlob.

- ★ **Key Features:**
  - ○ **Speed:** Optimized algorithms and data structures for fast performance.
  - ○ **Accuracy:** High accuracy in linguistic annotations, powered by deep learning.
  - ○ **Ease of Use:** Intuitive API and extensive documentation, suitable for beginners and professionals.

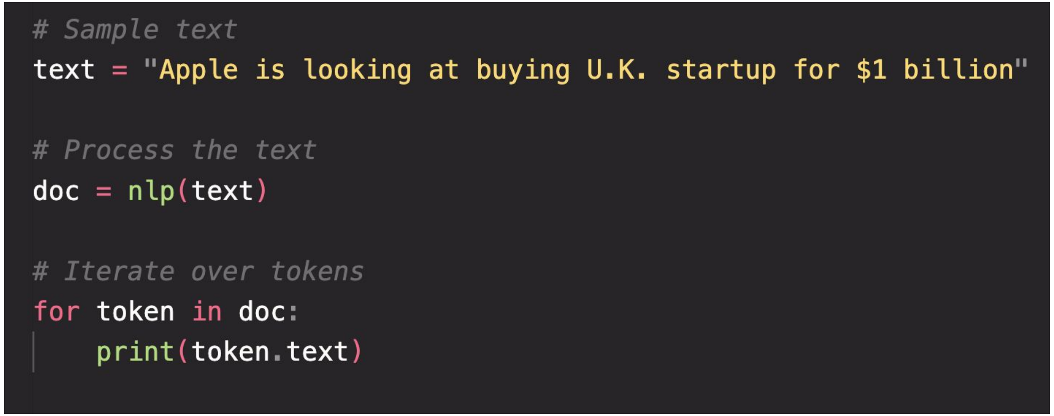
Installation and Setup



```
# !pip install spacy
# !python -m spacy download en_core_web_sm
!pip3 install spacy
!python3 -m spacy download en_core_web_sm
```

Tokenization

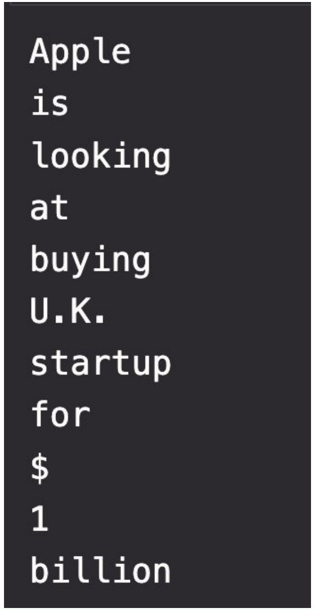
★ Tokenization is the process of **breaking down text into individual words, phrases, symbols, or other meaningful elements called tokens**.



```python
# Sample text
text = "Apple is looking at buying U.K. startup for $1 billion"

# Process the text
doc = nlp(text)

# Iterate over tokens
for token in doc:
    print(token.text)
```



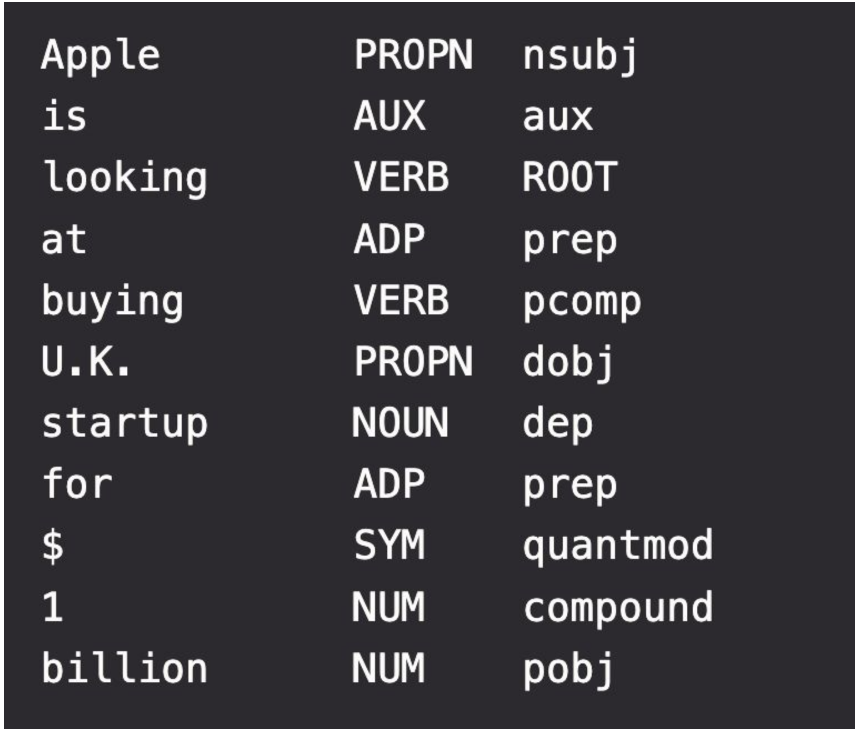
```
Apple
is
looking
at
buying
U.K.
startup
for
$
1
billion
```

Linguistic Features

★ **POS Tagging:** Part-of-Speech tagging **assigns parts of speech to each word, such as noun, verb, adjective, etc.**, based on its definition and context.

★ **Dependency Parsing:** Analyzes the grammatical structure of a sentence, establishing r**elationships between "head" words and words which modify those heads**.

Linguistic Features



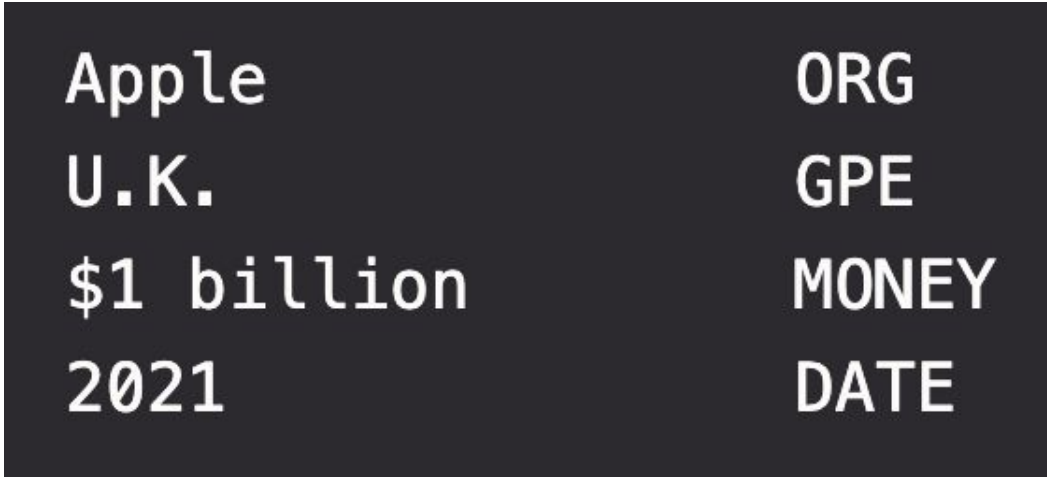
```
Apple        PROPN  nsubj
is           AUX    aux
looking      VERB   ROOT
at           ADP    prep
buying       VERB   pcomp
U.K.         PROPN  dobj
startup      NOUN   dep
for          ADP    prep
$            SYM    quantmod
1            NUM    compound
billion      NUM    pobj
```

Named Entity Recognition

- ★ NER is a **process of locating and classifying named entities** mentioned in unstructured text into pre-defined categories such as **person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.**

- ★ spaCy includes a **pre-trained NER model capable of recognizing various named entities.** It uses a combination of convolutional neural networks (CNNs) and conditional random fields (CRFs) for high accuracy.

Named Entity Recognition



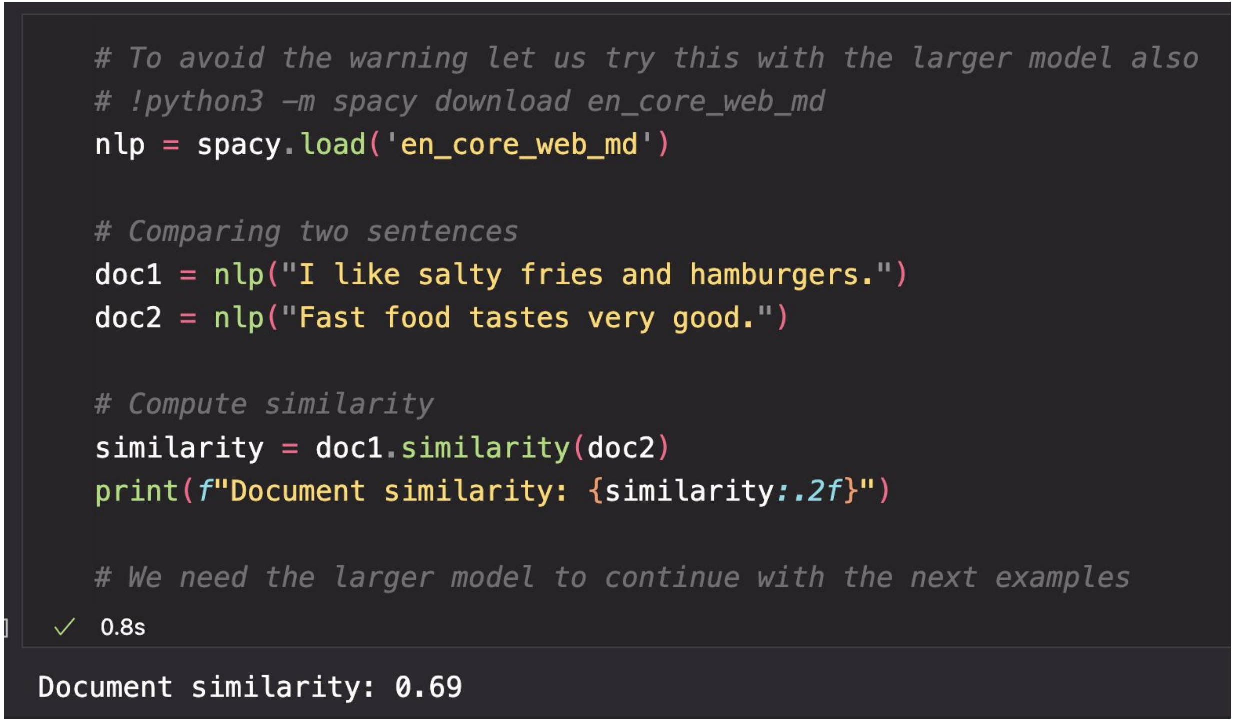
```
Apple        ORG
U.K.         GPE
$1 billion   MONEY
2021         DATE
```

Semantic Similarity

★ Semantic similarity measures **how much two pieces of text (documents, sentences, or words) are related to each other** in terms of meaning.

★ spaCy uses **word vectors, multidimensional representations of meanings of words, to calculate similarity**.

Semantic Similarity



```
# To avoid the warning let us try this with the larger model also
# !python3 -m spacy download en_core_web_md
nlp = spacy.load('en_core_web_md')

# Comparing two sentences
doc1 = nlp("I like salty fries and hamburgers.")
doc2 = nlp("Fast food tastes very good.")

# Compute similarity
similarity = doc1.similarity(doc2)
print(f"Document similarity: {similarity:.2f}")

# We need the larger model to continue with the next examples
```

✓ 0.8s

```
Document similarity: 0.69
```

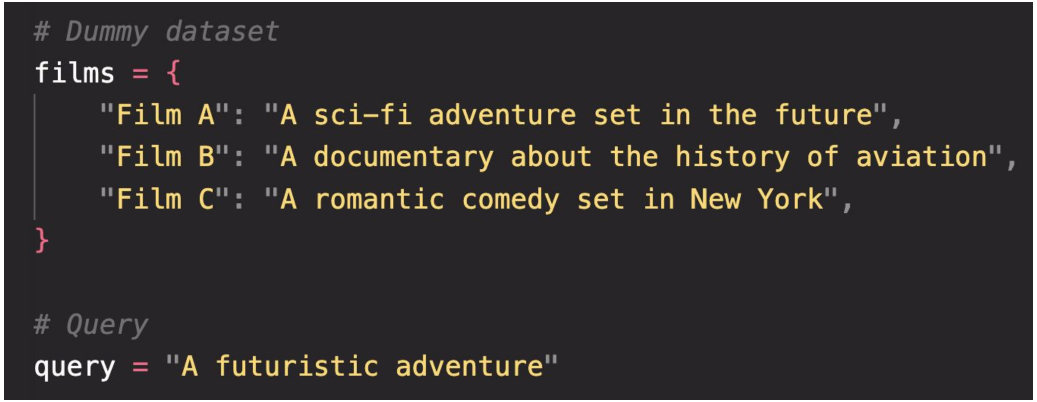
Let’s Breathe!

Let’s take a small break before moving on to the next topic.

Film Recommendation Engine

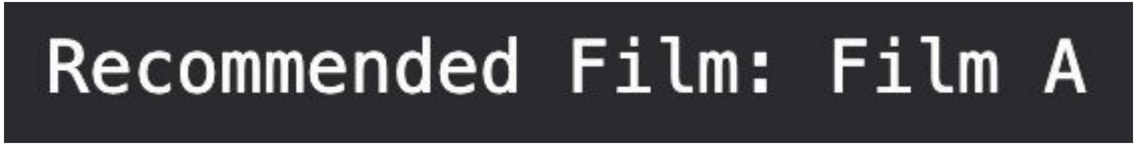
- ★ **Building the Engine with spaCy**:
  - ○ **Data Preparation:** Start with a dataset of film descriptions.
  - ○ **Feature Extraction:** Use spaCy to process descriptions and extract features such as named entities, keywords, and semantic vectors.
  - ○ **Similarity Calculation:** Compute similarity scores between a query (user's favorite film description) and the dataset.
  - ○ **Recommendation:** Recommend films with the highest similarity scores to the query.

Film Recommendation Engine



```python
# Dummy dataset
films = {
    "Film A": "A sci-fi adventure set in the future",
    "Film B": "A documentary about the history of aviation",
    "Film C": "A romantic comedy set in New York",
}


# Query
query = "A futuristic adventure"
```



```
Recommended Film: Film A
```

CoGrammar

Q & A SECTION

Please use this time to ask any questions relating to the topic, should you have any.

CoGrammar

Thank you for joining us

1. Take regular breaks
2. Stay hydrated
3. Avoid prolonged screen time
4. Practise good posture
5. Get regular exercise

“With great power comes great responsibility”