IMPERIAL

Word Representation in Biomedical Domain

Natural Language Processing Team Project

Weihang Zhang, Shuojie Fu

03/02/2025

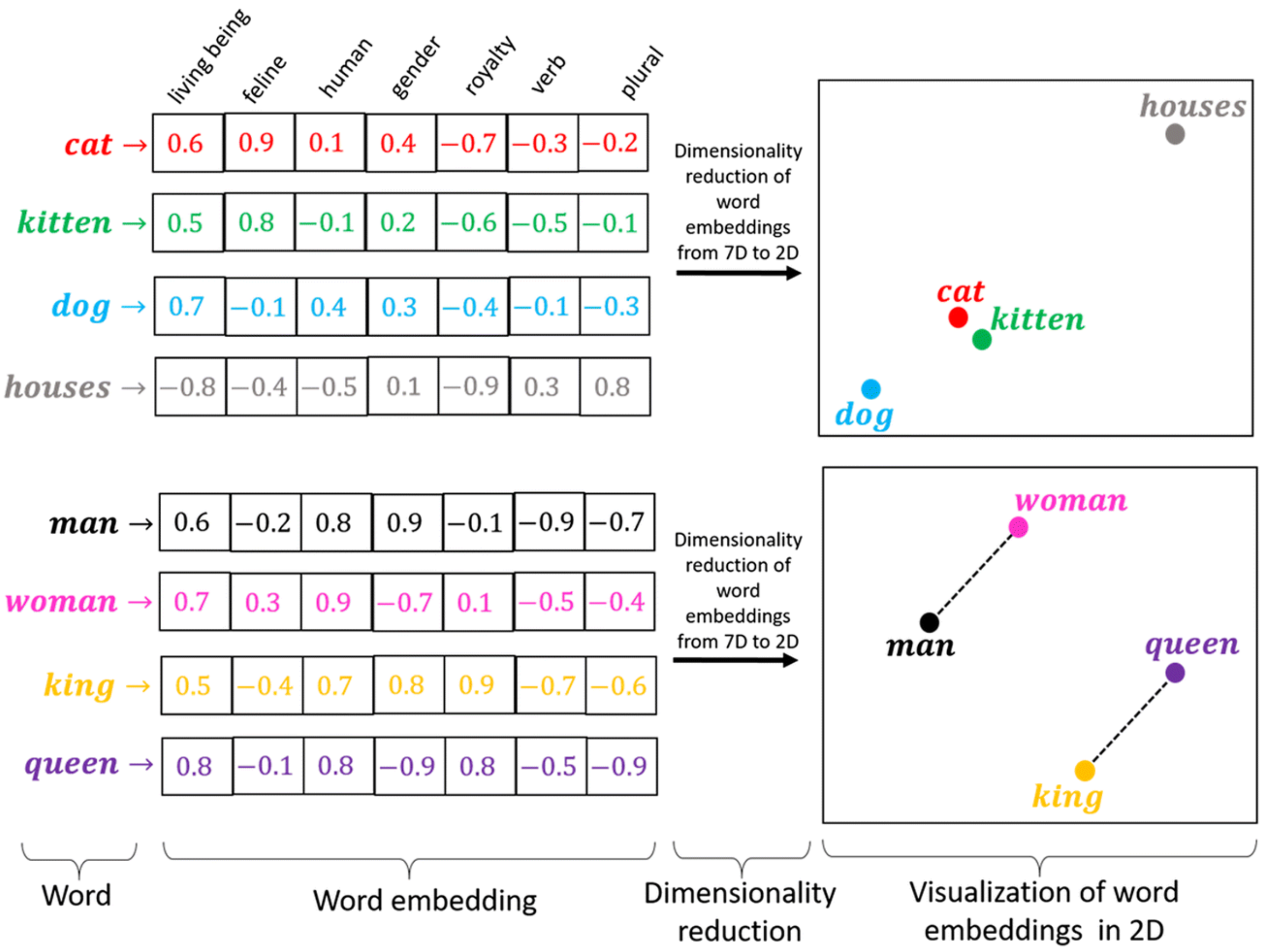
Natural Language Processing

Applications of NLP

- Question Answering
- Text summarization
- Sentiment analysis
- Speech recognition
- Neural Machine Translation
- Search engine
- Smart assistant
- Language translation
- Recommendation system used by major companies
- Healthcare, finance....

Natural Language Processing

Word Representation



	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

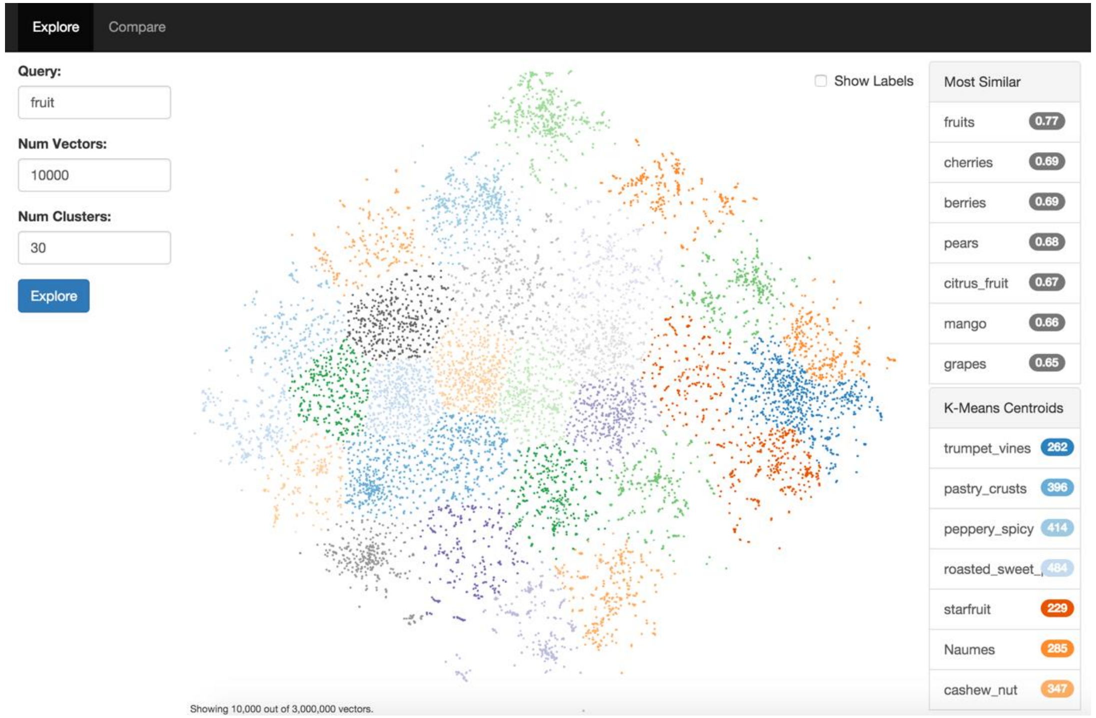
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

- How can machine learn to represent each word in the human language?
- How can we know if the machine has learnt the semantic meaning of the words?
- What can we do with the learnt word representations?

Natural Language Processing

Word Representation

- Mapping each word token to a vector that represents a point in a ‘space’.
- Using limited dimensions to represent semantics of words.
- Considering semantic similarity and difference between words.



Query: fruit

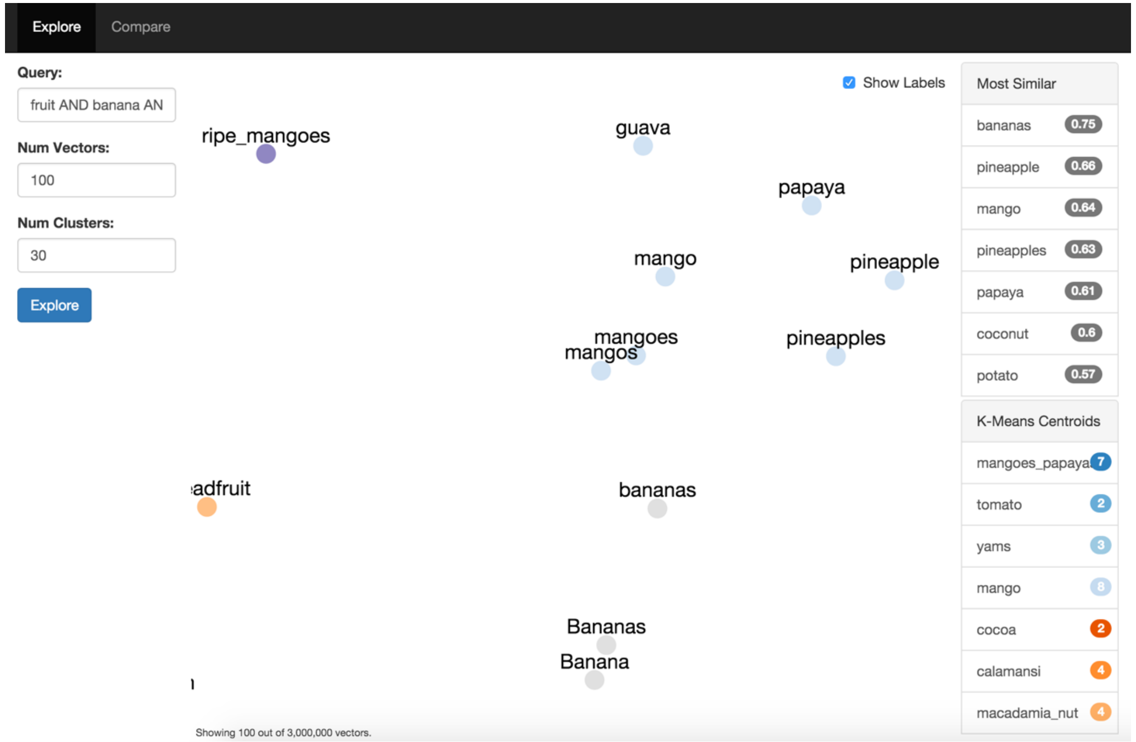
Num Vectors: 10000

Num Clusters: 30

Most Similar	
fruits	0.77
cherries	0.69
berries	0.69
pears	0.68
citrus_fruit	0.67
mango	0.66
grapes	0.65

K-Means Centroids	
trumpet_vines	262
pastry_crusts	396
peppery_spicy	414
roasted_sweet_	484
starfruit	229
Naumes	285
cashew_nut	347

Showing 10,000 out of 3,000,000 vectors.



Query: fruit AND banana AN

Num Vectors: 100

Num Clusters: 30

Most Similar	
bananas	0.75
pineapple	0.66
mango	0.64
pineapples	0.63
papaya	0.61
coconut	0.6
potato	0.57

K-Means Centroids	
mangoes_papaya	7
tomato	2
yams	3
mango	8
cocoa	2
calamansi	4
macadamia_nut	4

Showing 100 out of 3,000,000 vectors.

Natural Language Processing

Dataset – CORD-19

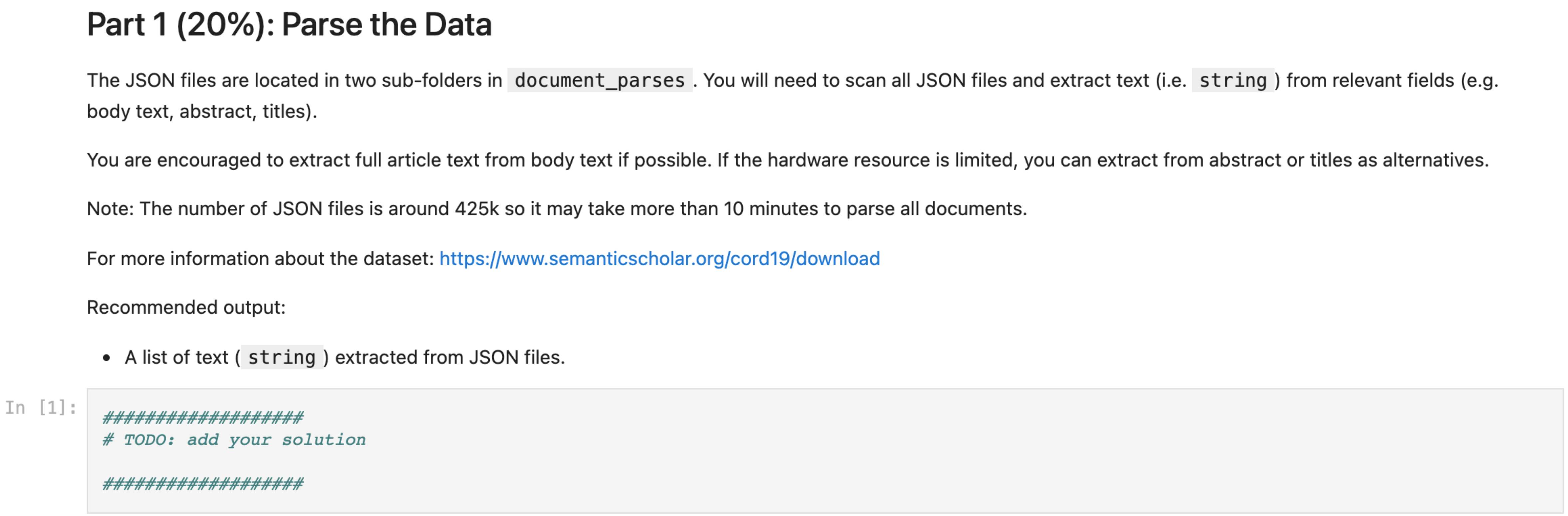
- COVID-19 Open Research Dataset (CORD-19)
  - Over 500,000 scholarly articles
  - including over 200,000 with full text
  - about COVID-19, SARS-CoV-2, and related coronaviruses

Natural Language Processing

Project Overview

- In this project, you will develop algorithms to learn word representations.
- Part 1: Parse the data
- Part 2: Tokenization
- Part 3: Build word representations
- Part 4: Explore word representations
- Part 5 (Bonus): Mine biomedical knowledge

Natural Language Processing
Jupyter Notebook



Part 1 (20%): Parse the Data

The JSON files are located in two sub-folders in `document_parses`. You will need to scan all JSON files and extract text (i.e. `string`) from relevant fields (e.g. body text, abstract, titles).

You are encouraged to extract full article text from body text if possible. If the hardware resource is limited, you can extract from abstract or titles as alternatives.

Note: The number of JSON files is around 425k so it may take more than 10 minutes to parse all documents.

For more information about the dataset: https://www.semanticscholar.org/cord19/download

Recommended output:

- A list of text (`string`) extracted from JSON files.

In [1]:

```
###################
# TODO: add your solution

###################
```

Natural Language Processing

Report

- For each part
  - Explain the method used
  - Show the results
  - Analyze and discuss the result

Natural Language Processing

Marking Scheme

- NLP Project – 40%
  - Completeness – 30%
  - Final report – 10%
  - It is fine to learn from online sources, tutorials, examples, etc. However, ensure that all relevant sources are properly cited.
- Breakdown scores for each part (1-5) are available in Jupyter Notebook
- Group work, collaboration is important
- NLP project code and final report
- Cite other's work wherever appropriate

IMPERIAL

Thank you