Dictionary for Culinary Terms

Getting all kinds of culinary terms from the internet, along with scraping their meanings.

Vaibhav Sharma(2019284)
Manav Saini(2020581)
Shivam Verma(2019272)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Scraping from www.englishclub.com/vocabulary/food

- Data present as:
  - Words related to food under different categories like types of food, cooking vocabulary, kitchen and kitchen ware, dining vocabulary
  - Words were present as captions under pictures and as lists
- Developed a function to scrape all words under pictures and in the list and ran the function on all the pages of different categories
- Formed a Dataframe with the words and cleaned the text
  - Removed redundant extra information in parentheses
  - Converted the text into lower case
- No. of words scraped = 786

Scraping from www.touchbistro.com

- Data present as:
  - Dictionary form
    - First word followed by a colon then meaning
- Approach:
  - Scraped all the lines splitted them into word and meaning using split function and using colon as a parameter
  - Appending the word in a list and meaning in another list
  - Created a dataframe
  - Cleaned the text
    - Lowercasing of words
    - Removal of parentheses with extra information
    - Words with a space were joined together by a hyphen(‘-’)
  - Converted Dataframe to a CSV file
- No. of words scraped = 103

Scraped a PDF

- We got a pdf with 222 culinary terms with definition in it. (pdf submitted as well).
- Used python to get text from the pdf, and then write it into the text file.
- Then made all terms+definition in one line, so that line by line reading can be done.
- Preprocessing the text:
  - Removed dots
  - Separated words and their meanings, and made two lists.
  - Words with a space were joined together by a hyphen(‘-’).

Scraping from www.food.com

- Data present as:
  - Dictionary form
    - Words are stored in the file line by line and which extracted and stored in a list
- Approach:
  - Scraped all the words in the different recipe
  - Appending the word in a list and then created a dataframe.
  - Automated the process of loading the subsequent pages
  - Cleaned the text
    - Lowercasing of words
    - Removal of parentheses and other symbols and splitting of the text
  - Converted Dataframe to a CSV file
- No. of words scraped = 1000

Word2Vec - Word Embeddings

- Now comes the best part, we took one step ahead and came up with an out of the box solution to get food related words

- Word2vec contains 30 lakh words, and through gensim library and loading word2vec weights we can use cosine_similarity to find the similarity of 2 words, an index of how much these two words are similar to each other

Word2Vec - Word Embeddings

- So we iterated over all the 30 lakh words and checked the cosine similarity index with ‘food’, ‘cuisine’ , ‘drink’, ‘culinary’, ‘utensil’, and 3-4 more food related terms.

- We got very good results with 4200+ words, now this too had some error(these embeddings are not always 100% errorless)

- We tried to do some preprocessing manually, and through python

  And we were able to get the corpus to 2500-3000 words.

Definition/Meaning of a Word

- We have used an api by freedictionary.com, that tends to return a json with full fledged details of a word from the english dictionary, and from there we fetch the meaning.

- If this api fails to give a response(status_code!=200), we use web scraping and scrape the definition from wikipedia Encyclopedia.(encyclopedia.thefreedictionary.com)

Deliverables

- Scraping Codes
- Preprocessing
- Dictionary
  - with meanings (3076 words)
  - without meanings (919 words, the meanings of these words were not available in the api we used, but they have meaningful)

Future Work

- Due to low CPU specifications and time boundations we didn’t get to work on other word embeddings

- Such as Glove Vectors, FastText, Universal Sentence Encoder, etc, all these embeddings provide a cosine similarity function similar to the one provided by gensim library word2vec. So, in a similar way we would be able to find more words from these embeddings and populate our dataset.

References

- www.archanaskitchen.com
- www.food.com
- www.touchbistro.com
- www.englishclub.com/voabulary/food
- models.word2vec – Word2vec embeddings — gensim (radimrehurek.com)
- UNIT 4 CULINARY TERMS.pdf (ihmnotes.in)
- Allrecipes | Recipes, How-Tos, Videos and More
- https://www.kaggle.com/datasets/leadbest/googlenewsvectorsnegative300
- https://towardsdatascience.com/word2vec-made-easy-139a31a4b8ae

Thank You !!

Team Members :

Shivam Verma, 2019272

Manav Saini, 2020581

Vaibhav Sharma, 2019284