Product Categorization

A Text Classifier for Food Data

Evelyn Trautmann, Sebastian Hansmann, Sumit Sidana

Takeaway

41 Million orders last quarter

~ 19 Million products

~ 40 k active restaurants

Active in 12 countries

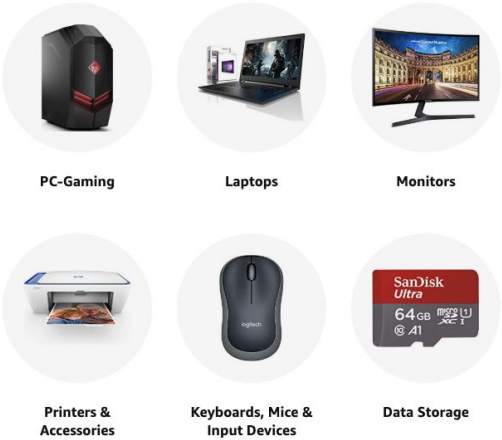
Looking for a Data Engineer right now

Agenda

- Introduction
- FastText
- Open Food Dataset
- Similarities
- Multilabel Classification
- Applications

Introduction

- Product catalogues are common entities in business
- Categorizing unstructured items according to catalogue
- Catalogues often contain a vast amount of classes



Takeaway.com

String Matching

- Misspellings

String Matching

- Misspellings
- Word Concatenations

String Matching

- Misspellings
- Word Concatenations
- Different Contextual Meaning

Text Classifier

fastText

recent NLP developments


FastText
Takeaway.com

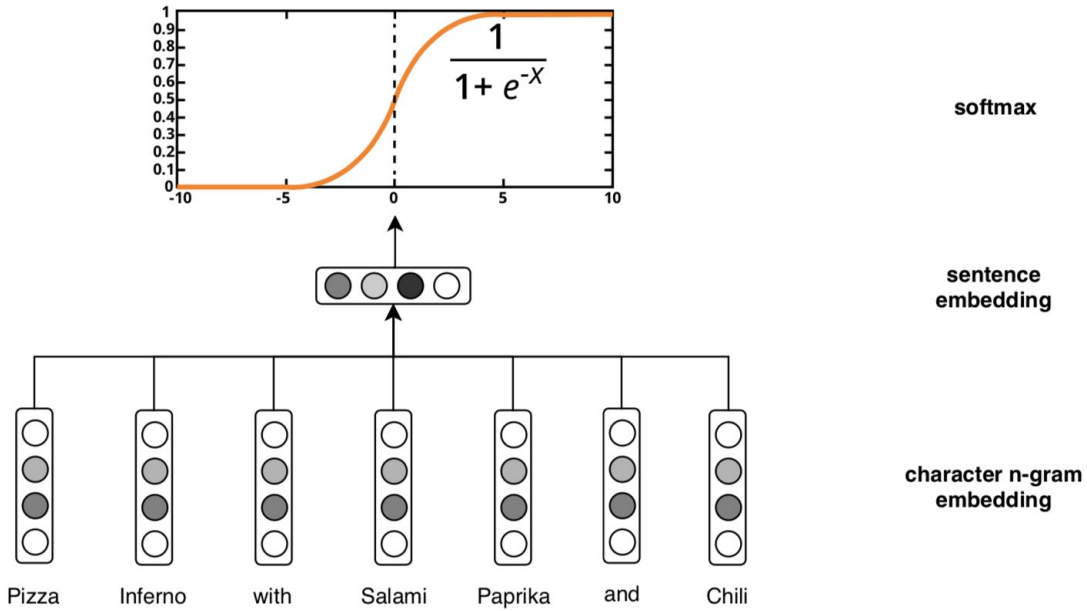
Model Architecture

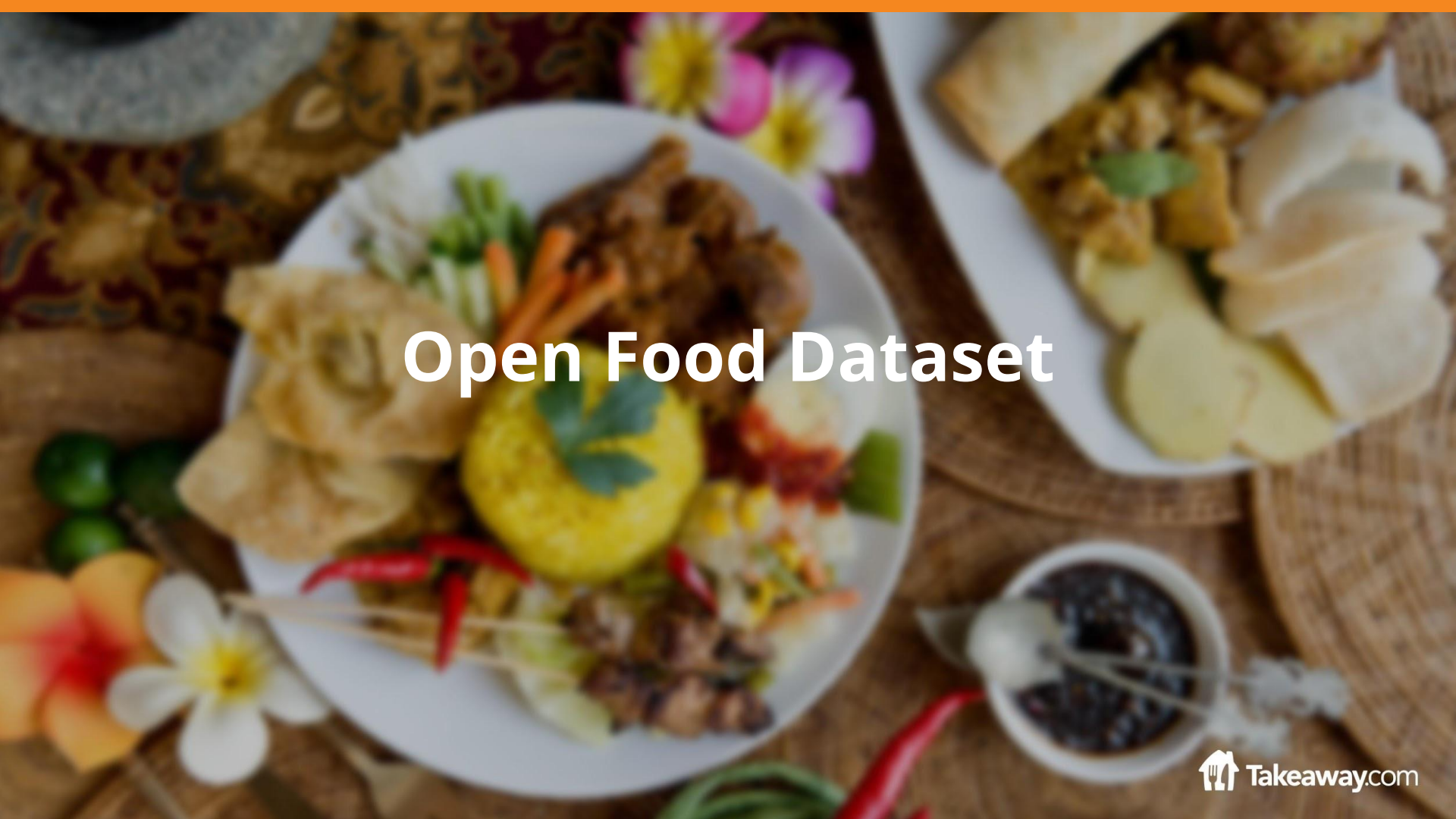
- Character n-gram Embeddings
  - CBOW
  - skip-gram
- Averaging n-gram features
- Softmax

Pizza Inferno with [Salami] Paprika and Chili

[Pizza] [Inferno] [with] Salami [Paprika] [and] [Chili]

Model Architecture




Open Food Dataset
Takeaway.com

Dataset

List of categories - World

25840 categories:

Search:

Category	Products	*
Plant-based foods and beverages	108367	
Plant-based foods	92833	
Snacks	53620	
Beverages	49224	
Sweet snacks	43829	
Dairies	39904	
Cereals and potatoes	31198	
Meats	30341	
Non-Alcoholic beverages	30099	
Fruits and vegetables based foods	30067	
Fermented foods	28201	
Fermented milk products	28118	
Meals	27465	
Groceries	22671	

▪ ▪ ▪

OPEN FOOD FACTS
the free food products database

Takeaway.com

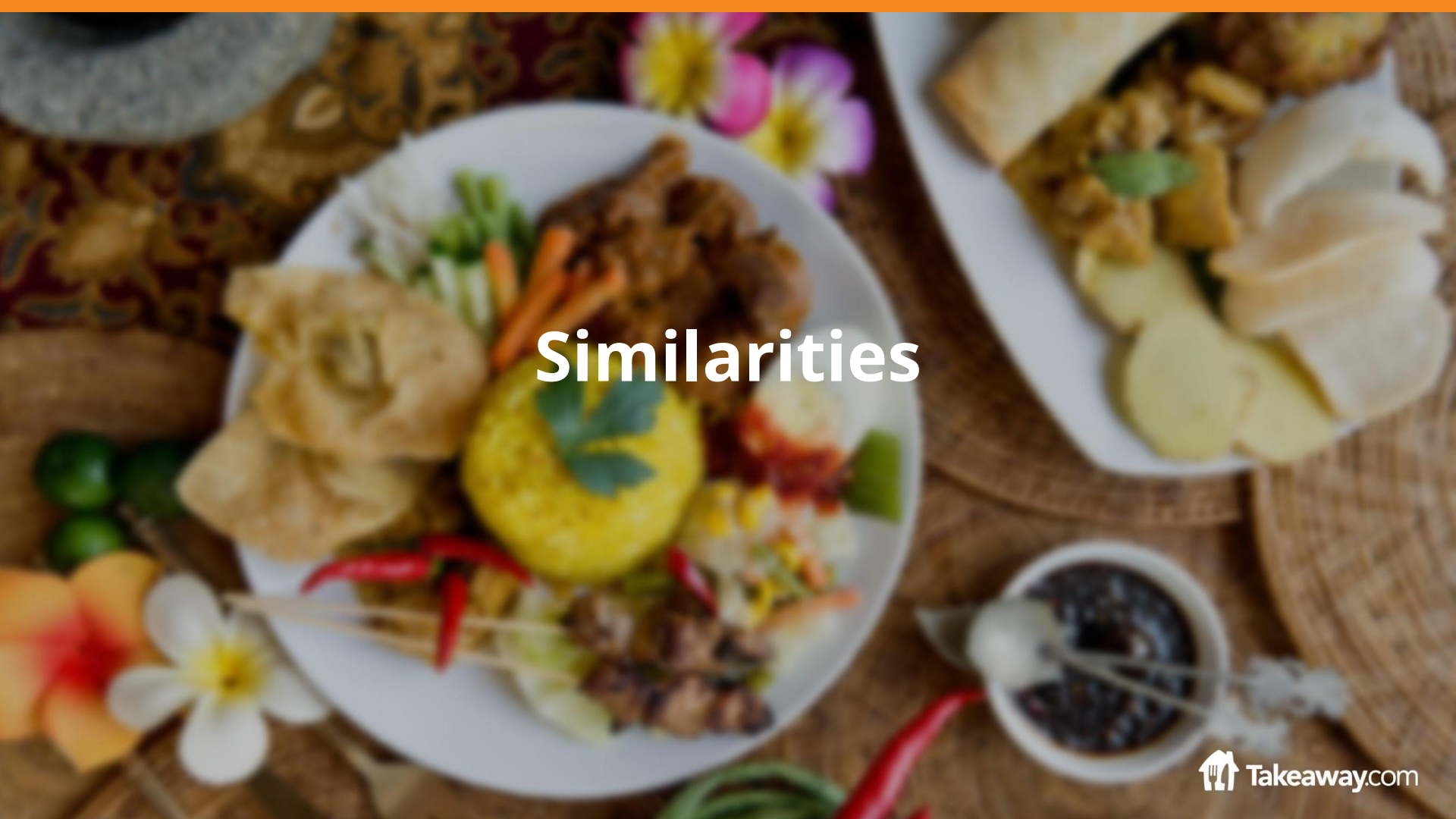
Problem Description

Can we assign appropriate categories to specific products given its product name, generic name and brand?

Classification Demo

http://localhost:8889/notebooks/OpenFoodFactExample.ipynb

GitHub: https://github.com/metterlein/evaluate_supervised/


Similarities
Takeaway.com

Similarity Matrix

- Not all classes miss-classifications are actual miss-classifications
- Some classes exhibit higher similarity

Classification Metrics wrt Similarities

- In Precision calculation numerator changes

From: $$Pr(i) = \frac{C_{ii}}{\sum_j C_{ij}}$$

- Recall likewise

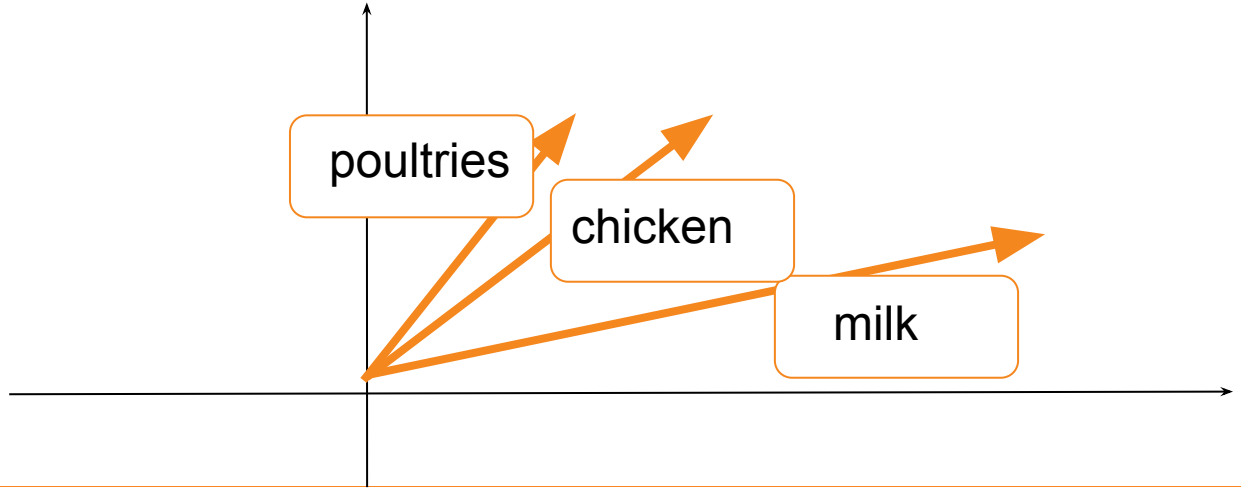
To: $$Pr(i) = \frac{\sum_j C_{ij} S_{ij}}{\sum_j C_{ij}}$$

- Weighted Truth - Prediction Pairs

Confusion Matrix: $$C = \begin{pmatrix} C_{11} & \dots & C_{1n} \\ \dots & & \dots \\ C_{n1} & \dots & C_{nn} \end{pmatrix}$$

Determine Similarity Matrix

- Word Embeddings contain context information
- Labels appearing in similar contexts are close to each other
- Distance between word vectors




Multilabel Classification
Takeaway.com

Evaluation

- Fasttext Default: Precision@k, Recall@k
- Only Average, not per class
- Scikit Learn: Confusion Matrix per Class (OvA)
- How can we evaluate approximately accurate predictions?
- Multilabel Confusion Matrix

Evaluation

- Classification Report
- Multilabel Confusion Matrix

Increase
C[snacks,snacks]
and
C[confectioneries,confectioneries]

Truth	Prediction	Matches
snacks, confectioneries, sweet-snacks	snacks, confectioneries, candies	snacks, confectioneries
chickens, appetizers, poultry-meals	chickens, poultries, prepared-meats	chickens

Matches:
Snacks, confectioneries

Evaluation

- Classification Report
- Multilabel Confusion Matrix

Increase
C[snacks,snacks]
and
C[confectioneries,confectioneries]

Increase
C[sweet-snacks,candies]

Truth	Prediction	Matches
snacks, confectioneries, sweet-snacks	snacks, confectioneries, candies	snacks, confectioneries
chickens, appetizers, poultry-meals	chickens, poultries, prepared-meats	chickens

Miss-classified:
candies

Not detected:
sweet-snacks

Evaluation

- Classification Report
- Multilabel Confusion Matrix

Truth	Prediction	Matches
snacks, confectioneries, sweet-snacks	snacks, confectioneries, candies	snacks, confectioneries
chickens, appetizers, poultry-meals	chickens, poultries, prepared-meats	chickens

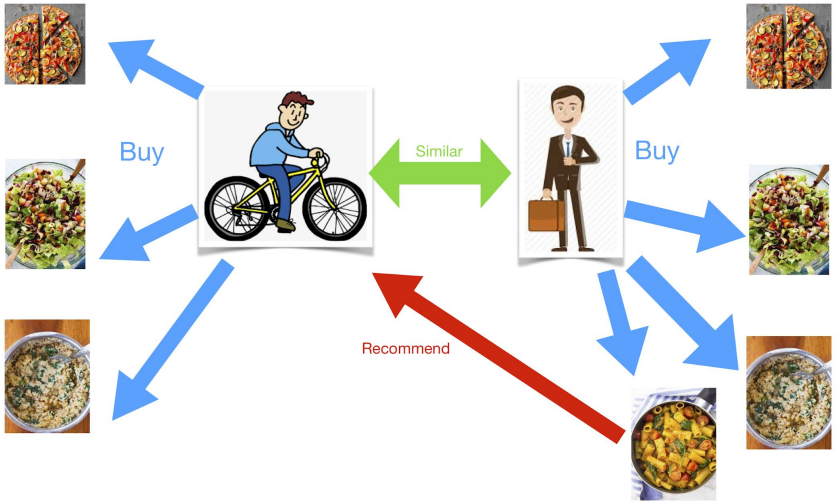
Misclassifications:

- appetizers,poultries
- appetizers,prepared-meats
- poultry-meals,poultries
- poultry-meals,prepared-meats


Application
Takeaway.com

Recommendations

- Collaborative filtering: Make use of past order history to make recommendations and no meta categories
- Recommendations without features: Collaborative filtering (an Ok solution, but not enough!)



Recommendations with categories

- Difficult to do Dish-Based recommendations with Interaction data alone!
  - 6 M distinct dishes
  - Pizza 12 inches = Pizza 24 inches (what we have now)
  - Both need to be classified as Pizza (using categories)
  - Need a common ground (such as a broad category!)
- Collaborative filtering with meta-information for restaurant recommendations
  - Embed categories in CF model for restaurants
  - Handle Restaurant cold start (problem of recommending new restaurants)
- Customer Segmentation
  - Clustering on the basis of categories

References

- Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification (cite arxiv:1607.01759) https://arxiv.org/abs/1607.01759
- https://ai.facebook.com/blog/fasttext-blog-post-open-source-in-brief/
- Wikipedia contributors. "Open Food Facts." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 13 Sep. 2019. Web. 2 Oct. 2019.
- https://github.com/metterlein/evaluate_supervised/

Thank You!

Questions?

Takeaway.com brands

Lieferando.de Lieferservice.at Lieferservice.ch Pizza.be Pizza.fr
Pizza.lu Pyszne.pl Pizza.pt Thuisbezorgd.nl and Vietnammm.com